(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110544035 A
(43)申请公布日 2019.12.06

(21)申请号 201910815700.9

(22)申请日 2019.08.30

(71)申请人 中国南方电网有限责任公司
地址 510000 广东省广州市萝岗区科学城科翔路11号

(72)发明人 刘菱琳 周祖斌 杨志清 曲成 王婷

(74)专利代理机构 佛山粤进知识产权代理事务所(普通合伙) 44463
代理人 王储

(51)Int.Cl.
G06Q 10/06(2012.01)
G06F 16/21(2019.01)
G06F 16/28(2019.01)
G06F 16/953(2019.01)

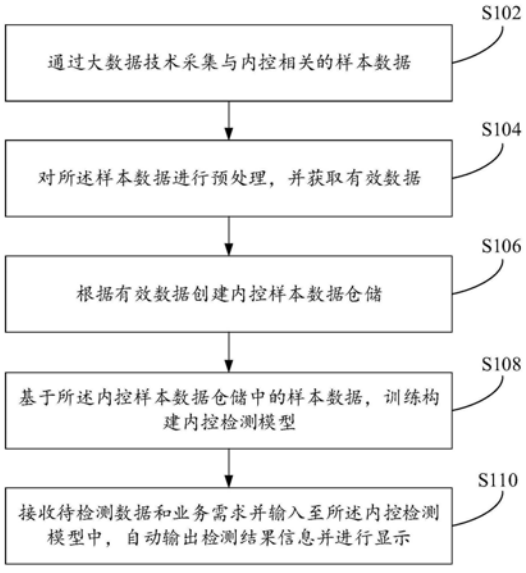
权利要求书2页 说明书12页 附图5页

(54)发明名称

一种内控检测方法、系统和计算机可读存储介质

(57)摘要

本发明提供一种内控检测方法、系统和计算机可读存储介质，所述方法包括：通过大数据技术采集与内控相关的样本数据；对所述样本数据进行预处理，并获取有效数据；根据有效数据创建内控样本数据仓储；基于所述内控样本数据仓储中的样本数据，训练构建内控检测模型；接收待检测数据和业务需求并输入至所述内控检测模型中，自动输出检测结果信息并进行显示。本发明突破技术瓶颈，解决内控管理的业务难题，基于分词技术和非结构化数据解析技术，创建内控检测模型，验证业务数据和财务数据、业务与关联业务等之间的信息正确性，逐步完善内控体系，提升内控测评的效率和效果，促进企业经营的改善。



CN 110544035 A

1.一种内控检测方法,其特征在于,所述方法包括:

通过大数据技术采集与内控相关的样本数据;

对所述样本数据进行预处理,并获取有效数据;

根据有效数据创建内控样本数据仓储;

基于所述内控样本数据仓储中的样本数据,训练构建内控检测模型;

接收待检测数据和业务需求并输入至所述内控检测模型中,自动输出检测结果信息并进行显示。

2.根据权利要求1所述的一种内控检测方法,其特征在于,对所述样本数据进行预处理,并获取有效数据,具体包括:

预先建立有关业务类型的表;

对结构化样本数据按照业务类型进行分类处理;

将结构化样本数据按照分类结果保存在相应的表中。

3.根据权利要求1所述的一种内控检测方法,其特征在于,对所述样本数据进行预处理,并获取有效数据,具体还包括:

当接收到的样本数据为非结构化数据时,定义非结构化数据载体文件格式,并生成标准结构文件;

提取所述标准结构文件的元数据,建立相应的文件模板并将模板信息写入到Oracle数据库的文件模板表中;

根据已经生成的文件模板,进行结构匹配操作,消除结构冲突、语义冲突和联系冲突,提取各专业输出文件中的数据内容写入到相应的Data XML文档中;

将得到的Data XML文档中的数据按照结构匹配、语义匹配和相关算法,写入到Oracle数据库的结果表中,所述结果表中的数据为非结构化数据转换后得到的结构化数据。

4.根据权利要求1所述的一种内控检测方法,其特征在于,对所述样本数据进行预处理,并获取有效数据,具体还包括:

利用一些具有分隔作用的符号将样本数据切分成较短的句子或字串;

输入字串Y_i,调用自适应隐马尔可夫模型进行分词,计算该字串所在段落包含术语集中专业术语的数量;

如果大于某一阈值,调用二阶隐马尔可夫模型进行分词,将字串Y_i分成单词序列$X_i=x_{i1},x_{i2},\cdots,x_{in}$,否则,调用一阶隐马尔可夫模型进行分词,将字串$Y_i$分成单词序列$X_i=x_{i1},x_{i2},\cdots,x_{in}$;

遍历X_i,判断$x_{ij}(j=1,\cdots,n)$是否在领域词典中,若在领域词典中,查找单词x_{ij}的相邻上下单词,并记录单词$x_{ij},x_{i,j-1},x_{i,j+1}$和字串编号i到数组S;

判断数组S是否完全遍历,若遍历结束,则结束分词,并输出分词结果。

5.根据权利要求1所述的一种内控检测方法,其特征在于,通过大数据技术采集与内控相关的样本数据,具体包括:

通过在企业信息系统中设置特定的接口以供转换平台连接和访问,实现对所述企业信息系统产生的非结构化样本数据进行自动采集;和/或

将非结构化样本数据批量复制到移动存储设备中,再从所述移动存储设备复制到转换平台中,实现非结构化样本数据的自动采集;和/或

通过网络爬虫从指定的网站将非结构化样本数据抽取出来，通过相应的转换处理，以结构化的方式存储到转换平台中；

其中，所述转换平台用于将非结构化样本数据转换为结构化样本数据并进行存储处理。

6.根据权利要求1所述的一种内控检测方法，其特征在于，在通过大数据技术采集与内控相关的样本数据之后，所述方法还包括：

将非结构化样本数据资源统一转换成XML文档，通过映射策略将XML文档中的数据加载到关系数据库；和/或

将占用空间较大的非结构化样本数据及其元数据信息存到非关系数据库中，将占用空间不大的元数据信息复制导入到关系数据库中进行管理，以保持样本数据之间的联系；和/或

将非结构化样本数据记录作为key/value进行存储，其中key作为主键，其余数据作为一个整体value，把key和value的每一个属性组成二维子表，将不定长的字段值的存储转化为定长的子块进行存储，并对每一个字段值分配定长空间进行存储。

7.根据权利要求1所述的一种内控检测方法，其特征在于，训练构建内控检测模型，具体包括：

从分析工程款项申请支付业务的内控风险入手，训练构建工程款项申请支付业务测评模型；和/或

从分析成本费用报销支付业务的内控风险入手，训练构建成本费用报销支付业务测评模型；和/或

从分析营销、外部银行、财务数据对账业务的内控风险入手，训练构建营银财三方对账业务内控自动测评模型；和/或

从分析账卡物一致性业务的内控风险入手，训练构建资产管理领域测评模型。

8.根据权利要求1所述的一种内控检测方法，其特征在于，所述内控检测模型训练用的数据库包括：

模型分词库，用于存储从样本数据描述信息中梳理出的内控业务专业词、同义词，以便进行分词解析使用；

HBASE数据库，用于存储内控检测模型所依赖的非结构化数据、结构化数据并经由结构化处理后的数据；

分析结果OLAP，用于存储内控检测模型的分析结果。

9.一种内控检测系统，其特征在于，所述内控检测系统包括：存储器及处理器，所述存储器中包括一种内控检测方法程序，所述内控检测方法程序被所述处理器执行时实现如权利要求1至8任意一项所述的一种内控检测方法的步骤。

10.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质中包括一种内控检测方法程序，所述内控检测方法程序被处理器执行时，实现如权利要求1至8中任意一项所述的一种内控检测方法的步骤。

一种内控检测方法、系统和计算机可读存储介质

技术领域

[0001] 本发明涉及大数据技术领域，尤其涉及一种内控检测方法、系统和计算机可读存储介质。

背景技术

[0002] 目前，一些公司的内控测评依然以线下管理为主、在线监控结果与缺陷整改脱节，未能实现真正意义的内控闭环管理效果，主要原因是缺乏技术手段，亟待开展新技术的应用研究，解决内控管理的业务难题如何发挥内控管理职能部门第二道防线的作用，确保企业风险管理得到落实，并对相关工作做出持续性监控，已经成为企业在开展内控管理过程中待解决的首要任务。

