(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 112861821 A
(43)申请公布日 2021.05.28

(21)申请号 202110364526.8

(22)申请日 2021.04.06

(71)申请人 刘羽
地址 230022 安徽省合肥市包河区宿松路金安花园7栋104

(72)发明人 刘羽 王贺 王辉 李姜晖 刘永付俐

(51)Int.Cl.
G06K 9/00(2006.01)
G06F 16/33(2019.01)

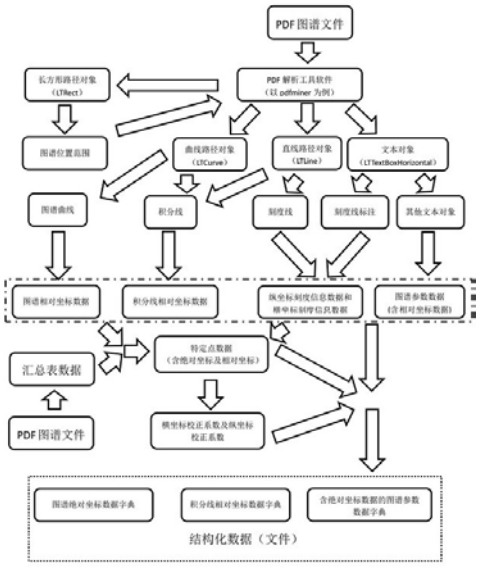
权利要求书5页 说明书5页 附图8页

(54)发明名称

基于PDF文件解析的图谱数据还原方法

(57)摘要

本发明公开了基于PDF文件解析的图谱数据还原方法，该方法包括：通过解析文件获得图谱位置范围；依据图谱中各类相关对象的位置属性，识别分类不同功能的数据及相对坐标；通过数据间的相互关系，获得图谱中的特定点的相对坐标和绝对坐标，并进一步获得相对坐标和绝对坐标所对应的横坐标校正系数和纵坐标校正系数；通过对所获得的相对坐标数据的换算，得到构建图谱的绝对坐标数据，从而实现对PDF图谱数据的还原。将PDF格式的图谱内容转换为反映图谱特征的、数值与原始数据接近、可操作可检索的数据，使图谱数据的使用脱离原专用系统、工作站、工作程序的限制，提升图谱数据的交换、查询、比对的便利性，方便进行数据的统一管理。



CN 112861821 A

1.基于PDF文件解析的图谱数据还原方法,其特征在于,所述处理方法包括以下步骤:

步骤一:使用计算机软件工具对PDF文件进行解析,依次获取图谱报告中存在的的路径对象(Path Object),读取并对路径对象的位置信息进行分析,寻找并确定图谱的位置范围;

步骤二:获取图谱报告中存在的的路径对象(Path Object),根据情况,并进一步识别为图谱曲线、坐标轴框架、积分线、积分线标记,将图谱曲线中的图谱坐标组合生成图谱相对坐标数据,将积分线的图谱坐标生成积分线相对坐标数据;

步骤三:在图谱所处的位置范围内,获取文件图谱的文本对象的文本信息和图谱坐标,识别为纵坐标刻度线标注,横坐标刻度线标注,及图谱其他文本对象;进一步的,对图谱其他文本对象根据对象位置判断对象功能,将文本对象的文本及对象相对坐标匹配,生成图谱参数数据;

步骤四:在图谱所处的位置范围内,获取文件图谱的的路径对象(Path Object),根据情况,并进一步识别为纵坐标刻度线,横坐标刻度线,积分线、积分线标记,将积分线的图谱坐标生成积分线相对坐标数据;

步骤五:根据步骤三和步骤四的结果分别形成纵坐标刻度信息数据和横坐标刻度信息数据;

步骤六:解析文件中数据汇总表的文本信息及文本的位置信息生成汇总表数据;

步骤七:读取或计算文件图谱中特定点的绝对坐标及对应的相对坐标;

步骤八:根据已知特定点组合的绝对坐标数据及相对坐标数据,计算图谱的横坐标校正系数与纵坐标校正系数;

步骤九:采用已知绝对坐标及相对坐标的特定点之一作为参照点,根据步骤八得到的横坐标校正系数与纵坐标校正系数,将步骤二得到的图谱相对坐标数据逐一进行换算,得到绝对坐标,生成图谱绝对坐标数据字典;

步骤十:采用已知的绝对坐标及相对坐标的特定点之一作为参照点,根据步骤八得到的横坐标与纵坐标的校正系数,将步骤三得到的图谱参数数据中的相对坐标数据逐一进行换算,得到绝对坐标,生成包含有绝对坐标的图谱参数数据字典;

步骤十一:采用已知的绝对坐标及相对坐标的特定点之一作为参照点,根据步骤八得到的横坐标与纵坐标的校正系数,将步骤二或步骤四得到的积分线相对坐标数据中的相对坐标逐一进行换算,得到绝对坐标,生成积分线绝对坐标数据字典;

步骤十二:将图谱绝对坐标数据字典,包含有绝对坐标的图谱参数数据字典,积分线绝对坐标数据字典合并打包生成结构化数据备用。

2.根据权利要求1所述基于PDF文件解析的图谱数据还原方法,其特征在于,所述PDF文件为由生成数据的仪器设备的工作站或专用软件的报告程序通过调用PDF虚拟打印功能直接生成的具备规范内部结构的可被程序解析其中所包含的各类对象位置和内容的PDF文件。

3.根据权利要求1所述基于PDF文件解析的图谱数据还原方法,其特征在于,所述步骤一到步骤四中涉及的计算机工具软件包括而不仅限于C、C#、Python、Java、Visual Studio等计算机语言所创建的可对PDF文件内容进行解析的工具软件,进一步要求为可实现对PDF中各类对象的位置信息进行解析的工具软件,更进一步要求为Python语言中的Pdfminer或

Pdfminer3K。

4.根据权利要求1所述基于PDF文件解析的图谱数据还原方法,其特征在于,所述步骤一到步骤四中的图谱坐标是基于PDF页面位置进行定位的相对坐标,为符合Pdfminer工具软件所定义对象属性中的x0、y0、x1、y1或pts数据包中的内容。

5.根据权利要求1所述基于PDF文件解析的图谱数据还原方法,其特征在于,所述步骤一中路径对象(Path Object)为符合Pdfminer工具软件所定义的LTRect对象。

6.根据权利要求1所述基于PDF文件解析的图谱数据还原方法,其特征在于,所述步骤一中图谱的准确位置范围系指由满足x1-x0最大且y1-y0最大的LTRect对象的x0,y0,x1,y1定义的矩形范围。

7.根据权利要求1所述基于PDF文件解析的图谱数据还原方法,其特征在于,所述步骤二中路径对象(Path Object)为符合Pdfminer工具软件所定义的LTCurve对象。

8.根据权利要求1所述基于PDF文件解析的图谱数据还原方法,其特征在于,所述步骤二中图谱坐标为LTCurve对象属性中的pts数据包内容。

9.根据权利要求1所述基于PDF文件解析的图谱数据还原方法,其特征在于,所述步骤二中将LTCurve对象识别为图谱曲线,坐标轴框架、积分线、积分线标记的方法具体为:

根据LTCurve对象属性中的pts数据包中坐标数量和坐标差值进行判断:

pts数据包中坐标数量为2,且所述两点之间的纵坐标之差与横坐标之差与其他包含坐标数量为2的pts数据包中的两点之间的纵坐标之差与横坐标之差均不同(偏差大于5%),则判

断为积分线;

pts数据包中坐标数量大于2且小于5,且数据包中相邻的两个坐标的纵坐标及横坐标相等,出现重叠,则判断为积分线;

pts数据包中坐标数量为2,且所述两点之间的纵坐标之差与横坐标之差与其他包含坐标数量为2的pts数据包中的两点之间的纵坐标之差与横坐标之差相同(偏差小于5%),则判断为积分线标记;

pts数据包中坐标数量大于4,且数据包中第一个坐标与最后一个坐标两点的纵坐标和横坐标不相等,则判断为图谱曲线;

pts数据包中坐标数量等于5,且数据包中第一个坐标与最后一个坐标两点的纵坐标和横坐标相等,则判断为坐标轴框架。

10.根据权利要求1所述基于PDF文件解析的图谱数据还原方法,其特征在于,所述步骤二中组合生成图谱相对坐标数据的方法具体为:

