(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 108021632 A
(43)申请公布日 2018.05.11

(21)申请号 201711183952.1

(22)申请日 2017.11.23

(71)申请人 中国移动通信集团河南有限公司
地址 450008 河南省郑州市金水区经三路48号
申请人 北京思诺博信息技术股份有限公司

(72)发明人 曾磊 杨冠强 杨建军 黄宇 贺延敏 王欣 辛朝 肖志立 宋亚丽 裴照华 杨继学 陈海伟 刘岩 陈健 高朗 韩志勇

(74)专利代理机构 郑州大通专利商标代理有限公司 41111
代理人 陈勇

(51)Int.Cl.
G06F 17/30(2006.01)

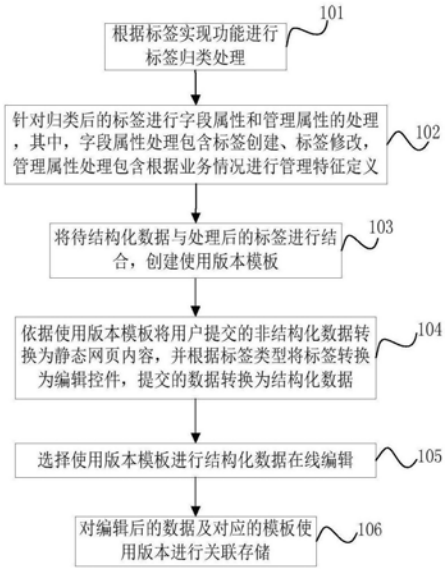
权利要求书2页 说明书6页 附图2页

(54)发明名称

非结构化数据与结构化数据相互转换处理方法

(57)摘要

本发明属于信息处理技术领域，特别涉及一种非结构化数据与结构化数据相互转换处理方法，包含：根据标签实现功能进行标签归类处理；针对标签进行字段属性和管理属性的处理；将待结构化数据与标签进行结合，创建使用版本模板；依据使用版本模板将用户提交的非结构化数据转换为静态网页内容，并根据标签类型将标签转换为编辑控件，提交的数据转换为结构化数据；选择使用版本模板进行结构化数据在线编辑；对编辑后的数据及对应的模板使用版本进行关联存储。本发明通过对非结构化数据文件的标签定义、模板化处理及非结构化数据文件的生成与转换存储，实现对非结构化数据与结构化数据之间的相互转换，便于后续数据分析处理，具有较好的应用价值。



CN 108021632 A

1.一种非结构化数据与结构化数据相互转换处理方法,其特征在于,包含如下内容:

根据标签实现功能进行标签归类处理;

针对归类后的标签进行字段属性和管理属性的处理,其中,字段属性处理包含标签创建、标签修改,管理属性处理包含根据业务情况进行管理特征定义;

将待结构化数据与处理后的标签进行结合,创建使用版本模板;

依据使用版本模板将用户提交的非结构化数据转换为静态网页内容,并根据标签类型将标签转换为编辑控件,提交的数据转换为结构化数据;

选择使用版本模板进行结构化数据在线编辑;

对编辑后的数据及对应的模板使用版本进行关联存储。

2.根据权利要求1所述的非结构化数据与结构化数据相互转换处理方法,其特征在于,标签归类处理包含:根据标签实现功能归类为对应的标签类型中,标签类型包含:字符型标签、数值型标签、日期型标签、引用型标签和选择型标签。

3.根据权利要求1所述的非结构化数据与结构化数据相互转换处理方法,其特征在于,字段属性处理包含依据信息抽取原则创建标签,及对标签类型及该类型的相关属性信息的修改,其中,该类型的相关属性信息至少包含标签长度和标签默认值。

4.根据权利要求3所述的非结构化数据与结构化数据相互转换处理方法,其特征在于,依据信息抽取原则创建标签中的信息抽取原则包含如下内容:信息内容存在变动趋势,信息内容中存在用于数据分析的关键字,信息内容中存在有标识数据特征的内容。

5.根据权利要求1所述的非结构化数据与结构化数据相互转换处理方法,其特征在于,管理属性处理包含根据业务情况进行管理特征定义,对定义标签赋予管理特质分类。

6.根据权利要求5所述的非结构化数据与结构化数据相互转换处理方法,其特征在于,管理特质分类的分类内容如下:通用性特质属性和管理策略性特质属性,其中,管理策略性特质属性包含采购策略、风险管控策略、交付策略、质量策略和评估策略。