[0003] 互联网、物联网、企业内部信息化技术的不断发展，带领人们步入了大数据的时代，大数据技术可以减少因人为因素主动规避不合规，同时能更全面地对信息数据进行自动判断，减少随机抽样带来的不合规信息遗漏。因此作为企业竞争力重要保障的企业内部控制，应充分借助于大数据进行改变与创新。

[0004] 1.大数据为全面内控管理提供强有力的数据支撑，内控管理部门作为一个综合性的经济监督部门，秉承了用数据说话的传统。内控报告中无论是综合评价，还是揭示问题，无一不是以数字为支撑。全覆盖内控管理的数据对象不仅包含全业务链的基础数据、交易数据的结构化数据，还包含各种报告的文本数据、原始凭证等的图像数据、监控视屏数据、物联网采集数据等，数据规模巨大及种类多样需要大数据的有力支撑。

[0005] 大数据的显著特点之一是其结构化数据、非结构化数据的适时性，在大数据技术下，企业可以适时采集来自于内部信息化平台、互联网、物联网等渠道的大量数据信息，以此为基础，对内部控制效果的适时评价就成为可能，定期报告式监督的时效缺陷就可以得到弥补。其二，大数据还有助于全面的内控监督。大数据另一个显著特点是总体数据的可得性与可分析性，传统审计中所进行的抽样评估的缺陷，在大数据下可以得到避免。基于这种技术的内部控制评价，将更为客观、全面。

[0006] 2.大数据为内控数据分析提供智能化的技术支撑，内控数据分析可以通过对相关领域长年累月形成的数据的分析，挖掘出某种群体行为的特点和问题线索，是未来内控管理的必备手段。在大数据时代，充分利用数据仓库、联机分析、云计算、数据挖掘和数据可视化等技术、把离散存储于不同系统中的海量数据彼此关系并进行深度挖掘分析，可以对企业经营情况、相关内控措施的效果进行评估，从而得出客观的内部控制结论，因此内控数据分析的智能化需要大数据的有力支撑。

[0007] 基于上述需求，目前急需提出一种基于大数据技术的内控检测方法。

发明内容

[0008] 为了解决上述至少一个技术问题，本发明提出了一种内控检测方法，所述方法包括：

[0009] 通过大数据技术采集与内控相关的样本数据；

[0010] 对所述样本数据进行预处理，并获取有效数据；

[0011] 根据有效数据创建内控样本数据仓储；

[0012] 基于所述内控样本数据仓储中的样本数据，训练构建内控检测模型；

[0013] 接收待检测数据和业务需求并输入至所述内控检测模型中，自动输出检测结果信息并进行显示。

[0014] 本方案中，对所述样本数据进行预处理，并获取有效数据，具体包括：

[0015] 预先建立有关业务类型的表；

[0016] 对结构化样本数据按照业务类型进行分类处理；

[0017] 将结构化样本数据按照分类结果保存在相应的表中。

[0018] 本方案中，对所述样本数据进行预处理，并获取有效数据，具体还包括：

[0019] 当接收到的样本数据为非结构化数据时，定义非结构化数据载体文件格式，并生成标准结构文件；

[0020] 提取所述标准结构文件的元数据，建立相应的文件模板并将模板信息写入到Oracle数据库的文件模板表中；

[0021] 根据已经生成的文件模板，进行结构匹配操作，消除结构冲突、语义冲突和联系冲突，提取各专业输出文件中的数据内容写入到相应的Data XML文档中；

[0022] 将得到的Data XML文档中的数据按照结构匹配、语义匹配和相关算法，写入到Oracle数据库的结果表中，所述结果表中的数据为非结构化数据转换后得到的结构化数据。

[0023] 本方案中，对所述样本数据进行预处理，并获取有效数据，具体还包括：

[0024] 利用一些具有分隔作用的符号将样本数据切分成较短的句子或字串；

[0025] 输入字串Y_i，调用自适应隐马尔可夫模型进行分词，计算该字串所在段落包含术语集中专业术语的数量；

[0026] 如果大于某一阈值，调用二阶隐马尔可夫模型进行分词，将字串Y_i分成单词序列$X_i=x_{i1},x_{i2},\cdots,x_{in}$，否则，调用一阶隐马尔可夫模型进行分词，将字串$Y_i$分成单词序列$X_i=x_{i1},x_{i2},\cdots,x_{in}$；

[0027] 遍历X_i，判断$x_{ij}(j=1,\cdots,n)$是否在领域词典中，若在领域词典中，查找单词x_{ij}的相邻上下单词，并记录单词x_{ij}，$x_{i,j-1}$，$x_{i,j+1}$和字串编号i到数组S；

[0028] 判断数组S是否完全遍历，若遍历结束，则结束分词，并输出分词结果。

[0029] 本方案中，通过大数据技术采集与内控相关的样本数据，具体包括：

[0030] 通过在企业信息系统中设置特定的接口以供转换平台连接和访问，实现对所述企业信息系统产生的非结构化样本数据进行自动采集；和/或

[0031] 将非结构化样本数据批量复制到移动存储设备中，再从所述移动存储设备复制到转换平台中，实现非结构化样本数据的自动采集；和/或

[0032] 通过网络爬虫从指定的网站将非结构化样本数据抽取出来，通过相应的转换处理，以结构化的方式存储到转换平台中；

[0033] 其中，所述转换平台用于将非结构化样本数据转换为结构化样本数据并进行存储处理。

[0034] 本方案中，在通过大数据技术采集与内控相关的样本数据之后，所述方法还包括：

[0035] 将非结构化样本数据资源统一转换成XML文档，通过映射策略将XML文档中的数据加载到关系数据库；和/或

[0036] 将占用空间较大的非结构化样本数据及其元数据信息存到非关系数据库中，将占用空间不大的元数据信息复制导入到关系数据库中进行管理，以保持样本数据之间的联系；和/或

[0037] 将非结构化样本数据记录作为key/value进行存储，其中key作为主键，其余数据作为一个整体value，把key和value的每一个属性组成二维子表，将不定长的字段值的存储转化为定长的子块进行存储，并对每一个字段值分配定长空间进行存储。

[0038] 本方案中，训练构建内控检测模型，具体包括：

[0039] 从分析工程款项申请支付业务的内控风险入手，训练构建工程款项申请支付业务测评模型；和/或

[0040] 从分析成本费用报销支付业务的内控风险入手，训练构建成本费用报销支付业务测评模型；和/或

[0041] 从分析营销、外部银行、财务数据对账业务的内控风险入手，训练构建营银财三方对账业务内控自动测评模型；和/或

[0042] 从分析账卡物一致性业务的内控风险入手，训练构建资产管理领域测评模型。

[0043] 本方案中，所述内控检测模型训练用的数据库包括：

[0044] 模型分词库，用于存储从样本数据描述信息中梳理出的内控业务专业词、同义词，以便进行分词解析使用；

[0045] HBASE数据库，用于存储内控检测模型所依赖的非结构化数据、结构化数据并经由结构化处理后的数据；

[0046] 分析结果OLAP，用于存储内控检测模型的分析结果。

[0047] 本发明第二方面还提出一种内控检测系统，所述内控检测系统包括：存储器及处理器，所述存储器中包括一种内控检测方法程序，所述内控检测方法程序被所述处理器执行时实现如上述的一种内控检测方法的步骤。

[0048] 本发明第三方面还提出一种计算机可读存储介质，所述计算机可读存储介质中包括一种内控检测方法程序，所述内控检测方法程序被处理器执行时，实现如上述的一种内控检测方法的步骤。

[0049] 本发明突破技术瓶颈，解决内控管理的业务难题，基于分词技术和非结构化数据解析技术，支撑对文本类资料样本的自动分析和测评。通过对结构化及非结构化业务数据的收集、解析和预处理，结合内控业务检查要求和控制标准，创建内控检测模型，验证业务数据和财务数据、业务与关联业务等之间的信息正确性，逐步完善内控体系，提升内控测评的效率和效果，促进企业经营的改善，实现企业经营目标最大化。

[0050] 本发明的附加方面和优点将在下面的描述部分中给出，部分将从下面的描述中变得明显，或通过本发明的实践了解到。

附图说明

[0051] 图1示出了本发明一种内控检测方法的流程图；

[0052] 图2示出了本发明非结构化数据到结构化数据转换平台的架构图；

[0053] 图3示出了本发明一种分词算法的流程图；

[0054] 图4示出了本发明大数据处理整体技术架构图；

[0055] 图5示出了本发明模型建模方法的流程图；

[0056] 图6示出了本发明内控应用部署的架构图；

[0057] 图7示出了本发明模型数据仓库的架构图；

[0058] 图8示出了本发明一种内控检测系统的框图。

具体实施方式

[0059] 为了能够更清楚地理解本发明的上述目的、特征和优点，下面结合附图和具体实施方式对本发明进行进一步的详细描述。需要说明的是，在不冲突的情况下，本申请的实施例及实施例中的特征可以相互组合。