对指定范围内的有序排列的绘制图谱曲线的一个或多个LTCurve对象进行逐一历遍,读取LTCurve对象属性中的pts数据包内容,添加到指定的数据列表生成组合pts数据列表。

11.根据权利要求1所述基于PDF文件解析的图谱数据还原方法,其特征在于,所述步骤三中文件图谱的文本对象为符合Pdfminer工具软件所定义的LTTextBox特征的数据对象;上述数据对象进一步优选为符合Pdfminer工具软件所定义的 LTTextBoxHorizontal特征的数据对象。

12.根据权利要求1所述基于PDF文件解析的图谱数据还原方法,其特征在于,所述步骤三中文本对象识别为纵坐标刻度线标注,横坐标刻度线标注,及图谱其他文本对象的方法

具体为：

文本对象之间位置属性中的x1相等，且对象的文本为文本型数字，则判断为纵坐标刻度线标注；

文本对象之间位置属性中的y0或y1相等，且对象的文本为文本型数字，则判断为横坐标刻度线标注；

文本对象不满足上述两种情况，则判断为图谱其他文本对象。

13.根据权利要求1所述基于PDF文件解析的图谱数据还原方法，其特征在于，所述步骤四中路径对象(Path Object)为符合Pdfminer工具软件所定义的LTLine对象。

14.根据权利要求1所述基于PDF文件解析的图谱数据还原方法，其特征在于，所述步骤四中识别为纵坐标刻度线，横坐标刻度线，积分线、积分线标记的具体方法为：

路径对象之间位置属性中的x1相等，且路径对象的y0=y1，则判断为纵坐标刻度线，其在纵坐标轴上的相对坐标为(x1,y0)或(x1,y1)；

路径对象之间位置属性中的y1相等，且路径对象的x0=x1，则判断为横坐标刻度线，其在横坐标轴上的相对坐标为(x0,y1)或(x1,y1)；

不满足上述两种情况的路径对象则判断为积分线或积分线标记，进一步的，路径对象的x0=x1则判断为积分线标记，路径对象的x0≠x1则判断为积分线。

15.根据权利要求1所述基于PDF文件解析的图谱数据还原方法，其特征在于，所述步骤五中刻度信息数据中的元素为刻度线标注的文本与对应的纵/横坐标轴上位置的相对坐标的配对。

16.根据权利要求1所述基于PDF文件解析的图谱数据还原方法，其特征在于，所述步骤五中生成纵坐标刻度信息数据和横坐标刻度信息数据的具体方法包括：

方法一：

通过对所保存的刻度线标注和对应的刻度线的历遍，循环比对刻度线标注的位置信息和对应的刻度线的位置信息：

计算|(Textbox$_z$.y0+ Textbox$_z$.y1)/2-LTline$_z$(i).y0| 或|(Textbox$_z$.y0+ Textbox$_z$.y1)/2-LTline$_z$(i).y1|的最小值，

满足上述条件则进行纵坐标刻度线标注的文本和对应的纵坐标刻度线的(x1,y0)或(x1,y1)的匹配；

计算|(Textbox$_h$.x0+Textbox$_h$.x1)/2-LTline$_h$(i).x0|或|(Textbox$_h$.x0+ Textbox$_h$.x1)/2-LTline$_h$(i).x1|的最小值，

满足上述条件则进行横坐标刻度线标注的文本和对应的横坐标刻度线的(x0,y1)或(x1,y1)的匹配；

其中：Textbox$_z$.y0为逐一历遍的纵坐标刻度线标注的y0，Textbox$_z$.y1为逐一历遍的纵坐标刻度线标注的y1，

LTline$_z$(i).y0为所进行循环比对的纵坐标刻度线的y0，LTline$_z$(i).y1为所进行循环比对的纵坐标刻度线的y1，

Textbox$_h$.x0为逐一历遍的横坐标刻度线标注的x0，Textbox$_h$.x1为逐一历遍的横坐标刻度线标注的x1，

LTline$_h$(i).x0为所进行循环比对的横坐标刻度线的x0，LTline$_h$(i).x1为所进行循环

比对的横坐标刻度线的x1，

将上述一一匹配的标注文本与对应的坐标点相对位置保存为列表元素；

方法二：

通过对所保存的纵坐标刻度线标注及横坐标刻度线标注分别进行历遍，实现纵坐标标注的文本与$(x_z,(y0 + y1)/2)$进行匹配；横坐标标注的文本与$((x0 + x1)/2,y_h)$进行匹配；

其中：x0、y0、x1、y1为当前文本对象的位置属性，

x_z为纵坐标轴的横坐标，由权利要求14中所述的纵坐标刻度线的x1得到，

y_h为横坐标轴的纵坐标，由权利要求14中所述的横坐标刻度线的y1得到，

将上述一一匹配的标注文本与对应点的相对坐标保存为列表元素；

优选采用方法一实现生成纵坐标刻度信息数据和横坐标刻度信息数据。

17.根据权利要求1所述基于PDF文件解析的图谱数据还原方法，其特征在于，所述步骤七中特定点绝对坐标为具有实验意义的，可以在图谱中对特定点进行定位的数据，和相对坐标为一一对应的关系，是对所述特定点基于不同参照系的描述。

18.根据权利要求1所述基于PDF文件解析的图谱数据还原方法，其特征在于，所述步骤七中特定点包括在图谱数据汇总表中有完整记录的点；在图谱数据汇总表中没有完整记录但可以依据图谱进行推算的点；位于坐标轴上有明确刻度标记及标注的可进行推算的点。

19.根据权利要求1所述基于PDF文件解析的图谱数据还原方法，其特征在于，所述步骤七中获得文件图谱中特定点的绝对坐标及对应的相对坐标的方法包括：

方法一：比对所述步骤五的纵坐标刻度信息数据和横坐标刻度信息数据其中的相对坐标位置信息，识别纵坐标轴与横坐标轴相交点；依据纵坐标轴与横坐标轴相交点的标记文本及相对坐标，通过读取或计算获得纵坐标轴与横坐标轴相交点的绝对坐标；根据相交点的绝对坐标将所述步骤五中刻度信息数据中的列表元素转化为绝对坐标与相对坐标相匹配的列表元素；

方法二：读取步骤六中所述汇总表数据中的数据，筛选获得绝对坐标；对步骤二中所述图谱相对坐标数据根据其中的相对坐标元素的纵坐标进行排序、筛选，获得对应相对坐标；形成绝对坐标与相对坐标相匹配的数据；

方法三：根据情况，对步骤二中所述图谱相对坐标数据根据其中的相对坐标元素的横坐标进行排序、筛选，选择横坐标最小的相对坐标元素作为相对坐标；选择(0,0)作为绝对坐标；形成绝对坐标与相对坐标相匹配的数据；

方法四：分析图谱获得特定点的相对坐标数据，并通过人工识别，手动介入的方法进行录入为绝对坐标与相对坐标相匹配的数据。

20.根据权利要求1所述基于PDF文件解析的图谱数据还原方法，其特征在于，所述步骤八中特定点组合至少为2个，进一步要求当特定点数量为2时，所选特定点的横坐标与纵坐标均不相同；当特定点组合的数量大于2时，至少其中1点的横坐标与其他点不同且至少其中1点的纵坐标与其他点不同；优选参与计算的两个特定点之间满足计算纵坐标校正系数时纵坐标差值最大或计算横坐标校正系数时横坐标差值最大，最优选参与计算的两个特定点之间满足纵坐标差值最大及横坐标差值最大。