7.根据权利要求1所述的非结构化数据与结构化数据相互转换处理方法,其特征在于,将待结构化数据与处理后的标签进行结合来创建使用版本模板,包含内容如下:使用WORD作为模板编译载体,在待结构化数据中需要结构化处理的地方插入处理后的标签,并记录标签所在待结构化数据中的位置,创建使用版本模板,并存储至后台数据库。

8.根据权利要求7所述的非结构化数据与结构化数据相互转换处理方法,其特征在于,创建的使用版本模板,每次存储至后台数据库均自动生成一个新的模板版本,用户使用时自动推荐最新版本或提供多个版本供用户选择。

9.根据权利要求1所述的非结构化数据与结构化数据相互转换处理方法,其特征在于,提交的数据转换为结构化数据,包含如下内容:读取提交数据中的各章节内容,将非常标签转换为静态数据,根据标签类型将字符型标签转换为文本输入框,将选择型标签转换为下拉列表,将日期型标签转换为日期控件;并将需要结构化的模板文件按章节、段落、条目进行拆解,将标签进行层次划分,将拆解后每一层次中需结构化处理的标签进行抽取形成导航目录;生成辅助信息窗口,当用户标签发生变动,根据后台数据库,自动生成辅助信息,该辅助信息包含历史数据部分、信息推荐部分、数据变动提醒部分。

10.根据权利要求9所述的非结构化数据与结构化数据相互转换处理方法,其特征在于,进行结构化数据在线编辑中,通过导航目录定位标签,根据标签类型规范用户输入数

据,并通过辅助信息窗口为用户提供输入帮助;进一步地,对编辑后的数据及对应的模板使用版本进行关联存储时,提交静态网页的form表单,将form表单数据与模板使用版本关联并保存至后台数据库,用于后期历史数据查询及结构化数据往非结构化数据的转换。

非结构化数据与结构化数据相互转换处理方法

技术领域

[0001] 本发明属于信息处理技术领域，特别涉及一种非结构化数据与结构化数据相互转换处理方法。

背景技术

[0002] 对于供应链管理系统中需要使用的非结构化数据文档，通常都是以人工写好文档以后，以附件的方式上传到系统中，文档的编写过程与系统无关，文档中能够体现管控要求的关键条款也无法在系统中以结构化的方式进行存储。供应链系统中非结构化数据与结构化数据之间无法相互转换，也就无法解决如下问题：第一，非结构化的采购文件信息，只能通过手工处理，效率低；第二，无法通过系统化的方式快速、高效对采购的合法性、合规性进行检查，采购的合法、合规性得不到及时、有效控制；第三，非结构化文件中可以体现供应链管理思想与策略的关键性条款，无法通过系统化的方法进行固化，也无法综合检验，对比分析，不利于供应链管理的整体循环改进。现有的技术方案，通常是解决非结构化数据的检索、传输和存储问题，未实现非结构化数据的自动生成以及与结构化数据之间的相互转换与验证分析。现有技术通常是直接将非结构化文件以附件的方式上传的系统中，系统仅是对非结构化文件提供了存储和下载功能，其主要的缺点如下：非结构化文件编制与系统脱节，效率低；非结构化文件中关键的条款无法与结构数据进行关联，因此无法以系统化的手段对业务的实际运行情况进行监督；非结构化文件数据与结构化数据相互隔离，不利于对系统数据进行后续的分析处理，数据价值挖掘受到限制。

发明内容

[0003] 针对现有技术中的不足，本发明提供一种非结构化数据与结构化数据相互转换处理方法，以结构化数据标签的方式，生成非结构化数据文件，并实现对非结构化数据与结构化数据之间的相互转换，提高文件编制的效率，实现信息制作和传播效益的最大化，利于后续数据的分析处理。