[0060] 在下面的描述中阐述了很多具体细节以便于充分理解本发明，但是，本发明还可以采用其他不同于在此描述的其他方式来实施，因此，本发明的保护范围并不受下面公开的具体实施例的限制。

[0061] 大数据的关键技术包括以下五部分：大数据感知、分布式数据存储、分布式数据计算、大数据分析、数据可视化。

[0062] 图1示出了本发明一种内控检测方法的流程图。

[0063] 如图1所示，本发明第一方面提出一种内控检测方法，所述方法包括：

[0064] S102，通过大数据技术采集与内控相关的样本数据；

[0065] S104，对所述样本数据进行预处理，并获取有效数据；

[0066] S106，根据有效数据创建内控样本数据仓储；

[0067] S108，基于所述内控样本数据仓储中的样本数据，训练构建内控检测模型；

[0068] S110，接收待检测数据和业务需求并输入至所述内控检测模型中，自动输出检测结果信息并进行显示。

[0069] 可以理解，待检测的数据可以为成本费用报销支付信息、工程款项申请支付信息、财务对账信息、营财对账信息等；业务需求可以为对账业务、申请支付业务等；检测结果可以是对业务的真实性、资料的完整性等所作出的判断结果。

[0070] 根据本发明的实施例，对所述样本数据进行预处理，并获取有效数据，具体包括：

[0071] 预先建立有关业务类型的表；

[0072] 对结构化样本数据按照业务类型进行分类处理；

[0073] 将结构化样本数据按照分类结果保存在相应的表中。

[0074] 需要说明的是，结构化数据(Structured Data)是指具有一定结构性、可以划分为固定的基本组成要素、能通过一个或多个二维表来表示的数据。结构化数据一般存储在关系数据库中，具有一定逻辑结构，可用关系数据库的表或视图表示，使用关系型数据库来管理结构化数据是目前最好的一种方法。

[0075] 具体的，将结构化数据信息存储在预先建立好的关系数据库中，再把数据按业务分类，并设计相应的表，然后将对应的信息保存到相应的表中。实际应用中，这些表便于查询统计，且操作简单、易于维护。

[0076] 根据本发明的实施例，对所述样本数据进行预处理，并获取有效数据，具体还包括：

[0077] 当接收到的样本数据为非结构化数据时，定义非结构化数据载体文件格式，并生成标准结构文件；

[0078] 提取所述标准结构文件的元数据，建立相应的文件模板并将模板信息写入到Oracle数据库的文件模板表中；

[0079] 根据已经生成的文件模板，进行结构匹配操作，消除结构冲突、语义冲突和联系冲突，提取各专业输出文件中的数据内容写入到相应的Data XML文档中；

[0080] 将得到的Data XML文档中的数据按照结构匹配、语义匹配和相关算法，写入到Oracle数据库的结果表中，所述结果表中的数据为非结构化数据转换后得到的结构化数据。

[0081] 需要说明的是，非结构化数据(Unstructured Data)是指结构化数据以外的数据，数据结构不固定，无法使用关系数据库存储，只能够以各种类型的文件形式存放，如office文档、文本文件、办公文档(包括PDF、Rtf、caj等格式的文档)、图片、财务报表、图像、音频和视频等等。

[0082] 由于非结构化数据大多都为文档、视频、图片等。这种数据容量巨大，并且难以通过二维表格分解逻辑存储。本发明选择Oracle数据库管理系统或者云存储的形式来存储非结构化数据。

[0083] Oracle数据存储针对存储体量相对小的文本、文档以及PDF具有速度快、管理和维护简单等优点；云存储是网格、并行和分布式计算等众多技术的发展和延伸，实现了存储的完全虚拟化，提供了更强大的存储和共享功能。

[0084] 如图2所示，非结构化数据到结构化数据的转换平台由数据库、文件系统、模板库、文件格式定义模块、元数据提取模块、模板创建及管理模块、中间数据表示模块和XML数据转换模块等组成。整个系统在架构上分为三个层次：界面应用层、程序逻辑层、数据存储层。

[0085] 界面应用层提供了图形化的数据转换界面给用户使用，通过应用界面，用户可以使用非结构化到结构化数据转换的相关操作，而不必关心数据转换的具体实现。

[0086] 程序逻辑层由系统的五个功能模块构成，工作重点是实现非结构化到结构化数据转换平台的业务逻辑。界面应用层客户端在获取文件系统中的输出文件后，发出数据转换请求，然后，应用程序接收客户端发出的请求，将需要转换的文件传递给数据转换模块。模块收到文件后，根据文件类型进行分类，确定使用哪个程序进行转换。然后，五个功能模块开始工作，首先定义文件格式，完成标准结构文件的生成，提取该文件的元数据，建立相应的文件模板并将模板信息写入到Oracle数据库的文件模板表中，创建结果表结构，然后实现非结构化到半结构化数据转换，最后完成半结构化到结构化数据转换，将处理后的数据写入到Oracle数据库的结果表中。应用程序再将转换结果返回给用户，并提示用户是否进行下一次数据转换，最后完成数据转换的全过程。

[0087] 数据存储层集合了系统中用到的Oracle数据库表，比如文件模板表、文件关联表、结果表。文件模板表、文件关联表需要在系统运行前创建。结果表中的数据就是非结构化数据转换后得到的结构化数据。在数据转换完成后，系统将相关信息写入文件关联表中。

[0088] 各模块的功能如下：

[0089] Oracle数据库:用于存储非结构化输出文件的属性信息、模板信息以及转换后生成的结果表;

[0090] 文件系统:用于存放各种非结构化的输出文件;

[0091] 模板库:用于存放通过系统转换过程中生成的文件模板;

[0092] 文件格式定义模块:将实践中遇到的多种结构类型的非结构化文件统一转换为严格的标准结构文件,利用它可以转换为统一的标准格式的XML文档;

[0093] 元数据提取模块:对输出文件中的元数据直接有效提取,以此来建立元数据及其对应的数据库字段名、字段类型、约束条件等信息;

[0094] 模板创建及管理模块:根据非结构化文件的元数据定义,构建对应的文件模板,然后将这些模板信息存放到模板库和文件模板表中进行管理;

[0095] 中间数据表示模块:根据已经生成的文件模板,进行结构匹配操作,消除结构冲突、语义冲突和联系冲突,提取各专业输出文件中的数据内容写入到相应的Data XML文档中;

[0096] XML数据转换模块:将得到的Data XML文档按照结构匹配、语义匹配和相关算法,将数据写入到Oracle数据库的结果表中。

[0097] 根据本发明的实施例,对所述样本数据进行预处理,并获取有效数据,具体还包括:

[0098] 利用一些具有分隔作用的符号将样本数据切分成较短的句子或字串;

[0099] 输入字串Y_i,调用自适应隐马尔可夫模型进行分词,计算该字串所在段落包含术语集中专业术语的数量;

[0100] 如果大于某一阈值,调用二阶隐马尔可夫模型进行分词,将字串Y_i分成单词序列$X_i=x_{i1},x_{i2},\cdots,x_{in}$,否则,调用一阶隐马尔可夫模型进行分词,将字串$Y_i$分成单词序列$X_i=x_{i1},x_{i2},\cdots,x_{in}$;

[0101] 遍历X_i,判断x_{ij}(j=1,…,n)是否在领域词典中,若在领域词典中,查找单词x_{ij}的相邻上下单词,并记录单词$x_{ij},x_{i,j-1},x_{i,j+1}$和字串编号i到数组S;

[0102] 判断数组S是否完全遍历,若遍历结束,则结束分词,并输出分词结果。

[0103] 首先定义一个术语集,术语集是一系列具有代表性的专业术语集合。对本领域(如电力)中的各个子学科从1到i进行编号,统计每个子学科最常用的n个专业术语。然后将各个子学科的代表性专业术语构成一个集合,编号为i的子学科的术语集为Q_i,则总的术语集为$Q=\bigcup_{i-1}^{N}Q_i$。