21.根据权利要求1所述基于PDF文件解析的图谱数据还原方法，其特征在于，所述步骤八中计算图谱的横坐标校正系数与纵坐标校正系数的方法具体为：

$T_X=(X_{S1}-X_{S2})/(X_{P1}-X_{P2})$；

$T_Y=(Y_{S1}-Y_{S2})/(Y_{P1}-Y_{P2})$；

所述T_X为横坐标校正系数，T_Y为纵坐标校正系数；

所述(X_{S1},Y_{S1})与(X_{S2},Y_{S2})分别为两个所选特定点的绝对坐标，X_{S1}与X_{S2}为横坐标，Y_{S1}与Y_{S2}为纵坐标；

所述(X_{P1},Y_{P1})与(X_{P2},Y_{P2})分别为两个所选特定点的相对坐标，X_{P1}与X_{P2}为横坐标，Y_{P1}与Y_{P2}为纵坐标；

进一步要求，所述横坐标校正系数与纵坐标校正系数为符合计算机定义的整数型或浮点型数值，优选为浮点型的单精度或者双精度型数值，更优选为双精度型数值。

22.根据权利要求1所述基于PDF文件解析的图谱数据还原方法，其特征在于，所述步骤九、步骤十、步骤十一换算得到绝对坐标的方法具体为：

$X=T_{X*}(X_p-X_{P1})+X_{S1}$；

$Y=T_{Y*}(Y_p-Y_{P1})+Y_{S1}$；

所述(X,Y)为所选目标点的绝对坐标，X为横坐标，Y为纵坐标；

所述(X_{S1},Y_{S1})为所选特定点的绝对坐标，X_{S1}为横坐标，Y_{S1}为纵坐标；

所述(X_{P1},Y_{P1})为所选特定点的相对坐标，X_{P1}为横坐标，Y_{P1}为纵坐标；

所述(X_p,Y_p)为目标点的相对坐标，X_p为横坐标，Y_p为纵坐标；

所述T_X为权利要求21所述横坐标校正系数，T_Y为权利要求21所述纵坐标校正系数。

23.根据权利要求1所述基于PDF文件解析的图谱数据还原方法，其特征在于，所述步骤十二中的结构化数据包括而不仅限于XML、Json等的符合计算机领域定义并可被相应的规则解析的便于传递、保存的特定结构文件；进一步要求上述文件可在设定解析策略后被数据处理绘图软件，包括而不仅限于Origin、EXCEL、Matlab等工具软件，识别、解析并绘制为矢量图谱。

基于PDF文件解析的图谱数据还原方法

技术领域

[0001] 本发明涉及基于PDF文件解析的图谱数据还原方法,属于文件数据解析领域。

背景技术

[0002] 图谱作为科学研究的重要手段,在分析实验中的作用巨大。图谱通常以包含纵坐标与横坐标的散点图形式出现,通常呈现连续性变化,其纵坐标与横坐标具有特征性的相关性。例如液相图谱:洗脱物质的吸收值与洗脱时间的对应关系;紫外分光光度的扫描图谱:样品吸光值与步进变化的波长之间的对应关系;晶体的X衍射:步进变化的衍射角2θ与强度标值I之间的对应关系等等。

[0003] 这种特征性的相关性直接或间接的反映出了被研究对象特定的物理化学性质,因此图谱解析作为现代实验室的主要研究手段异常重要。

[0004] 现代分析仪器通常采用安装于PC机、工作站或者网络服务器的专用工作软件进行数据抓取和分析,仪器与专用软件之间属于一对一的匹配,具有专属性;综合型实验室因为研究目标、研究手段、设备更新等原因,又存在实际应用场景中的多样性。例如实验室会配备多种研究设备采用不同的方法(如液相色谱、气相色谱、质谱、核磁共振、热分析等)对同一研究目标从多方面进行研究,且由于商业竞争、仪器更新迭代、软件版本升级等原因,相同原理的检测设备也会出现不同品牌设备并存或是同品牌新老设备并存的现象。

[0005] 由于上述的种种原因,现代实验室中的图谱数据文件格式五花八门。对综合型实验室,多类型图谱综合性的数据管理及报告生成并没有较好的解决方案。

[0006] 现有LIMS系统(实验室信息管理系统)及SDMS(科学数据管理系统)或是属于第三方开发系统,因为缺乏对仪器硬件底层技术及图谱数据原始文件的数据结构的了解,易造成数据采集的不完整;或是因为硬件供应商自行开发的控制及管理系统,其专属性太强,无法实现跨品牌、跨硬件类型的数据采集,会因为网络内部的仪器类型,仪器型号,厂商的不同而产生兼容性问题。

[0007] 综合性实验室需要实现数据采集的专属性与仪器系统兼容性的平衡。

[0008] PDF(Portable Document Format,便携式文档格式)是一种独立于硬件、操作系统、应用程序的电子文档。上述的图谱数据都可以通过虚拟打印的方式实现图谱报告的输出,生成图谱的PDF文件。PDF文件因其独特的优点成了事实上的实验室通用报告文本。通过对PDF文件的解析和数据还原,可以实现对综合性实验室电子数据的管理。目前所知对PDF文件的解析通常只是针对文件中的字符型数据按照规则进行解析,对以图形式展现的图谱并没有较好的解析,这使得所得到的报告数据并不全面。

发明内容

[0009] 发明目的:本发明针对综合型实验室存在的问题,提出一种图谱数据还原方法,将PDF格式的图谱报告还原为包含图谱绝对坐标数据、积分线及图谱参数的集合,并打包生成可进行数据传输XML、Json等的特定数据结构文件。上述文件可在设定解析策略后被通用的

Origin、EXCEL、Matlab等数据处理软件识别、解析并绘制为矢量图谱。并可以在上述软件中进行进一步的数据标记、面积积分操作。

[0010] 技术方案:多数电子仪器的基本原理是通过传感器接收特定信号,如特定波长,温度,压力等,转换为电信号,并通过数模转换成为可被计算机记录处理的数字信号,该数字信号与对应的参数如保留时间,转角步进,变化波长等一一匹配,形成以二维数组列表形式的原始数据记录;上述记录通过特定软件/算法进行处理、计算、压缩最终生成图谱报告。

[0011] 生成图谱报告的“数据散点”是由安装于PC机、工作站或者网络服务器的专用工作软件根据报告输出的分辨率采用特定的算法所生成的压缩数据,这种压缩没有将所记录的所有数据点一一体现,但在所输出的特定分辨率下并没有改变图谱的特征性,图谱依旧可以被识别并依此进行判断。

[0012] 实验室图谱一般都是以直线和曲线来描述的图形,通过以二维数组(X,Y)为坐标生成“数据散点”,其中的横坐标X与纵坐标Y为具有相关性的特征性数据。上述二维数组中的X,Y来源于具有实验研究意义的数据,其坐标为固定的“绝对坐标”,不会因为参照物的不同而不同。

[0013] PDF是从PS语言(Postscript语言,即页面描述语言)发展而来的一种结构化的文档格式。通过页面描述指令对指定区域进行着色绘制页面,PDF支持5种类型的的位图对象(Graphics Object)包括:路径对象(Path Object),文本对象(Text Object),图像对象(Image Object)和外部对象(External Object:XObject)。

[0014] PDF文件中的图谱通常是采用路径对象(Path Object)形式实现,其中直线(Line),曲线(Curve),长方形(Rectangle)都属于路径对象(Path Object)。

[0015] PDF文件在绘制图谱时,会根据图谱的“数据散点”基于页面位置布局对进行处理。各“绘图散点”的坐标为在PDF页面中的位置,其坐标为“相对坐标”(相对于页面位置),会因为图谱报告排版的变化而发生变化。

[0016] 由于“绘图散点”是由“数据散点”依据坐标变换而来,并没有改变“数据散点”所反映的特征性及“数据散点”之间的相关性,因此通过“绘图散点”所绘制的图谱在视觉上不会发生变形。

[0017] 依据坐标变换的原理,在图谱文件中找到相关的数据通过设立参照点和修正系数,建立“数据散点”和“绘图散点”转换公式,就可以通过“绘图散点”找到“数据散点”。从而得到与原数据文件输出数值及效果接近、可以反映出检测物质的特征性的数据。