[0004] 按照本发明所提供的设计方案，一种非结构化数据与结构化数据相互转换处理方法，包含如下内容：

[0005] 根据标签实现功能进行标签归类处理；

[0006] 针对归类后的标签进行字段属性和管理属性的处理，其中，字段属性处理包含标签创建、标签修改，管理属性处理包含根据业务情况进行管理特征定义；

[0007] 将待结构化数据与处理后的标签进行结合，创建使用版本模板；

[0008] 依据使用版本模板将用户提交的非结构化数据转换为静态网页内容，并根据标签类型将标签转换为编辑控件，提交的数据转换为结构化数据；

[0009] 选择使用版本模板进行结构化数据在线编辑；

[0010] 对编辑后的数据及对应的模板使用版本进行关联存储。

[0011] 上述的，标签归类处理包含：根据标签实现功能归类为对应的标签类型中，标签类

型包含:字符型标签、数值型标签、日期型标签、引用型标签和选择型标签。

[0012] 上述的,字段属性处理包含依据信息抽取原则创建标签,及对标签类型及该类型的相关属性信息的修改,其中,该类型的相关属性信息至少包含标签长度和标签默认值。

[0013] 优选的,依据信息抽取原则创建标签中的信息抽取原则包含如下内容:信息内容存在变动趋势,信息内容中存在用于数据分析的关键字,信息内容中存在有标识数据特征的内容。

[0014] 上述的,管理属性处理包含根据业务情况进行管理特征定义,对定义标签赋予管理特质分类。

[0015] 优选的,管理特质分类的分类内容如下:通用性特质属性和管理策略性特质属性,其中,管理策略性特质属性包含采购策略、风险管控策略、交付策略、质量策略和评估策略。

[0016] 上述的,将待结构化数据与处理后的标签进行结合来创建使用版本模板,包含内容如下:使用WORD作为模板编译载体,在待结构化数据中需要结构化处理的地方插入处理后的标签,并记录标签所在待结构化数据中的位置,创建使用版本模板,并存储至后台数据库。

[0017] 优选的,创建的使用版本模板,每次存储至后台数据库均自动生成一个新的模板版本,用户使用时自动推荐最新版本或提供多个版本供用户选择。

[0018] 上述的,提交的数据转换为结构化数据,包含如下内容:读取提交数据中的各章节内容,将非常标签转换为静态数据,根据标签类型将字符型标签转换为文本输入框,将选择型标签转换为下拉列表,将日期型标签转换为日期控件;并将需要结构化的模板文件按章节、段落、条目进行拆解,将标签进行层次划分,将拆解后每一层次中需结构化处理的标签进行抽取形成导航目录;生成辅助信息窗口,当用户标签发生变动,根据后台数据库,自动生成辅助信息,该辅助信息包含历史数据部分、信息推荐部分、数据变动提醒部分。

[0019] 优选的,进行结构化数据在线编辑中,通过导航目录定位标签,根据标签类型规范用户输入数据,并通过辅助信息窗口为用户提供输入帮助;进一步地,对编辑后的数据及对应的模板使用版本进行关联存储时,提交静态网页的form表单,将form表单数据与模板使用版本关联并保存至后台数据库,用于后期历史数据查询及结构化数据往非结构化数据的转换。

[0020] 本发明的有益效果:

[0021] 本发明解决了非结构化向结构化转换时,关注点的问题,目前非结构化转换为结构化数据时,都是从非结构化的东西,全量的抽取,不准备,没有关注点;本发明提出标签的概念,不仅解决了非结构化向结构化转换的问题,而且解决了一般转换过程中,抽出来的数据,无重点,无语义,混乱的问题,根据标签抽出来的数据,语义明确;本发明把现在的结构化只能本地处理的功能,改为在线网页中处理,通过转换功能,把现有的非结构化数据(word)转换为结构化的数据(html),从而实现在线的编辑;结合标签类型,对标签的类型进行定义,实现标签的填写的规范化,并赋予了标签在系统中管理内涵;查询统计方便,数据保存到数据库后,方便查询统计;可以实现向非结构化数据转换,可实现把现有的结构化数据方便的输出为pdf,word等;使用方便,提供导航目录功能,方便结构化数据的定位,提供辅助信息窗口,方便实时为用户提供帮助和甚辅助的信息;结构化处理后的数据,大大减少信息冗余,便于数据分析处理,提高数据后期处理的效率,具有较好的实际应用价值。

附图说明：

[0022] 图1为本发明的方法流程图；

[0023] 图2为实施例中业务框架示意图；

[0024] 图3为实施例中实现原理示意图。

具体实施方式：

[0025] 下面结合附图和技术方案对本发明作进一步清楚、完整的说明，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其它实施例，都属于本发明保护的范围。

[0026] 针对现有供应链系统中非结构化数据转换效率低、无法有效控制及不利于整体循环等情形，本发明实施例一，参见图1所示，提供一种非结构化数据与结构化数据相互转换处理方法，包含如下内容：