[0104] 自适应隐马尔可夫模型的作用原理为:首先根据术语集预先判断待分文档包含专业术语的多少,然后与阈值进行比较,若所包含专业术语的数量大于阈值,则说明该段落包含的专业术语较多,应该进行准确分词,此时调用二阶隐马尔可夫模型进行分词;否则进行快速分词,调用一阶隐马尔可夫模型进行分词。

[0105] 首先根据输入的文档判断本文档属于哪一个子学科领域,假设输入文档为D,所属的子学科编号为m,提取术语集$Q=Q_m$。假设Q_m中含有的代表性术语个数为n,遍历Q_m搜索文档D中含有的代表性专业术语数量,$X=[x_1,x_2,...,x_n]$,x_i表示文档D中含有的编号为i的代表性术语的数量,则文档D含有的代表性专业术语的数量为:$num_D=\sum_{i=1}^{n}x_i$,根据文档字

数num以及比例系数α确定专业术语数量阈值s＝num•α。

[0106] $$model=\begin{cases}一阶隐马尔可夫模型, & num_D<s\\ 二阶隐马尔可夫模型, & num_D>s\end{cases}$$

[0107] 图3示出了本发明一种分词算法的流程图。

[0108] 如图3所示，该算法流程包括数据预处理阶段和分词阶段。

[0109] 在分词计算前，需要对待分词文档进行预处理，即利用一些具有分隔作用的符号将文档切分成较短的句子或字符串，以减少匹配次数，提高分词效率，降低分词难度。先按照段落分隔符进行段落切分，将文档分成多个段落，然后利用标点符号、数字、英文以及构词能力差的单字等分隔符再将段落细分成较短的句子或字符串。

[0110] 假设待分文档为D，经过预处理后有r个段落和s个最小字符串。字串$Y_i=y_{i1},y_{i2},\cdots,y_{in}$，$y_{ij}$表示一个单字。调用自适应隐马尔可夫模型进行分词，然后判断分好的单词是否在领域词典中，若在领域词典中，则判断与该单词相邻的上下单词之间的紧密度，以及其是否需要进行重新分词，否则代入约束矩阵再进行语法和语义约束校准，最后输出分词结果。具体分词步骤如下：

[0111] 步骤1，输入字串Y_i，调用自适应隐马尔可夫模型进行分词，计算该字串所在段落包含术语集中专业术语的数量，如果大于某一阈值，则转入步骤2，否则转入步骤3；

[0112] 步骤2，调用二阶隐马尔可夫模型进行分词，将字串Y_i分成单词序列$X_i=x_{i1},x_{i2},\cdots,x_{in}$，转入步骤4；

[0113] 步骤3，调用一阶隐马尔可夫模型进行分词，将字串Y_i分成单词序列$X_i=x_{i1},x_{i2},\cdots,x_{in}$，转入步骤4；

[0114] 步骤4，遍历X_i，判断x_{ij}(j＝1，…，n)是否在领域词典中，若在领域词典中，则转入步骤5，否则转入步骤6；

[0115] 步骤5，查找单词x_{ij}的相邻上下单词，并记录单词x_{ij}，$x_{i,j-1}$，$x_{i,j+1}$和字串编号i到数组S，转入步骤8；

[0116] 步骤6，将单词代入约束矩阵进行验证，若满足约束矩阵，则转入步骤8，否则记录并剔除该分词方式，转入步骤1；

[0117] 步骤7，判断组合词频率是否大于阈值，如大于阈值，则将编号为i的句子作为字符串输入，转入步骤1，否则转入步骤6；

[0118] 步骤8，判断数组S是否完全遍历，若遍历结束，则结束分词，并输出分词结果，否则遍历整个文档统计组合词$x_{i,j-1}x_{ij}$，$x_{i,j}x_{i,j+1}$的频率，转入步骤7。

[0119] 根据本发明的实施例，通过大数据技术采集与内控相关的样本数据，具体包括：

[0120] 通过在企业信息系统中设置特定的接口以供转换平台连接和访问，实现对所述企业信息系统产生的非结构化样本数据进行自动采集；和/或

[0121] 将非结构化样本数据批量复制到移动存储设备中，再从所述移动存储设备复制到转换平台中，实现非结构化样本数据的自动采集；和/或

[0122] 通过网络爬虫从指定的网站将非结构化样本数据抽取出来，通过相应的转换处理，以结构化的方式存储到转换平台中；

[0123] 其中，所述转换平台用于将非结构化样本数据转换为结构化样本数据并进行存储

处理。

[0124] 需要说明的是，非结构化数据可以来自下属单位生成、采集或购买的数据，以及一些必要的外部数据，比如来自互联网的相关信息等。对于已有的非结构化数据，可以利用系统接口传输或者批量复制数据的方式进行数据采集，对于来自互联网的外部数据，可以采用网络爬取的方式采集数据。

[0125] (1)建立特定系统接口传输数据。对于企业信息系统产生的非结构化数据，在保密性要求不高的情况下，可以在该信息系统中设置特定的接口转换平台连接和访问，便于根据需求，按照一定的频度、内容、范围等限定条件，实现非结构化数据的自动采集。

[0126] (2)批量复制。在保密性要求很高的情况下，出于数据安全考虑，对于此类非结构化数据，可以采用批量复制到移动存储设备中，再从移动存储设备复制到转换平台中对应的企业子系统的方式实现数据采集。

[0127] (3)网络爬取。对于外部网络资源数据，可以采用爬网技术，通过网络爬虫或网站公开API等方式，根据设置好的爬网作业从指定的网站将非结构化数据抽取出来，通过相应的转换处理，以结构化的方式存储到转换平台中。该方式可以支持图片、音频、视频等文件或附件的采集，并将附件与正文进行自动关联。

[0128] 根据本发明的实施例，在通过大数据技术采集与内控相关的样本数据之后，所述方法还包括：

[0129] 将非结构化样本数据资源统一转换成XML文档，通过映射策略将XML文档中的数据加载到关系数据库；和/或

[0130] 将占用空间较大的非结构化样本数据及其元数据信息存到非关系数据库中，将占用空间不大的元数据信息复制导入到关系数据库中进行管理，以保持样本数据之间的联系；和/或

[0131] 将非结构化样本数据记录作为key/value进行存储，其中key作为主键，其余数据作为一个整体value，把key和value的每一个属性组成二维子表，将不定长的字段值的存储转化为定长的子块进行存储，并对每一个字段值分配定长空间进行存储。

[0132] 本发明通过相关大数据技术，采集已确定的业务系统中的业务数据(包括结构化及非结构化数据)，并进行初步分析、预处理。例如，在资金管理中，数据抽取包括银行对账单抽取、对账记录抽取、余额调节表抽取等。

[0133] 图4示出了本发明大数据处理整体技术架构图；

[0134] 如图4所示，本发明的大数据处理整体技术架构主要分为数据源、数据集成、大数据平台、结果存储、数据应用五层。

[0135] 数据源层，支持多个数据源的集成，支持财务管控系统等业务系统的结构化数据源的集成。也支持日志、邮件等半结构化数的集成，还支持办公文件、合同、图片、视频等非结构化数据的集成。主要业务数据包括内控体系、会计凭证、工程报销支付附件、单据信息等；

[0136] 数据集成层，通过ETL抽取、文件适配器、实时数据采集等多种技术从外部数据源获取结构化数据(关系库记录)、半结构化数据(日志、邮件等)、非结构化数据(文件、视频、音频、网络数据流等)及特定需要的实时数据。

[0137] 大数据平台层，数据存储，负责进行大数据的存储，针对全数据类型和多样计算需

求，以海量规模存储、快速查询读取为特征，存储来自各个业务系统的各类数据，支撑数据计算和分析层的高级应用。数据计算，利用分布式计算技术，结合特定场景需要的流式计算和内存计算等技术，对应用涉及计算任务工作流进行合理分配和优化，实现计算任务的及时有效完成。数据分析，提供数据分析、数据挖掘等分析引擎。

[0138] 结果存储，把分析和挖掘结果存储在关系型数据库中，供应用调用。

[0139] 数据应用层，支撑资金支付、营银财三方对账、财卡物一致性三个模型前端应用功能。

[0140] 根据本发明的实施例，训练构建内控检测模型，具体包括：

[0141] 从分析工程款项申请支付业务的内控风险入手，训练构建工程款项申请支付业务测评模型；和/或

[0142] 从分析成本费用报销支付业务的内控风险入手，训练构建成本费用报销支付业务测评模型；和/或

[0143] 从分析营销、外部银行、财务数据对账业务的内控风险入手，训练构建营银财三方对账业务内控自动测评模型；和/或

[0144] 从分析账卡物一致性业务的内控风险入手，训练构建资产管理领域测评模型。

[0145] 在资金管理领域的资金支付和营银财三方对账方面抓取成本费用报销支付业务活动、工程款项申请支付业务活动、财银一次对账、营财二次对账、营财月度对账的业务活动，构建模型；在资产管理领域的账卡物一致性方面抓取资产新增、停产、合并、报废、盘盈盘亏的业务活动，构建模型。