[0018] 随着信息技术的发展,目前已经有工具可以实现对PDF文件中各类资源的解析,更进一步的可以实现对路径对象的识别和位置确定。这就为本发明的实现创造了必要的条件。

[0019] 由于采用了上述技术方案,本发明的有益效果为:将PDF格式的图谱内容转换为反映图谱特征的、数值与原始数据接近、可操作可检索的数据;使图谱数据的使用脱离原专用系统、工作站、工作程序的限制;提升图谱数据的交换、查询、比对的便利性,方便进行数据的统一管理;可以统一实验室的报告形式,无需通过附件形式附加图谱,有助于形成更规范的报告文本;所生成的数据便于进行自动化分析,结合AI技术,可以更迅速的分析结果。

附图说明

[0020] 此处的附图用于解释具体实施例,以便于更好的理解本发明,并不构成对本发明的不当限定。

[0021] 图1为本发明的流程示意图;

图2为本发明中实施例1的目标图谱PDF页面;

图3为本发明中的相关术语示意图(以实施例1为例);

图4为本发明中实施例1中的图谱曲线的局部放大图;

图5为本发明中实施例1处理过程的示例1;

图6为本发明中实施例1处理过程的示例2;

图7为本发明中实施例1的结果展示;

图8为本发明中实施例4的目标图谱PDF页面;

图9为本发明中的相关术语示意图(以实施例4为例);

图10为本发明中实施例4处理过程的示例;

附图标记:1、图谱范围(LTRect对象)示例;2、坐标轴框架(LTCurve对象)示例;3、图谱其他文本对象(LTTextBoxHorizontal对象)示例;4、图谱曲线之一示例(LTCurve对象);5、积分线示例(LTLine对象);6、积分线标记示例(LTLine对象);7、绝对坐标示例;8、特定点示例1;9、数据汇总表;10、纵坐标轴与横坐标轴相交点;11、特定点示例2;12、特定点示例3;13、特定点示例4;14、横坐标刻度线示例(LTLine对象);15、横坐标刻度线标注(LTTextBoxHorizontal对象)示例;16、纵坐标刻度线示例(LTLine对象);17、纵坐标刻度线标注(LTTextBoxHorizontal对象)示例;18、积分线示例(LTCurve对象)。

具体实施方式

[0022] 为便于理解,实施例采用Python语言编写的Pdfminer作为PDF解析软件,采用Python语言编写的Matplotlib作为绘图软件。

[0023] 需要说明的是,因生成PDF图谱文件的报告程序不同,所调用的虚拟打印的方法不同,不同的图谱PDF文件在绘图细节上也存在差异,如是否采用对坐标轴框架、对积分线所采用的描绘方式(LTCurve对象或是LTLine对象)、积分线标记方式(以LTLine对象描述的直线型或者是以LTCurve对象描述的箭头型)。在处理方法上应进行相应的调整。基本流程见图1。

[0024] 下面结合附图描述本发明的具体实施例。

[0025] 实施例1

目标PDF页面见图2,该实施例PDF图谱存在一个由LTCurve对象绘制的坐标轴框架2以及LTLine对象绘制的积分线5。参见图3。

[0026] 1、采用软件对PDF进行解析,通过解析文件的中以PDF页面为参照物生成的路径对象(Path Object),该类路径对象在Pdfminer中定义为LTRect对象,获得计算该类对象属性中的x1-x0及y1-y0的最大值,对符合条件的LTRect对象的位置信息进行解析,获得图谱范围1。

[0027] 2、采用软件对PDF进行解析,通过解析文件的中以PDF页面为参照物生成的用于显示图谱的路径对象(Path Object),该类路径对象在Pdfminer中定义为LTCurve对象,对

LTCurve对象进行识别，区分坐标轴框架2与图谱曲线4，参见图5。解析获得图谱曲线的LTCurve对象的路径点，参见图4，该路径点信息包含在LTCurve对象的属性中的pts中，将所获的多个绘制图谱曲线的LTCurve对象的路经点的相对坐标组合生成包含目标图谱上各点的位置的图谱相对坐标数据。

[0028] 3、在图谱所处的位置范围内，分析LTTextBoxHorizontal类型的文本对象，参见图6，根据相对坐标 (x0,y0,x1,y1)进行区分，判断文本对象的功能。识别为纵坐标刻度线标注17，横坐标刻度线标注15及图谱其他文本对象3。所述实施例1的图谱其他文本对象3包括纵/横坐标轴单位，样品名称，测定参数信息。将所获的图谱其他文本及对应相对坐标按照功能分类进行保存，生成图谱参数数据。

[0029] 4、在图谱所处的位置范围内，分析LTLine类型的路径对象。根据相对坐标 (x0,y0,x1,y1)进行区分，判断路经对象的功能。识别为纵坐标刻度线16，横坐标刻度线14、图谱的积分线5及积分线标记6。将所获的路径对象相对坐标的按照功能分类进行保存。将积分线5的图谱坐标生成积分线相对坐标数据。

[0030] 5、匹配得到的坐标刻度线与刻度线标注。通过位置比较，得到纵坐标轴与横坐标轴相交点10的标注信息，分别获得纵坐标轴的横坐标(绝对坐标)和横坐标轴的纵坐标(绝对坐标)；进一步的，获得位于纵/横坐标轴上的刻度点的绝对坐标与对应的相对坐标，形成相匹配的数据。选择其中一点作为特定点11。

[0031] 6、通过分析图谱文件中的数据汇总表9中的记录，解析目标图谱中特定点之一8的绝对坐标7：横坐标为实验图谱上定义的出峰位置，纵坐标为实验图谱上定义的峰高。通过解析所生成目标图谱的各点的相对坐标，通过排序法进行比较筛选，找到相对坐标的纵坐标的最大值，所对应的坐标即为特定点之一8的相对坐标。

[0032] 7、通过图谱上已知绝对坐标及相对坐标的上述两点8和11，计算相对坐标和绝对坐标的所对应的横坐标校正系数和纵坐标校正系数。

[0033] 8、采用已知绝对坐标与对应的相对坐标的所述特定点之一11作为参照点，将上述保存数据(图谱相对坐标数据、图谱参数数据、积分线相对坐标数据)中的相对坐标通过横坐标校正系数和纵坐标校正系数进行换算，得到对应的各点绝对坐标。根据功能构建为字典进行保存。

[0034] 9、进一步的，将所得到的字典生成结构化文件如XML或Json等便于传输和转移的文件。

[0035] 10、将上述文件数据解析后导入Matplotlib软件生成解析图谱，见图7。

[0036] 实施例2：

所分析得图谱与实施例1相同，实施思路相近，不同之处在于所选择的用于计算的特定点分别为具有可识别刻度标记的位于纵坐标轴的特定点13与横坐标轴的特定点11，参见附图3，而不需要再通过读取数据汇总表的方式找到特定点的绝对坐标。

[0037] 实施例3

所分析的图谱与实施例1相同，实施思路相近，不同之处在于所选择的用于计算的特定点之一为图谱的起始点：特定点12。在类似本实施例的图谱中，起始点通常默认为原点，因此其相对坐标为图谱相对坐标数据中的第一个坐标。而该位置绝对坐标为(0,0)；另一特定点为具有可识别刻度标记的位于横坐标轴的特定点11，参见图3，而不需要再通过读

取数据汇总表的方式找到特定点的绝对坐标。

[0038] 实施例4

目标PDF页面见图8，该PDF图谱不存在由LTCurve对象绘制的坐标轴框架，其积分线18由LTCurve对象绘制，参见图9。

[0039] 处理方法与实施例1基本一致。在对LTCurve对象的处理上略有所不同，需要进行对象的识别，区分积分线18与图谱曲线4，识别的结果参见图10。

[0040] 针对本实施例的其他操作与实施例1一致。

[0041] 所述实施例仅为本发明的部分实施例，并非因此限制本发明的专利范围，在本发明的技术构思范围内，采用不同的编程语言，对技术方案进行的变换或直接间接应用于其他技术领域均在本发明的专利保护范围内。