[0027] 101、根据标签实现功能进行标签归类处理；

[0028] 102、针对归类后的标签进行字段属性和管理属性的处理，其中，字段属性处理包含标签创建、标签修改，管理属性处理包含根据业务情况进行管理特征定义；

[0029] 103、将待结构化数据与处理后的标签进行结合，创建使用版本模板；

[0030] 104、依据使用版本模板将用户提交的非结构化数据转换为静态网页内容，并根据标签类型将标签转换为编辑控件，提交的数据转换为结构化数据；

[0031] 105、选择使用版本模板进行结构化数据在线编辑；

[0032] 106、对编辑后的数据及对应的模板使用版本进行关联存储。

[0033] 通过对非结构化数据文件的标签定义、模板化处理及非结构化数据文件的生成与转换存储，以结构化数据标签的方式，生成非结构化数据文件，并实现对非结构化数据与结构化数据之间的相互转换，节省人力投入成本，提高效率，便于后续数据分析处理。

[0034] 实施例二，一种非结构化数据与结构化数据相互转换处理方法，参见图2和3所示，包含如下内容：

[0035] 一、根据标签实现功能进行标签归类处理；

[0036] 标签归类处理包含：根据标签实现功能归类为对应的标签类型中，标签类型包含：字符型标签、数值型标签、日期型标签、引用型标签和选择型标签。字段属性处理包含依据信息抽取原则创建标签，及对标签类型及该类型的相关属性信息的修改，其中，该类型的相关属性信息至少包含标签长度和标签默认值。优选的，依据信息抽取原则创建标签中的信息抽取原则包含如下内容：信息内容存在变动趋势，信息内容中存在用于数据分析的关键字，信息内容中存在有标识数据特征的内容。

[0037] 二、针对归类后的标签进行字段属性和管理属性的处理，其中，字段属性处理包含标签创建、标签修改，管理属性处理包含根据业务情况进行管理特征定义。

[0038] 管理属性处理包含根据业务情况进行管理特征定义，对定义标签赋予管理特质分类。管理特质分类的分类内容如下：通用性特质属性和管理策略性特质属性，其中，管理策略性特质属性包含采购策略、风险管控策略、交付策略、质量策略和评估策略。

[0039] 三、将待结构化数据与处理后的标签进行结合，创建使用版本模板。

[0040] 将待结构化数据与处理后的标签进行结合来创建使用版本模板，包含内容如下：使用WORD作为模板编译载体，在待结构化数据中需要结构化处理的地方插入处理后的标签，并记录标签所在待结构化数据中的位置，创建使用版本模板，并存储至后台数据库。

[0041] 优选的，创建的使用版本模板，每次存储至后台数据库均自动生成一个新的模板版本，用户使用时自动推荐最新版本或提供多个版本供用户选择。

[0042] 四、依据使用版本模板将用户提交的非结构化数据转换为静态网页内容，并根据标签类型将标签转换为编辑控件，提交的数据转换为结构化数据。

[0043] 提交的数据转换为结构化数据，包含如下内容：读取提交数据中的各章节内容，将非常标签转换为静态数据，根据标签类型将字符型标签转换为文本输入框，将选择型标签转换为下拉列表，将日期型标签转换为日期控件；并将需要结构化的模板文件按章节、段落、条目进行拆解，将标签进行层次划分，将拆解后每一层次中需结构化处理的标签进行抽取形成导航目录；生成辅助信息窗口，当用户标签发生变动，根据后台数据库，自动生成辅助信息，该辅助信息包含历史数据部分、信息推荐部分、数据变动提醒部分。

[0044] 五、选择使用版本模板进行结构化数据在线编辑。

[0045] 进行结构化数据在线编辑中，通过导航目录定位标签，根据标签类型规范用户输入数据，并通过辅助信息窗口为用户提供输入帮助。

[0046] 六、对编辑后的数据及对应的模板使用版本进行关联存储。

[0047] 对编辑后的数据及对应的模板使用版本进行关联存储时，提交静态网页的form表单，将form表单数据与模板使用版本关联并保存至后台数据库，用于后期历史数据查询及结构化数据往非结构化数据的转换。