[0146] 资金管理在财务管理中占据重要地位，资金安全风险是财务风险中的重中之重，资金安全管控措施须保证资金安全万无一失。因此采用有效的测评手段，经常性对资金安全管控措施的有效性进行测评，全面客观的评价资金安全内控措施是否落实、内控体系设计是否完善，对于及时进行管理改善、杜绝资金安全风险发生至关重要。

[0147] 在资金管理内控领域构建资金支付业务内控自动测评模型和营银财三方对账业务内控自动测评模型，提供支付单据资料的完整性、及时性检查模型、支付审批流程的合规性检查模型等。其中，资金支付业务内控自动测评模型包括工程款项申请支付业务测评模型和成本费用报销支付业务测评模型。

[0148] 如图5所示，基于大数据的资金管理支付业务、资金管理营银财三方对账业务、资产管理账卡物一致性的模型建模步骤如下：

[0149] 步骤1，数据模型定义：数据模型定义描述系统如何设置、配置业务模型；

[0150] 步骤2，数据模型存储：数据的物理数据模型；

[0151] 步骤3，数据模型建立：建立数据的逻辑模型；

[0152] 步骤4，算法模型建立：结合语义分析、数据挖掘等建立三个模型；

[0153] 步骤5，任务执行：主要包括如下两步：系统自动生成任务、系统手动执行任务；

[0154] 步骤6，可视化：资金管理支付模型、资金管理营银财三方对账模型、资产管理账卡物一致性模型前台展示。

[0155] 如图6所示，模型分析的业务从内控应用服务中获取后存储在大数据平台，并通过非结构化转结构化处理输入到资金管理支付模型、资金管理营银财三方对账模型、资产管理账卡物一致性模型，通过模型数据挖挖掘与分析并将结果数据存储在模型结果OLAP中，

最后通过内控业务分析模型应用服务进行前台展示。

[0156] 如图7所示，所述内控检测模型训练用的数据库包括：

[0157] 模型分词库，用于存储从样本数据描述信息中梳理出的内控业务专业词、同义词，以便进行分词解析使用；具体的，可以根据内控体系、会计凭证等描述信息梳理内控业务专业词、同义词等并建立模型分词库，以便进行分词解析使用；

[0158] HBASE数据库，用于存储内控检测模型所依赖的非结构化数据、结构化数据并经由结构化处理后的数据；

[0159] 分析结果OLAP，用于存储内控检测模型的分析结果。

[0160] 图8示出了本发明一种内控检测系统的框图。

[0161] 如图8所示，本发明第二方面还提出一种内控检测系统8，所述内控检测系统8包括：存储器81及处理器82，所述存储器81中包括一种内控检测方法程序，所述内控检测方法程序被所述处理器执行时实现如下步骤：

[0162] 通过大数据技术采集与内控相关的样本数据；

[0163] 对所述样本数据进行预处理，并获取有效数据；

[0164] 根据有效数据创建内控样本数据仓储；

[0165] 基于所述内控样本数据仓储中的样本数据，训练构建内控检测模型；

[0166] 接收待检测数据和业务需求并输入至所述内控检测模型中，自动输出检测结果信息并进行显示。

[0167] 进一步的，对所述样本数据进行预处理，并获取有效数据，具体包括：

[0168] 预先建立有关业务类型的表；

[0169] 对结构化样本数据按照业务类型进行分类处理；

[0170] 将结构化样本数据按照分类结果保存在相应的表中。

[0171] 进一步的，对所述样本数据进行预处理，并获取有效数据，具体还包括：

[0172] 当接收到的样本数据为非结构化数据时，定义非结构化数据载体文件格式，并生成标准结构文件；

[0173] 提取所述标准结构文件的元数据，建立相应的文件模板并将模板信息写入到Oracle数据库的文件模板表中；

[0174] 根据已经生成的文件模板，进行结构匹配操作，消除结构冲突、语义冲突和联系冲突，提取各专业输出文件中的数据内容写入到相应的Data XML文档中；

[0175] 将得到的Data XML文档中的数据按照结构匹配、语义匹配和相关算法，写入到Oracle数据库的结果表中，所述结果表中的数据为非结构化数据转换后得到的结构化数据。

[0176] 进一步的，对所述样本数据进行预处理，并获取有效数据，具体还包括：

[0177] 利用一些具有分隔作用的符号将样本数据切分成较短的句子或字串；

[0178] 输入字串Y_i，调用自适应隐马尔可夫模型进行分词，计算该字串所在段落包含术语集中专业术语的数量；

[0179] 如果大于某一阈值，调用二阶隐马尔可夫模型进行分词，将字串Y_i分成单词序列$X_i=x_{i1},x_{i2},\cdots,x_{in}$，否则，调用一阶隐马尔可夫模型进行分词，将字串$Y_i$分成单词序列$X_i=x_{i1},x_{i2},\cdots,x_{in}$；

[0180] 遍历X_i,判断x_{ij}(j＝1,…,n)是否在领域词典中,若在领域词典中,查找单词x_{ij}的相邻上下单词,并记录单词x_{ij},$x_{i,j-1}$,$x_{i,j+1}$和字串编号i到数组S;

[0181] 判断数组S是否完全遍历,若遍历结束,则结束分词,并输出分词结果。

[0182] 进一步的,通过大数据技术采集与内控相关的样本数据,具体包括:

[0183] 通过在企业信息系统中设置特定的接口以供转换平台连接和访问,实现对所述企业信息系统产生的非结构化样本数据进行自动采集;和/或

[0184] 将非结构化样本数据批量复制到移动存储设备中,再从所述移动存储设备复制到转换平台中,实现非结构化样本数据的自动采集;和/或

[0185] 通过网络爬虫从指定的网站将非结构化样本数据抽取出来,通过相应的转换处理,以结构化的方式存储到转换平台中;

[0186] 其中,所述转换平台用于将非结构化样本数据转换为结构化样本数据并进行存储处理。

[0187] 进一步的,在通过大数据技术采集与内控相关的样本数据之后,还包括:

[0188] 将非结构化样本数据资源统一转换成XML文档,通过映射策略将XML文档中的数据加载到关系数据库;和/或

[0189] 将占用空间较大的非结构化样本数据及其元数据信息存到非关系数据库中,将占用空间不大的元数据信息复制导入到关系数据库中进行管理,以保持样本数据之间的联系;和/或

[0190] 将非结构化样本数据记录作为key/value进行存储,其中key作为主键,其余数据作为一个整体value,把key和value的每一个属性组成二维子表,将不定长的字段值的存储转化为定长的子块进行存储,并对每一个字段值分配定长空间进行存储。

[0191] 进一步的,训练构建内控检测模型,具体包括:

[0192] 从分析工程款项申请支付业务的内控风险入手,训练构建工程款项申请支付业务测评模型;和/或

[0193] 从分析成本费用报销支付业务的内控风险入手,训练构建成本费用报销支付业务测评模型;和/或

[0194] 从分析营销、外部银行、财务数据对账业务的内控风险入手,训练构建营银财三方对账业务内控自动测评模型;和/或

[0195] 从分析账卡物一致性业务的内控风险入手,训练构建资产管理领域测评模型。

[0196] 进一步的,所述内控检测模型训练用的数据库包括:

[0197] 模型分词库,用于存储从样本数据描述信息中梳理出的内控业务专业词、同义词,以便进行分词解析使用;

[0198] HBASE数据库,用于存储内控检测模型所依赖的非结构化数据、结构化数据并经由结构化处理后的数据;

[0199] 分析结果OLAP,用于存储内控检测模型的分析结果。

[0200] 本发明第三方面还提出一种计算机可读存储介质,所述计算机可读存储介质中包括一种内控检测方法程序,所述内控检测方法程序被处理器执行时,实现如上述的一种内控检测方法的步骤。

[0201] 本发明突破技术瓶颈,解决内控管理的业务难题,基于分词技术和非结构化数据

解析技术，支撑对文本类资料样本的自动分析和测评。通过对结构化及非结构化业务数据的收集、解析和预处理，结合内控业务检查要求和控制标准，创建内控检测模型，验证业务数据和财务数据、业务与关联业务等之间的信息正确性，逐步完善内控体系，提升内控测评的效率和效果，促进企业经营的改善，实现企业经营目标最大化。