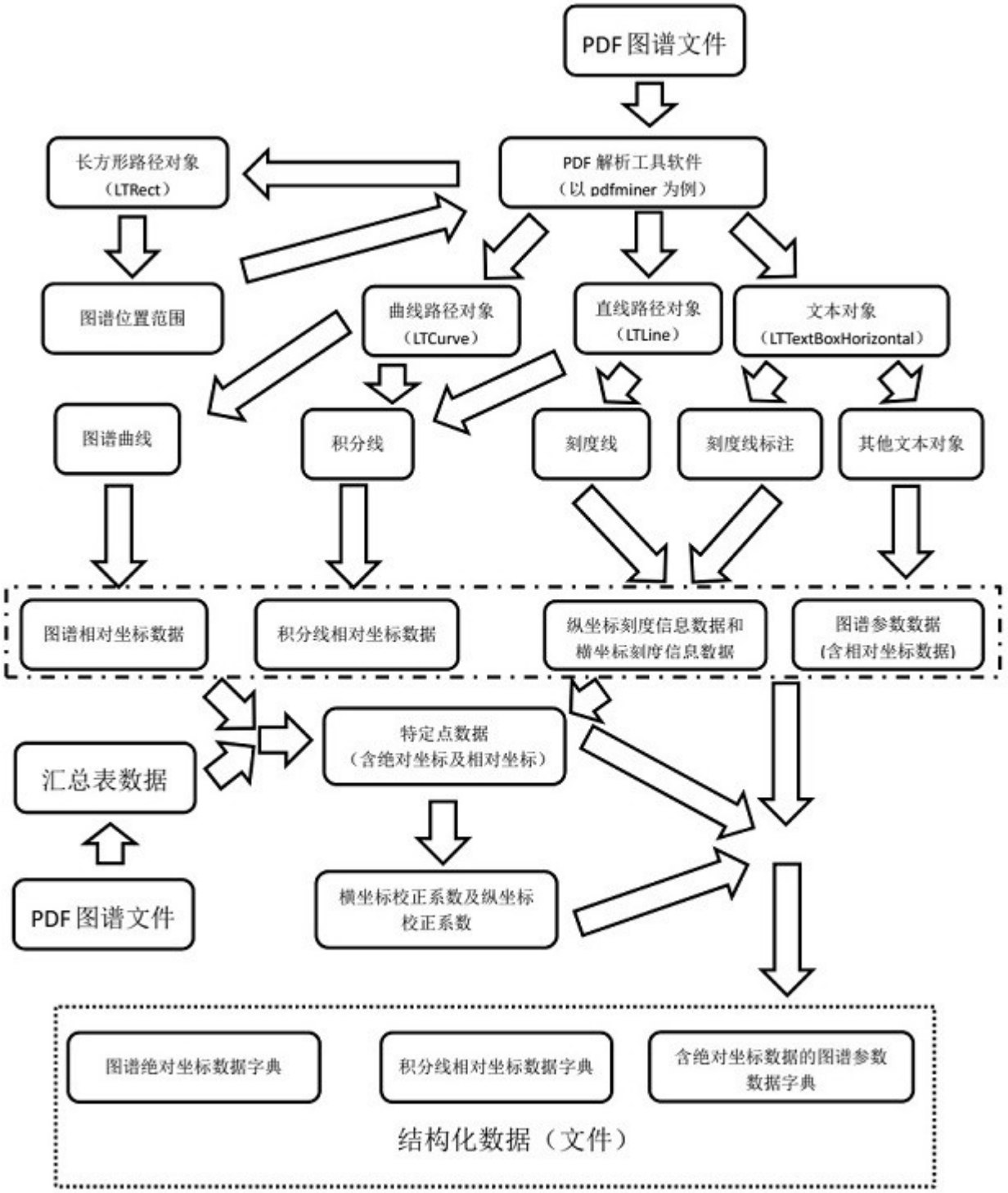

PDF图谱文件
PDF 解析工具软件
（以 pdfminer 为例）
长方形路径对象
（LTRect）
图谱位置范围
曲线路径对象
（LTCurve）
直线路径对象
（LTLine）
文本对象
（LTTextBoxHorizontal）
图谱曲线
积分线
刻度线
刻度线标注
其他文本对象
图谱相对坐标数据
积分线相对坐标数据
纵坐标刻度信息数据和
横坐标刻度信息数据
图谱参数数据
(含相对坐标数据)
特定点数据
（含绝对坐标及相对坐标）
汇总表数据
PDF图谱文件
横坐标校正系数及纵坐标
校正系数
图谱绝对坐标数据字典
积分线相对坐标数据字典
含绝对坐标数据的图谱参数
数据字典
结构化数据（文件）

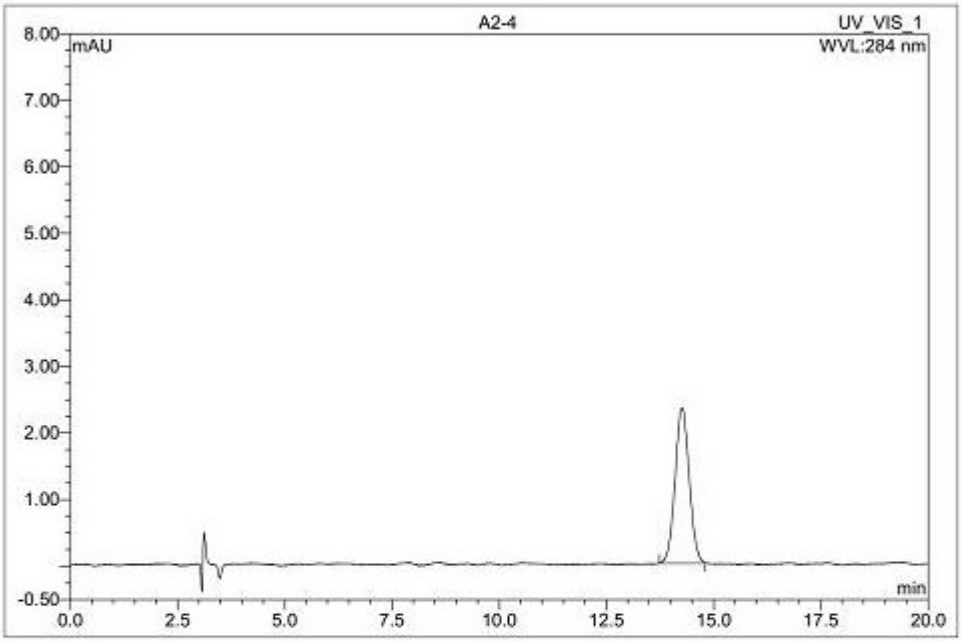
图1

Operator:Administrator Timebase:HPLC Sequence: 某苷

Page 1-1
2020-11-8 11:05 上午

4 A2-4

Sample Name:	**4**	*Channel:*	**UV_VIS_1**
Sample Type:	**unknown**	*Wavelength:*	**284**
Control Program:	**某苷**	*Bandwidth:*	**n.a.**
Quantif. Method:	**方法**	*Inject Volume*	**1.0**
Run Time (min):	**20.00**	*Sample Amount:*	**1.0000**



No.	Ret.Time min	Peak Name	Height mAU	Area mAU*min	S/N	esolution(USP	Plates(EP)
1	14.31	n.a.	2.340	0.897	91.21	n.a.	8941
Total:			2.340	0.897	91.21	0.000	

单组分 myp 2/Integration

Chromeleon (c) Dionex 1996-2006
Version 6.80 SR10 Build 2818 (166959)