[0048] 从非结构化数据将关注的信息抽取为标签，转换到其它载体中，从而实现数据结构化处理的方法。在对非结构化数据文件中的关键条款进行标签化处理，并自动化生成、转换、存储非结构化数据文件，为进一步进行数据价值挖掘分析奠定基础。

[0049] 根据常用的标签实现的功能进行归类，常用类型如下：字符型，主要属性有，最大长度，默认值。数值型，主要属性有：最大长度，是否允许小数，默认值。日期型，主要属性有：日期选择的区间。引用型，有时页面上几个标签的值应该保持一致，主要属性有：引用标签名。选择型，主要属性：选择的备选值。

[0050] 标签管理主要实现对标签的创建，修改等。包括标签的类型(字符型，数值型，日期型等)及该类型的相关属性信息(如长度，默认值等)。一般把信息抽取为标签定义的原则是：将来会变动的信息；对文档分析比较关键的信息；可以标识出文档特征的信息。标签的管理属性：标签的管理属性是对所定义的标签赋予管理的特质分类，具体的管理特质可以按照业务实际情况进行灵活定义：通用性特质属性和管理策略性特质属性，其中，管理策略性特质属性比如：采购策略、风险管控策略、交付策略、质量策略、评估策略等

[0051] 模板管理与标签文本位控制主要实现把需要结构化的数据和定义的标签进行结合和对模板的版本控制。使用大家比较熟悉的word作为模板编译的载体。在非结构化文件中，在需要结构化处理的地方插入我们已定义的标签，并记录标签所在文档的具体位置，也即标注出需要结构化转化的地方，从而创建一个模板的使用版本，并可以提交保存，存储到后台数据库。每次保存系统会自动生成一个新的版本，用户使用时自动推荐最新版本，但也

可以选择原有老版本。

[0052] 用户提交后，会触发转换功能，过程如下：将非常标签内容转换为网页的静态部分，读取各章节的内容，根据模板格式如：章节( )，段落(<p>)，文本()，换行(
)字体字号(style＝'font:宋体';font-size:xx')，等转换为HTML网页；把相应的标签根据类型转换为相应的网页编辑控件(如：文本框，下拉列表，日期控件，复选框等)，如字符型的标签转换为输入框<input type＝'text'，选择型的转换为下拉列表<select，日期型的转换为日期控件；自动生成左侧的标签导航，一般处理方法为：对需要结构化的模板文件按章节、段落、条目进行拆解，从而把标签进行层次划分，把拆解后每一层次中需结构化处理的标签抽取出来形成一个导航目录；根据当前编辑的标签实现帮助，推荐等功能，在右侧生成一个信息窗口，当用户的标签发生变成后，根据后台的数据，自动生成相应的帮助信息，历史数据，信息推荐等。

[0053] 使用结构化在线编辑，用户选择相应的模板类型，即可在网页中编辑现有的内容。通过左边的导航来定位标签；通过标签的类型，来规范用户的输入；通过右侧的信息窗口为用户提供帮助和一些辅助信息。

[0054] 结构化存储，用户在网页中编辑完信息后，单击“保存”。提交时其实是提交一个HTML的form表单，后台将提交的数据保存到数据库。将表单数据和关联的模板编号一起保存到后台数据库。模板和数据的隔离存储既实现了数据的结构化同时又节省空间。方便历史数据的查询。方便将现有的结构化数据的转换为任意的非结构化数据，如pdf,word等。

[0055] 本领域内的技术人员应明白，本发明的实施例可提供为方法、系统、或计算机程序产品。因此，本发明可采用完全硬件实施例、完全软件实施例、或结合软硬件方面的实施例的形式。而且，本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0056] 本发明是参照根据本发明实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其它可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0057] 这些计算机程序指令也可存储在能引导计算机或其它可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0058] 这些计算机程序指令也可装载到计算机或其它可编程数据处理设备上，使得在计算机或其它可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其它可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0059] 对所公开的实施例的上述说明，使本领域专业技术人员能够实现或使用本申请。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的，本文中所定义的

一般原理可以在不脱离本申请的精神或范围的情况下，在其它实施例中实现。因此，本申请将不会被限制于本文所示的这些实施例，而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