[0202] 在本申请所提供的几个实施例中，应该理解到，所揭露的设备和方法，可以通过其它的方式实现。以上所描述的设备实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，如：多个单元或组件可以结合，或可以集成到另一个系统，或一些特征可以忽略，或不执行。另外，所显示或讨论的各组成部分相互之间的耦合、或直接耦合、或通信连接可以是通过一些接口，设备或单元的间接耦合或通信连接，可以是电性的、机械的或其它形式的。

[0203] 上述作为分离部件说明的单元可以是、或也可以不是物理上分开的，作为单元显示的部件可以是、或也可以不是物理单元；既可以位于一个地方，也可以分布到多个网络单元上；可以根据实际的需要选择其中的部分或全部单元来实现本实施例方案的目的。

[0204] 另外，在本发明各实施例中的各功能单元可以全部集成在一个处理单元中，也可以是各单元分别单独作为一个单元，也可以两个或两个以上单元集成在一个单元中；上述集成的单元既可以采用硬件的形式实现，也可以采用硬件加软件功能单元的形式实现。

[0205] 本领域普通技术人员可以理解：实现上述方法实施例的全部或部分步骤可以通过程序指令相关的硬件来完成，前述的程序可以存储于计算机可读取存储介质中，该程序在执行时，执行包括上述方法实施例的步骤；而前述的存储介质包括：移动存储设备、只读存储器(ROM，Read-Only Memory)、随机存取存储器(RAM，Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0206] 或者，本发明上述集成的单元如果以软件功能模块的形式实现并作为独立的产品销售或使用时，也可以存储在一个计算机可读取存储介质中。基于这样的理解，本发明实施例的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备(可以是个人计算机、服务器、或者网络设备等)执行本发明各个实施例所述方法的全部或部分。而前述的存储介质包括：移动存储设备、ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

[0207] 以上所述，仅为本发明的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应以所述权利要求的保护范围为准。

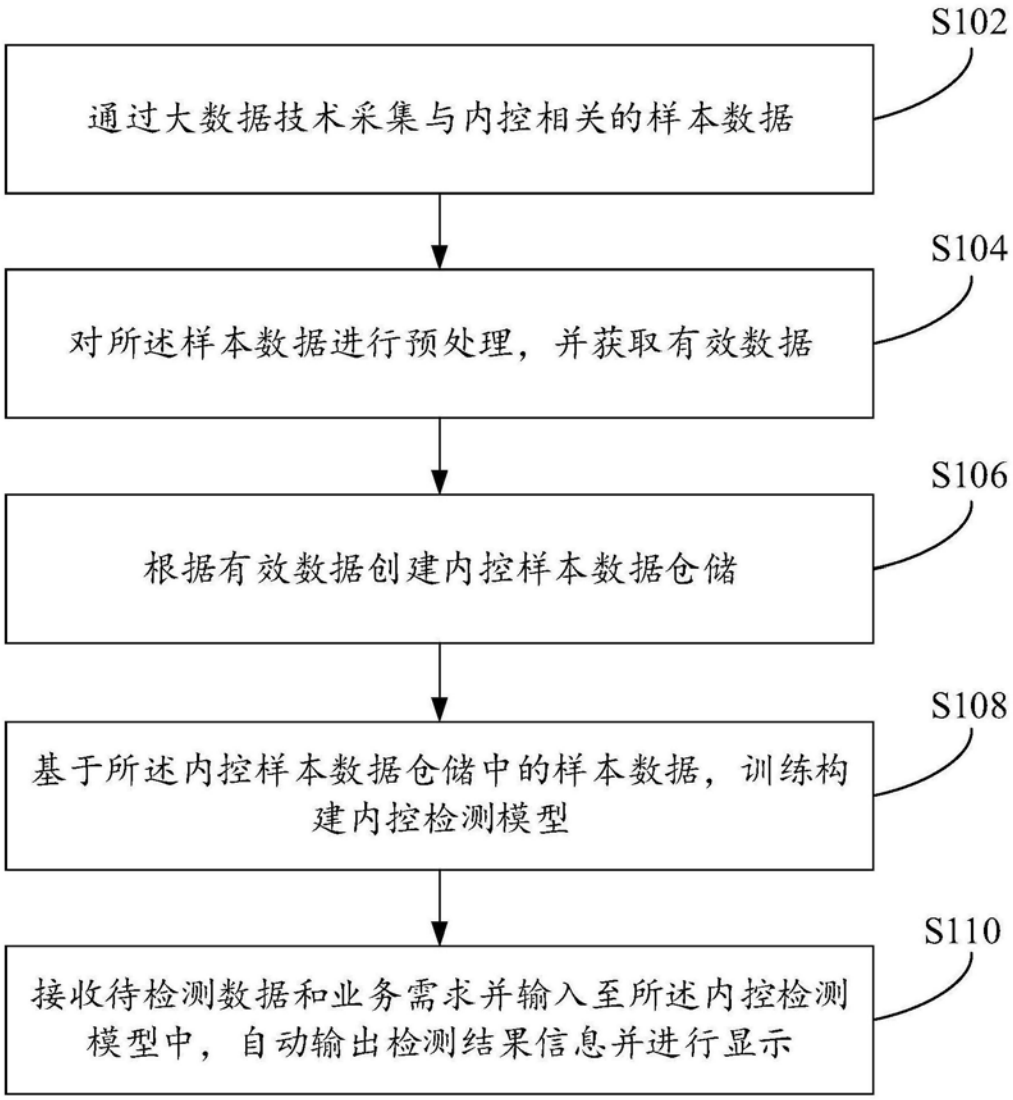

S102
通过大数据技术采集与内控相关的样本数据
S104
对所述样本数据进行预处理，并获取有效数据
S106
根据有效数据创建内控样本数据仓储
S108
基于所述内控样本数据仓储中的样本数据，训练构建内控检测模型
S110
接收待检测数据和业务需求并输入至所述内控检测模型中，自动输出检测结果信息并进行显示