图2

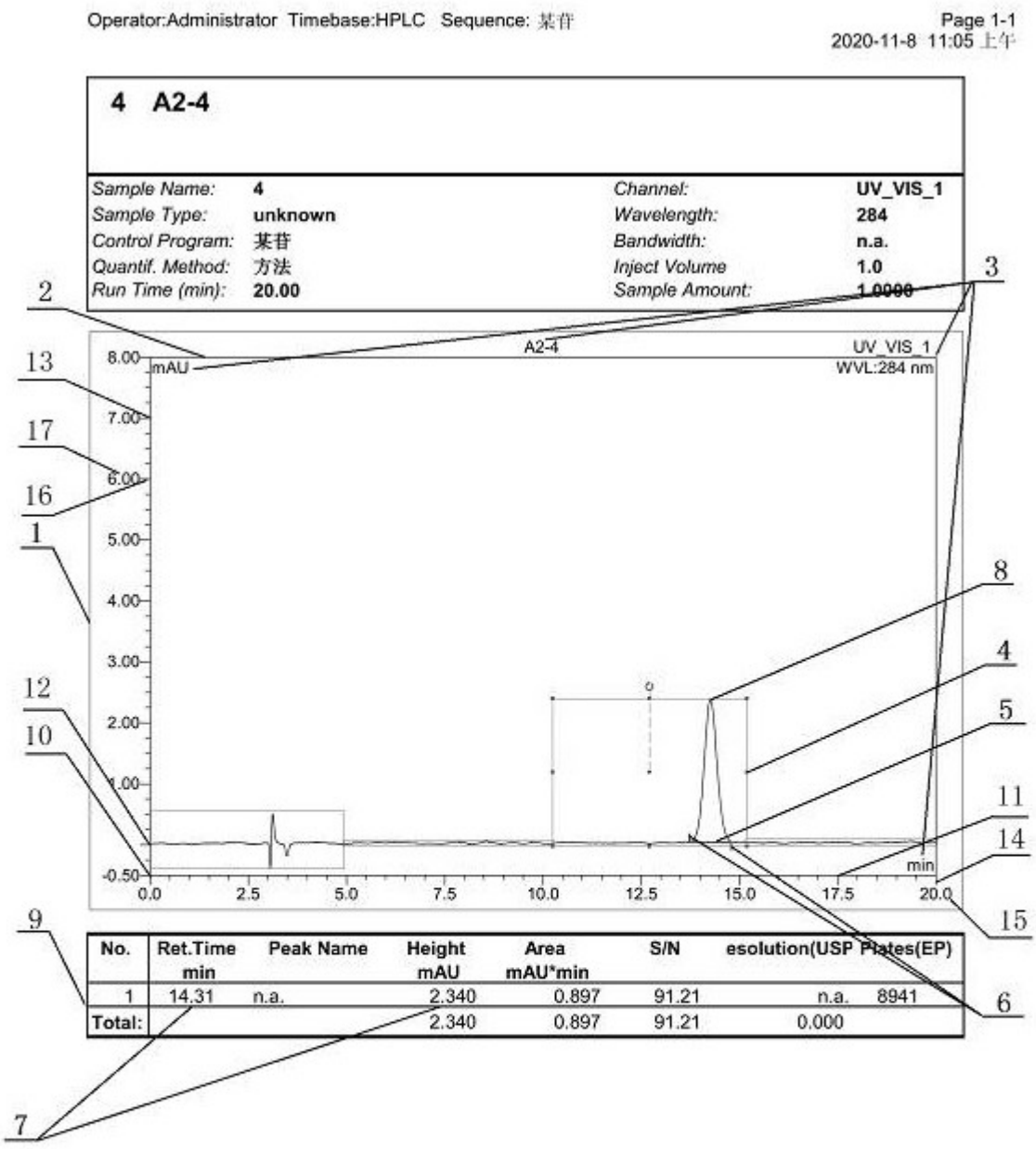

Operator:Administrator Timebase:HPLC Sequence: 某苷
Page 1-1
2020-11-8 11:05 上午
4 A2-4
Sample Name: 4
Sample Type: unknown
Control Program: 某苷
Quantif. Method: 方法
Run Time (min): 20.00
Channel: UV_VIS_1
Wavelength: 284
Bandwidth: n.a.
Inject Volume 1.0
Sample Amount: 1.0000
A2-4
UV_VIS_1
WVL:284 nm
mAU
min
No. Ret.Time min Peak Name Height mAU Area mAU*min S/N esolution(USP Plates(EP)
1 14.31 n.a. 2.340 0.897 91.21 n.a. 8941
Total: 2.340 0.897 91.21 0.000
单组分 myp 2/Integration
Chromeleon (c) Dionex 1996-2006
Version 6.80 SR10 Build 2818 (166959)

图3

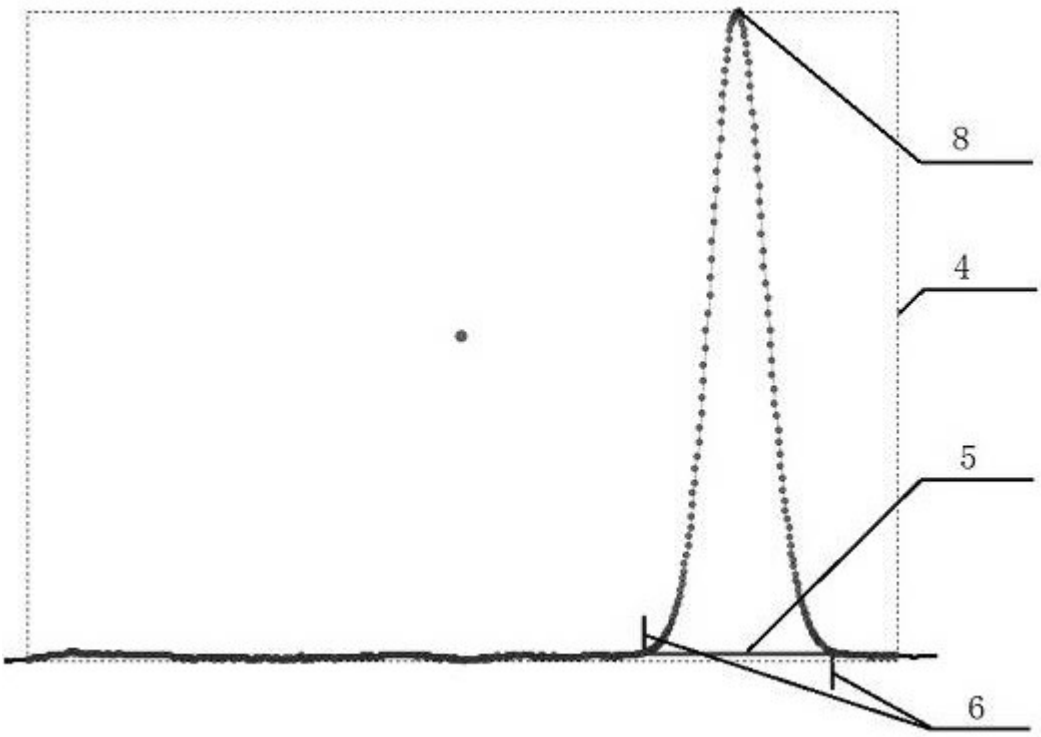

8
4
5
6

图4


管理员: C:\Windows\System32\cmd.exe
Microsoft Windows [版本 6.1.7601]
版权所有 (c) 2009 Microsoft Corporation。保留所有权利。
O:\>python pdfanax
<LTCurve 105.150,345.345,523.327,621.128> :坐标轴框架
<LTCurve 105.150,349.185,208.178,372.962> :图谱曲线
<LTCurve 208.178,361.890,319.538,363.570> :图谱曲线
<LTCurve 319.538,362.130,422.063,438.870> :图谱曲线
<LTCurve 422.063,362.130,523.327,363.570> :图谱曲线
O:\>python pdfanax
<LTCurve 105.150,345.345,523.327,621.128> :坐标轴框架
<LTCurve 105.150,345.345,523.327,621.128> 对象x0,y0,x1,y1: 105.15 345.345 523.327 621.128
<LTCurve 105.150,345.345,523.327,621.128> 对象pts内容: [(105.15, 621.128), (105.15, 345.345), (523.
327, 345.345), (523.327, 621.128), (105.15, 621.128)]
O:\>