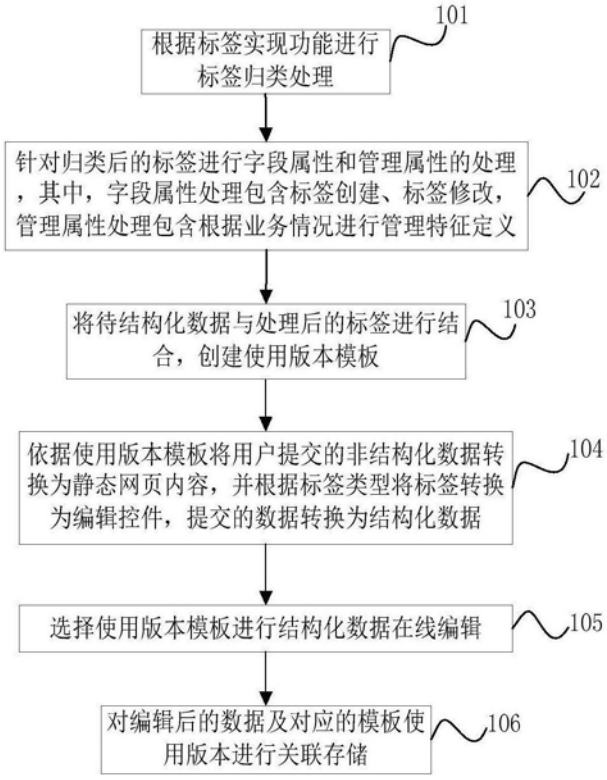

101
根据标签实现功能进行标签归类处理
针对归类后的标签进行字段属性和管理属性的处理，其中，字段属性处理包含标签创建、标签修改，管理属性处理包含根据业务情况进行管理特征定义
102
将待结构化数据与处理后的标签进行结合，创建使用版本模板
103
依据使用版本模板将用户提交的非结构化数据转换为静态网页内容，并根据标签类型将标签转换为编辑控件，提交的数据转换为结构化数据
104
选择使用版本模板进行结构化数据在线编辑
105
对编辑后的数据及对应的模板使用版本进行关联存储
106