图1

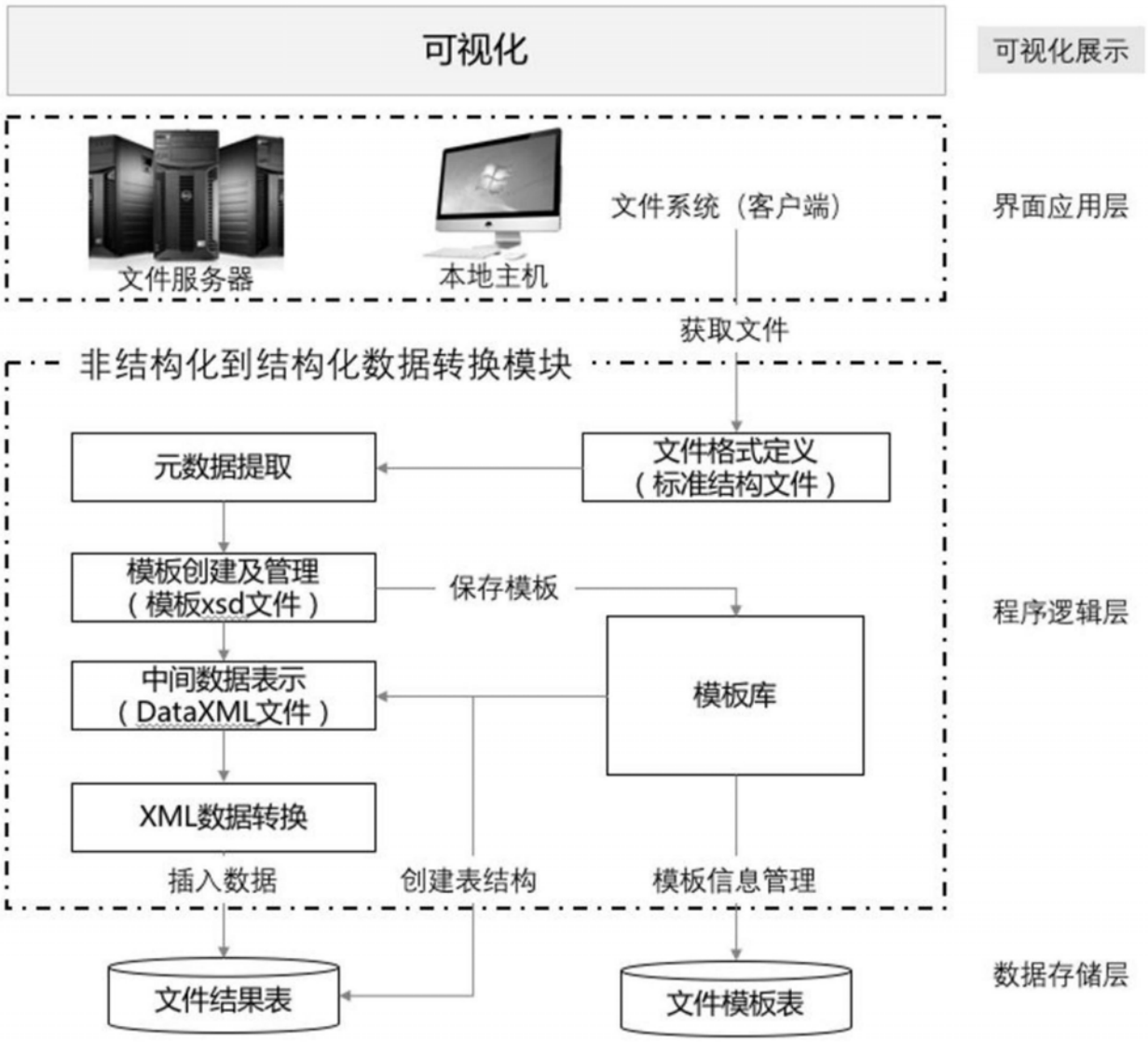

可视化
可视化展示
文件服务器
本地主机
文件系统（客户端）
界面应用层
获取文件
非结构化到结构化数据转换模块
元数据提取
文件格式定义
（标准结构文件）
模板创建及管理
（模板xsd文件）
保存模板
中间数据表示
（DataXML文件）
模板库
XML数据转换
程序逻辑层
插入数据
创建表结构
模板信息管理
文件结果表
文件模板表
数据存储层

图2

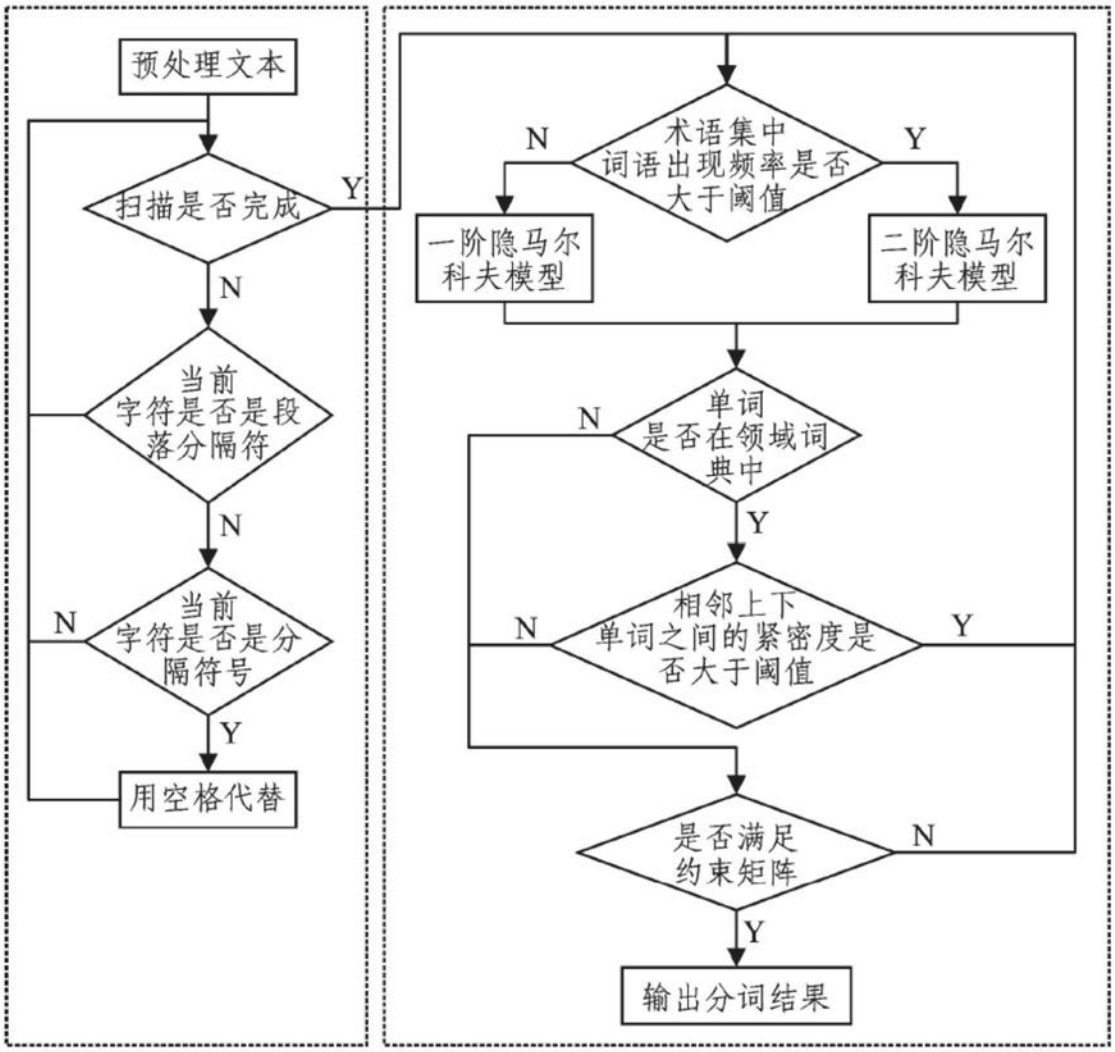

预处理文本
扫描是否完成
Y
N
当前字符是否是段落分隔符
N
当前字符是否是分隔符号
N
Y
用空格代替
术语集中词语出现频率是否大于阈值
N
Y
一阶隐马尔科夫模型
二阶隐马尔科夫模型
单词是否在领域词典中
N
Y
相邻上下单词之间的紧密度是否大于阈值
N
Y
是否满足约束矩阵
N
Y
输出分词结果

图3

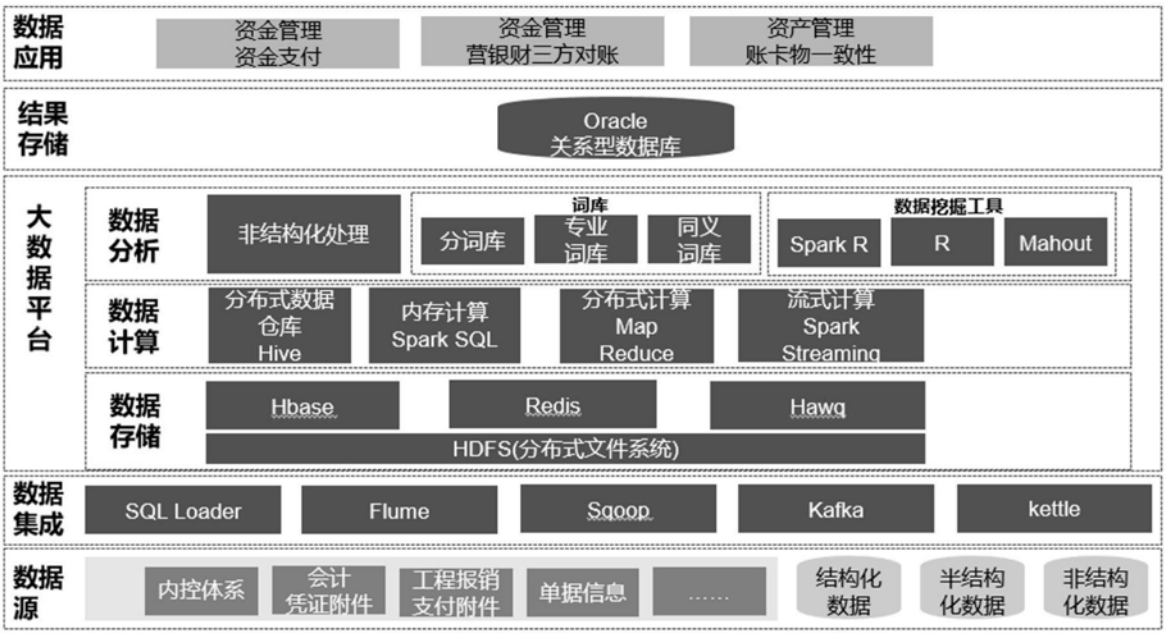

数据应用
资金管理 资金支付
资金管理 营银财三方对账
资产管理 账卡物一致性
结果存储
Oracle 关系型数据库
大数据平台
数据分析
非结构化处理
词库
分词库
专业词库
同义词库
数据挖掘工具
Spark R
R
Mahout
数据计算
分布式数据仓库 Hive
内存计算 Spark SQL
分布式计算 Map Reduce
流式计算 Spark Streaming
数据存储
Hbase
Redis
Hawq
HDFS(分布式文件系统)
数据集成
SQL Loader
Flume
Sqoop
Kafka
kettle
数据源
内控体系
会计凭证附件
工程报销支付附件
单据信息
……
结构化数据
半结构化数据
非结构化数据

