(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 110807449 A
(43)申请公布日 2020.02.18

(21)申请号 202010015896.6

(22)申请日 2020.01.08

(71)申请人 杭州皓智天诚信息科技有限公司
地址 310000 浙江省杭州市余杭区五常街道文一西路998号1幢1206B室

(72)发明人 江峰 李缙航

(74)专利代理机构 杭州创智卓英知识产权代理事务所(普通合伙) 33324
代理人 郑思思

(51)Int.Cl.
G06K 9/00(2006.01)
G06K 9/62(2006.01)
G06T 5/00(2006.01)
G06T 5/20(2006.01)
G06T 7/13(2017.01)
G06T 7/155(2017.01)

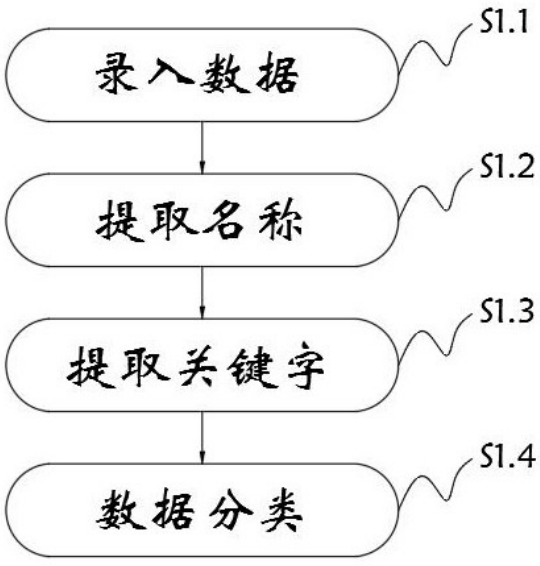
权利要求书2页 说明书7页 附图5页

(54)发明名称

一种科技项目申报线上服务终端

(57)摘要

本发明涉及服务终端技术领域，具体地说，涉及一种科技项目申报线上服务终端。其包括资料收集单元、资料预检查单元和信息查询单元，资料收集单元用于对申报的科技项目数据资料进行收集和归类，资料预检查单元用于对申报的科技项目数据资料进行预处理检查。该科技项目申报线上服务终端中，基于边缘文字检测算法提取录入的科技项目名称信息，并对录入的科技项目名称信息的关键字进行提取，按照关键字的相似度对科技项目进行分类，完成科技项目的录入，便于后期分类处理，提高处理效率，采用资料预检查单元对申报的科技项目数据资料进行预处理检查，提高申报的科技项目数据的完整性。



CN 110807449 A

1.一种科技项目申报线上服务终端，包括资料收集单元、资料预检查单元和信息查询单元，其特征在于：所述资料收集单元用于对申报的科技项目数据资料进行收集和归类，所述资料预检查单元用于对申报的科技项目数据资料进行预处理检查，所述信息查询单元用于对申报的科技项目数据资料处理流程溯源信息进行查询。

2.根据权利要求1所述的科技项目申报线上服务终端，其特征在于：所述资料收集单元包括如下流程步骤：

S1.1、录入数据：录入科技项目数据；

S1.2、提取名称：提取录入的科技项目名称数据；

S1.3、提取关键字：提取科技项目名称数据中的关键词；

S1.4、数据分类：根据提取关键字的相似度对录入科技项目数据进行分类。

3.根据权利要求2所述的科技项目申报线上服务终端，其特征在于：所述S1.2中，提取名称选用边缘文字检测算法，其算法流程如下：

S1.2.1、使用边缘检测算子检测出名称文字边缘特征；

S1.2.2、对边缘特征进行滤波处理；

S1.2.3、通过形态学操作将边缘合并呈区域；

S1.2.4、根据水平投影算法提取文字区域。

4.根据权利要求3所述的科技项目申报线上服务终端，其特征在于：所述边缘检测算子采用Sobel算子检测文字边缘特征，其算子公式为：

$$S_x=\begin{bmatrix}-1 & 0 & 1\\ -2 & 0 & 2\\ -1 & 0 & 1\end{bmatrix} \quad (1)$$

$$S_y=\begin{bmatrix}1 & 2 & 1\\ 0 & 0 & 0\\ -1 & -2 & -1\end{bmatrix} \quad (2)$$

$$K=\begin{bmatrix}a_0 & a_1 & a_2\\ a_7 & (i,\ j) & a_3\\ a_6 & a_5 & a_4\end{bmatrix} \quad (3)$$