图5

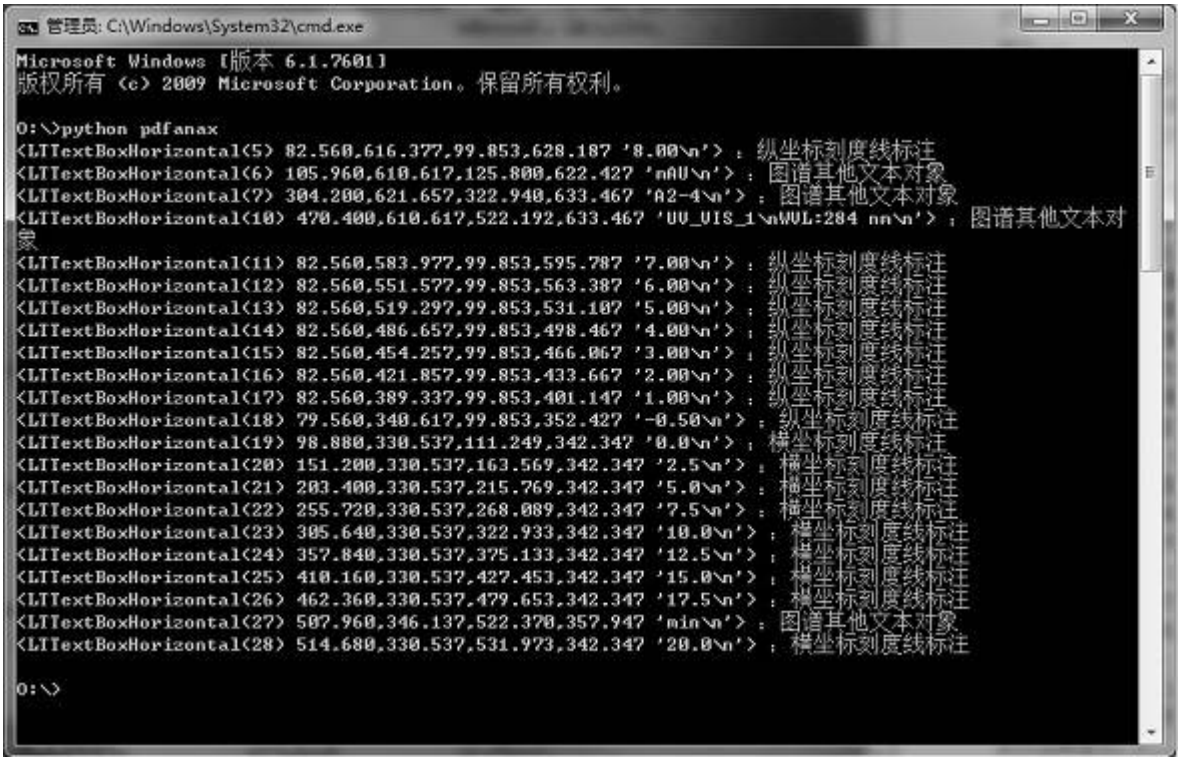

管理员: C:\Windows\System32\cmd.exe
Microsoft Windows [版本 6.1.7601]
版权所有 (c) 2009 Microsoft Corporation。保留所有权利。
O:\>python pdfanax
<LTTextBoxHorizontal(5) 82.560,616.377,99.853,628.187 '8.00\n'> : 纵坐标刻度线标注
<LTTextBoxHorizontal(6) 105.960,610.617,125.800,622.427 'mAU\n'> : 图谱其他文本对象
<LTTextBoxHorizontal(7) 304.200,621.657,322.940,633.467 'A2-4\n'> : 图谱其他文本对象
<LTTextBoxHorizontal(10) 470.400,610.617,522.192,633.467 'UV_VIS_1\nWVL:284 nm\n'> : 图谱其他文本对象
<LTTextBoxHorizontal(11) 82.560,583.977,99.853,595.787 '7.00\n'> : 纵坐标刻度线标注
<LTTextBoxHorizontal(12) 82.560,551.577,99.853,563.387 '6.00\n'> : 纵坐标刻度线标注
<LTTextBoxHorizontal(13) 82.560,519.297,99.853,531.107 '5.00\n'> : 纵坐标刻度线标注
<LTTextBoxHorizontal(14) 82.560,486.657,99.853,498.467 '4.00\n'> : 纵坐标刻度线标注
<LTTextBoxHorizontal(15) 82.560,454.257,99.853,466.067 '3.00\n'> : 纵坐标刻度线标注
<LTTextBoxHorizontal(16) 82.560,421.857,99.853,433.667 '2.00\n'> : 纵坐标刻度线标注
<LTTextBoxHorizontal(17) 82.560,389.337,99.853,401.147 '1.00\n'> : 纵坐标刻度线标注
<LTTextBoxHorizontal(18) 79.560,340.617,99.853,352.427 '-0.50\n'> : 纵坐标刻度线标注
<LTTextBoxHorizontal(19) 98.880,330.537,111.249,342.347 '0.0\n'> : 横坐标刻度线标注
<LTTextBoxHorizontal(20) 151.200,330.537,163.569,342.347 '2.5\n'> : 横坐标刻度线标注
<LTTextBoxHorizontal(21) 203.400,330.537,215.769,342.347 '5.0\n'> : 横坐标刻度线标注
<LTTextBoxHorizontal(22) 255.720,330.537,268.089,342.347 '7.5\n'> : 横坐标刻度线标注
<LTTextBoxHorizontal(23) 305.640,330.537,322.933,342.347 '10.0\n'> : 横坐标刻度线标注
<LTTextBoxHorizontal(24) 357.840,330.537,375.133,342.347 '12.5\n'> : 横坐标刻度线标注
<LTTextBoxHorizontal(25) 410.160,330.537,427.453,342.347 '15.0\n'> : 横坐标刻度线标注
<LTTextBoxHorizontal(26) 462.360,330.537,479.653,342.347 '17.5\n'> : 横坐标刻度线标注
<LTTextBoxHorizontal(27) 507.960,346.137,522.370,357.947 'min\n'> : 图谱其他文本对象
<LTTextBoxHorizontal(28) 514.680,330.537,531.973,342.347 '20.0\n'> : 横坐标刻度线标注
O:\>