图4

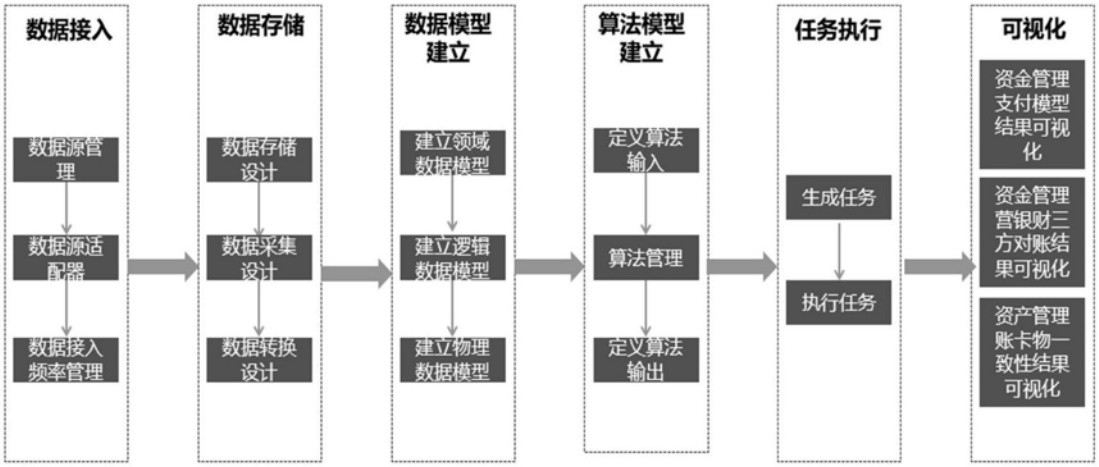

数据接入
数据源管理
数据源适配器
数据接入频率管理
数据存储
数据存储设计
数据采集设计
数据转换设计
数据模型建立
建立领域数据模型
建立逻辑数据模型
建立物理数据模型
算法模型建立
定义算法输入
算法管理
定义算法输出
任务执行
生成任务
执行任务
可视化
资金管理支付模型结果可视化
资金管理营银财三方对账结果可视化
资产管理账卡物一致性结果可视化

图5

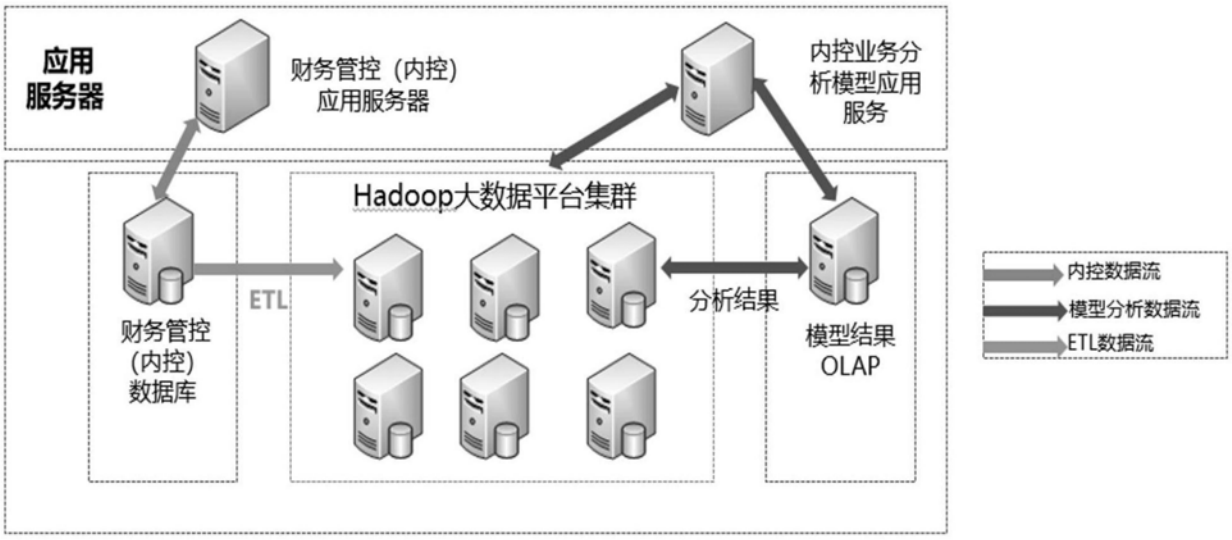

应用
服务器
财务管控（内控）
应用服务器
内控业务分
析模型应用
服务
Hadoop大数据平台集群
ETL
财务管控
（内控）
数据库
分析结果
模型结果
OLAP
内控数据流
模型分析数据流
ETL数据流

图6

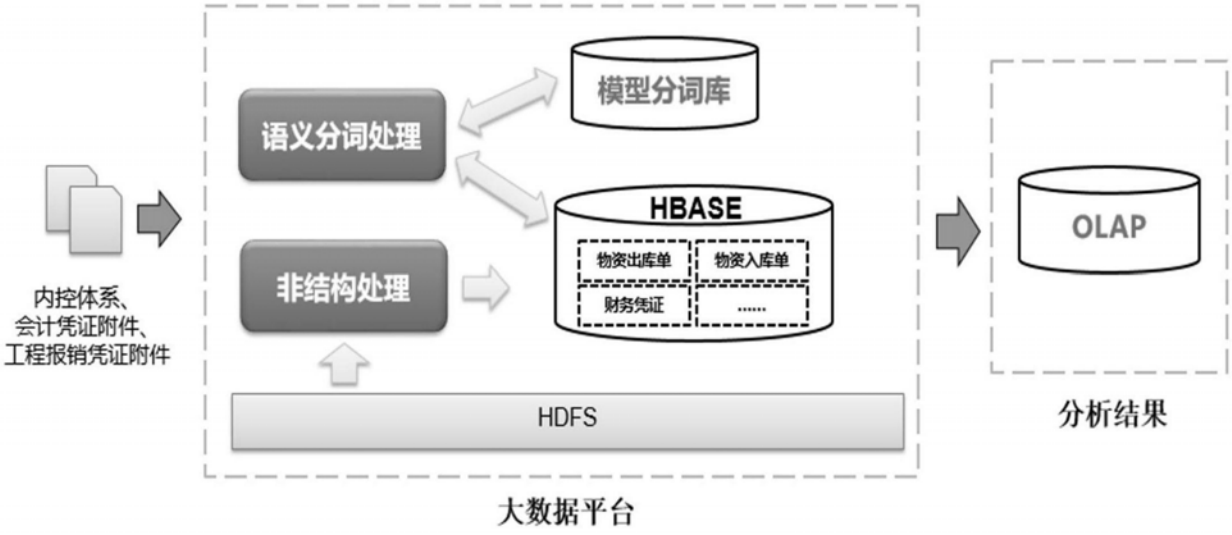

模型分词库
语义分词处理
HBASE
物资出库单
物资入库单
财务凭证
.....
非结构处理
内控体系、
会计凭证附件、
工程报销凭证附件
HDFS
OLAP
分析结果
大数据平台

图7

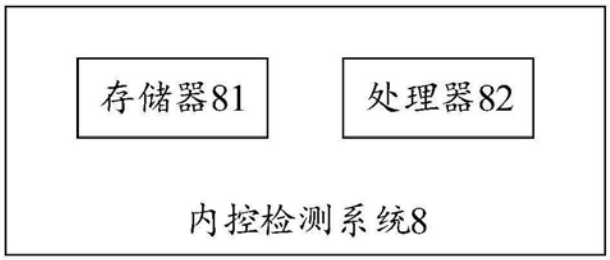

存储器81
处理器82
内控检测系统8

图8