图1

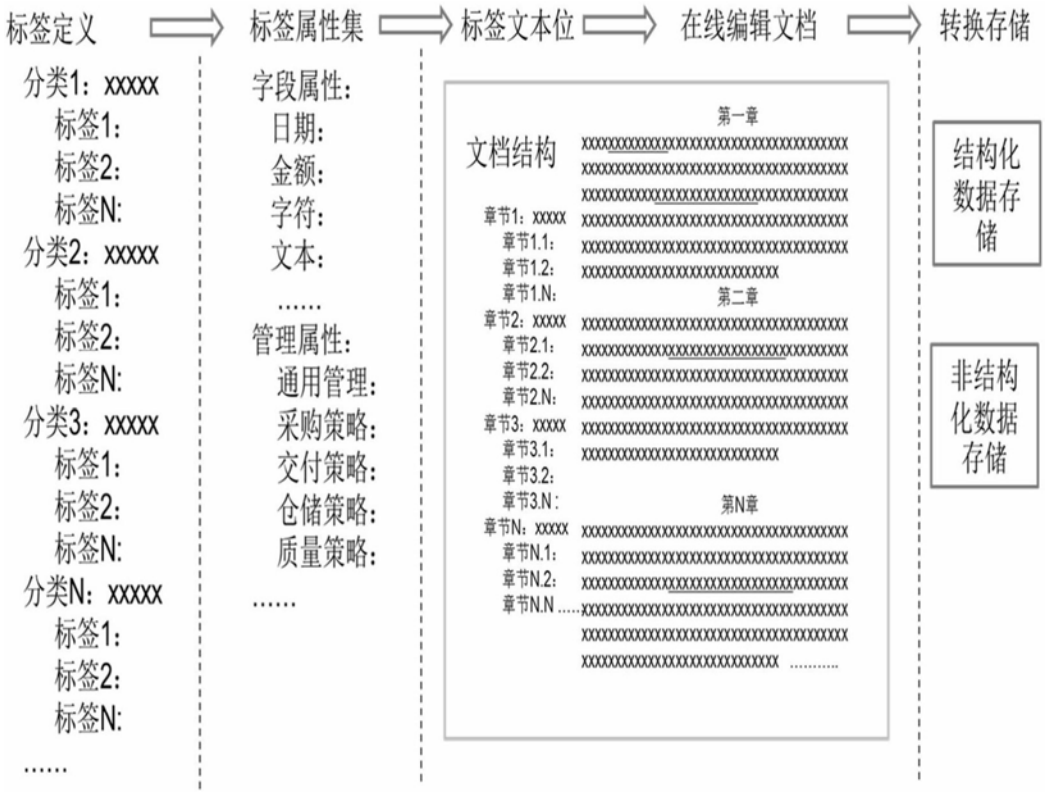

标签定义
标签属性集
标签文本位
在线编辑文档
转换存储
分类1：xxxxx
标签1：
标签2：
标签N：
分类2：xxxxx
标签1：
标签2：
标签N：
分类3：xxxxx
标签1：
标签2：
标签N：
分类N：xxxxx
标签1：
标签2：
标签N：
.....
字段属性：
日期：
金额：
字符：
文本：
.....
管理属性：
通用管理：
采购策略：
交付策略：
仓储策略：
质量策略：
.....
文档结构
第一章
章节1：xxxxx
章节1.1：
章节1.2：
章节1.N：
第二章
章节2：xxxxx
章节2.1：
章节2.2：
章节2.N：
章节3：xxxxx
章节3.1：
章节3.2：
章节3.N：
第N章
章节N：xxxxx
章节N.1：
章节N.2：
章节N.N .....
结构化数据存储
非结构化数据存储

图2

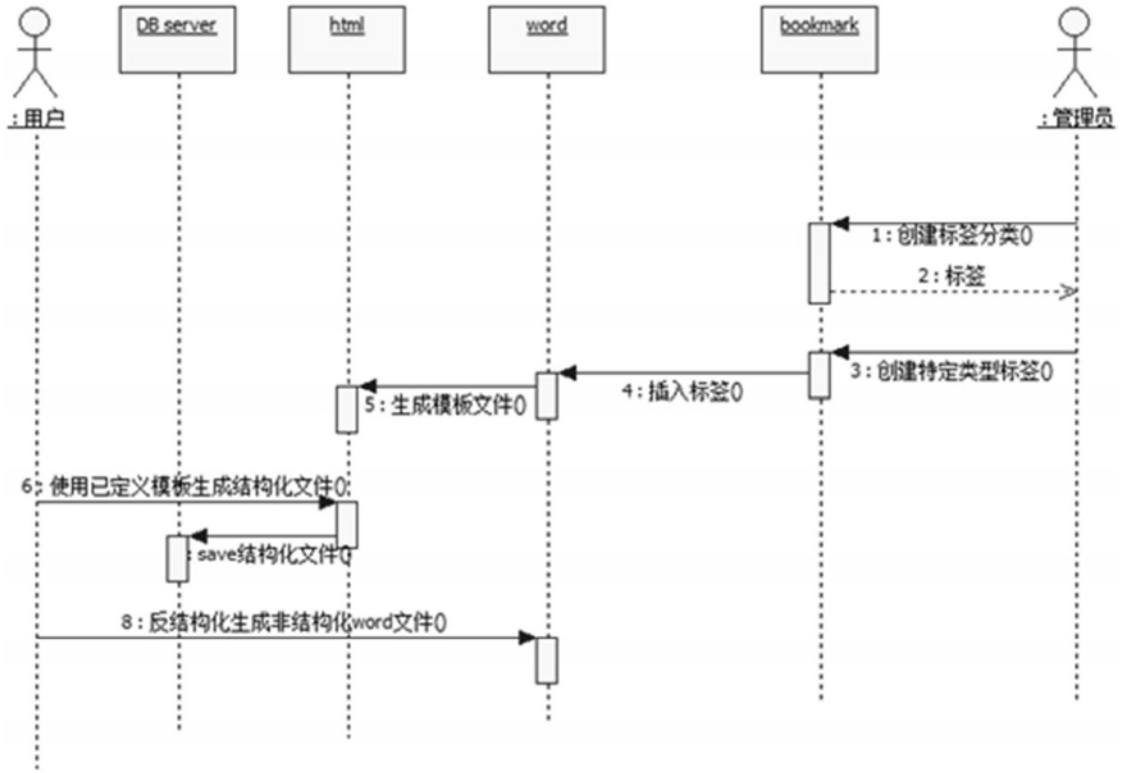

:用户
DB server
html
word
bookmark
:管理员
1:创建标签分类0
2:标签
3:创建特定类型标签0
4:插入标签0
5:生成模板文件0
6:使用已定义模板生成结构化文件0
save结构化文件0
8:反结构化生成非结构化word文件0

图3