图6

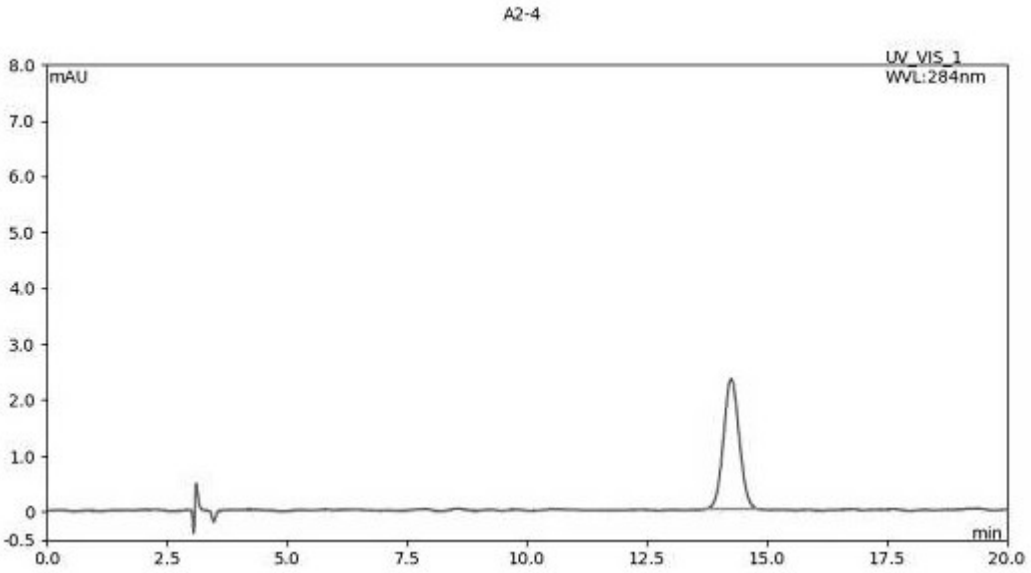
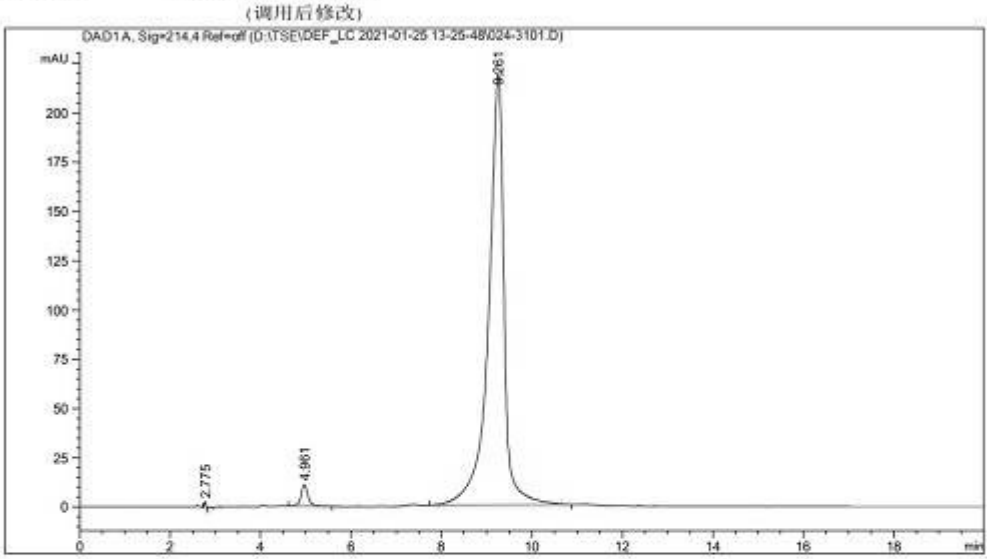

A2-4
UV_VIS_1
WVL:284nm
mAU
min
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0
-0.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

图7

数据文件: D:\TSE\DEF_LC 2021-01-25 13-25-48\024-3101.D
样品名称: T21

操作者 :
仪器 : 仪器 1
进样日期 : 2021/1/25 13:28:40
序列行 : 1
位置 : 样品瓶 21
进样次数 : 1
进样量 : 10.000 µl
采集方法 : D:\TSE\DEF_LC 2021-01-25 13-25-48\含测.M
最后修改 : 2021/1/25 13:20:42
分析方法 : E:\TSE\DEF_LC 2018-09-25 16-25-48\含测.M (序列方法)
最后修改 : 2021/1/29 14:18:59
(调用后修改)



面积百分比报告

排序 : 信号
乘积因子: : 1.0000
稀释因子: : 1.0000
内标使用乘积因子和稀释因子

信号 1: DAD1 A, Sig=214,4 Ref=off

峰 #	保留时间 [min]	类型	峰宽 [min]	峰面积 [mAU*s]	峰高 [mAU]	峰面积 %
1	2.775	VB	0.0501	9.94864	3.27696	0.1846
2	4.961	BB	0.1625	118.20660	10.92964	2.1937
3	9.261	BB	0.3574	5260.20264	219.14706	97.6216

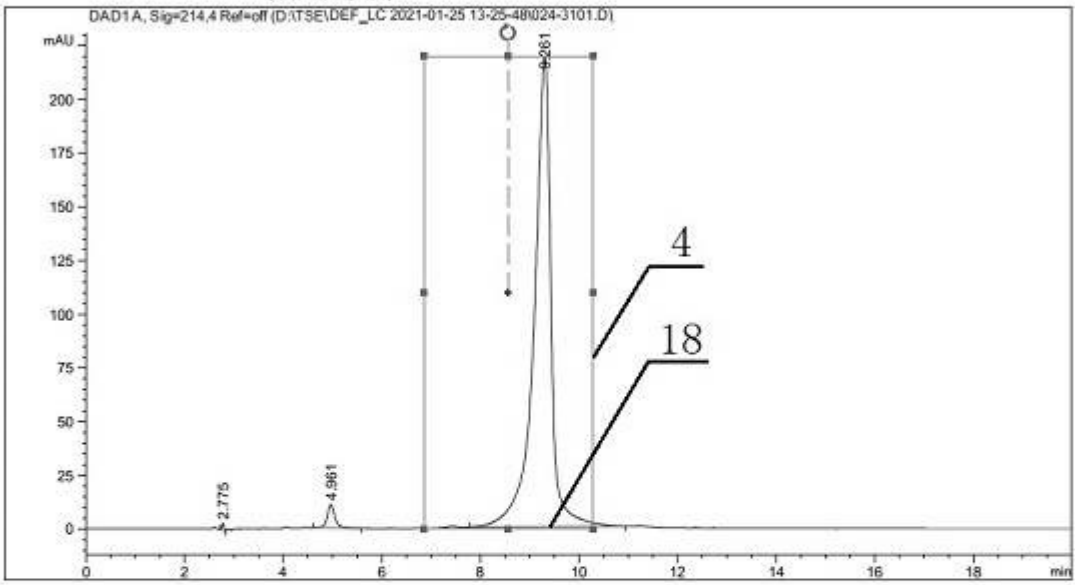
总量 : 5388.35787 233.35366

仪器 1 2021/1/29 14:19:31 页 1/1

图8

数据文件: D:\TSE\DEF_LC 2021-01-25 13-25-48\024-3101.D
样品名称: T21

操作者 :
仪器 : 仪器 1
进样日期 : 2021/1/25 13:28:40
序列行 : 1
位置 : 样品瓶 21
进样次数 : 1
进样量 : 10.000 µl
采集方法 : D:\TSE\DEF_LC 2021-01-25 13-25-48\含测.M
最后修改 : 2021/1/25 13:20:42
分析方法 : E:\TSE\DEF_LC 2018-09-25 16-25-48\含测.M (序列方法)
最后修改 : 2021/1/29 14:18:59
(调用后修改)



面积百分比报告

排序 : 信号
乘积因子: : 1.0000
稀释因子: : 1.0000
内标使用乘积因子和稀释因子

信号 1: DAD1 A, Sig=214,4 Ref=off

峰 #	保留时间 [min]	类型	峰宽 [min]	峰面积 [mAU*s]	峰高 [mAU]	峰面积 %
1	2.775	VB	0.0501	9.94864	3.27696	0.1846
2	4.961	BB	0.1625	118.20660	10.92964	2.1937
3	9.261	BB	0.3574	5260.20264	219.14706	97.6216

总量 : 5388.35787 233.35366

仪器 1 2021/1/29 14:19:31 页 1/1

图9


C:\Windows\System32\cmd.exe
Microsoft Windows [版本 6.1.7601]
版权所有 (c) 2009 Microsoft Corporation。保留所有权利。
E:\>python pdfanax2
<LTCurve 67.440,387.360,156.960,391.440> : 图谱曲线
<LTCurve 156.960,388.320,246.720,401.520> : 图谱曲线
<LTCurve 246.720,388.560,336.480,641.040> : 图谱曲线
<LTCurve 336.480,388.800,426.240,391.680> : 图谱曲线
<LTCurve 426.240,388.560,516.000,388.800> : 图谱曲线
<LTCurve 516.000,388.800,594.000,388.800> : 图谱曲线
<LTCurve 138.960,387.360,141.840,388.080> : 积分线
<LTCurve 188.880,388.800,214.320,389.040> : 积分线
<LTCurve 271.200,389.280,354.000,389.760> : 积分线
E:\>python pdfanax2
<LTCurve 138.960,387.360,141.840,388.080> : 积分线
<LTCurve 138.960,387.360,141.840,388.080> 对象pts内容: [(138.96, 388.08), (138.96, 387.84), (141.84
, 387.36), (141.84, 387.36)]
E:\>

图10