K代表邻域点标记矩阵模板，以(i,j)为中心3×3 邻域矩阵，a为是条件中的控制因子，取值范围为0至1，通过多个的a取值来控制边缘的宽度；

矩阵(1)、(2)和(3)分别为该算子的X向卷积模板、Y向卷积模板以及待处理点的邻域点标记矩阵。

5.根据权利要求3所述的科技项目申报线上服务终端，其特征在于：所述边缘特征进行滤波处理采用高斯滤波处理，其公式如下：

$$\mathrm{K}=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma*\sigma}} \quad (4)$$

其中，σ高斯滤波器宽度，σ决定着平滑程度，x为坐标，控制高斯核形状。

6.根据权利要求3所述的科技项目申报线上服务终端，其特征在于：所述水平投影算法的公式如下：

$$HP(i)=\sum_{i-1}^{h}E(i,j) \quad (5)$$

其中,E表示文本区域的边缘图,i,j是图像中像素点的坐标,h为图像的高度,$HP(i)$为横坐标为i的水平投影。

7.根据权利要求2所述的科技项目申报线上服务终端,其特征在于:所述S1.3中,提取关键字采用TFIDF算法,其算法流程如下:

S1.3.1、先给本聚类内的所有文档进行分词,然后用一个字典保存每个词出现的次数;

S1.3.2、遍历每个词,得到每个词在所有文档里的IDF值以及在本聚类内出现的次数TF相乘的值;

S1.3.3、用一个字典来保存所有的词信息,然后按value对字典排序,最后取权重排名靠前的几个词作为关键词。

8.根据权利要求2所述的科技项目申报线上服务终端,其特征在于:所述关键字的相似度采用汉明距离的文本相似度计算方法,其计算方法公式如下:

$$D(x,y)=\sum_{k-1}^{n} x_k \oplus y_k \quad (6)$$

其中,$\oplus$表示模2加运算,$x_k \in \{0,1\}$,$y_k \in \{0,1\}$,$D(x,y)$表示两码字在相同位置上不同码符号的数目的总和,n为两个长码字之间的距离,k为码字个数。

9.根据权利要求2所述的科技项目申报线上服务终端,其特征在于:所述数据分类采用K-means聚类算法,其方法步骤如下:

S1.4.1、对于等待聚类的文本集D,确定要生成的簇的数目k;

S1.4.2、生成k个聚类中心作为聚类的初始中心点,$S=\{s_1,s_2,\dots\dots s_j,\dots\dots,s_k\}$;

S1.4.3、对D中的每一个文本d_j,依次计算它与各个中心点s_j的相似度$sim(d_j,s_j)$;

S1.4.4、选取具有最大的相似度的中心点$\arg\ max\, sim(d_j,s_j)$,将d_i归入以s_j为聚类中心的簇C_j,从而得到D一个聚类$C=\{C_1,\dots,C_k\}$;

S1.4.5、重新确定每个簇的中心点;

S1.4.6、反复执行S1.4.3-S1.4.5,到中心点不再改变,文本不再重新被分配为止。

一种科技项目申报线上服务终端

技术领域

[0001] 本发明涉及服务终端技术领域，具体地说，涉及一种科技项目申报线上服务终端。

背景技术

[0002] 项目申报是指政府机关针对企业或其他研究单位作出的一系列优惠政策，企业或相关研究单位再根据政府的政策进行编写申报文件然后根据相关申报要求和流程进行申报。随着人们知识产权保护意识的提升，对科技项目的申报数量日益加剧，而现有的科技项目申报终端仅仅只能对科技项目申报信息进行收集，但科技项目申报信息种类繁多，且其中含有的无效数据较多，后期处理困难，处理效率低。

发明内容

[0003] 本发明的目的在于提供一种科技项目申报线上服务终端，以解决上述背景技术中提出的问题。

[0004] 为实现上述目的，本发明提供一种科技项目申报线上服务终端，包括资料收集单元、资料预检查单元和信息查询单元，所述资料收集单元用于对申报的科技项目数据资料进行收集和归类，所述资料预检查单元用于对申报的科技项目数据资料进行预处理检查，所述信息查询单元用于对申报的科技项目数据资料处理流程溯源信息进行查询。

[0005] 作为优选，所述资料收集单元包括如下流程步骤：

S1.1、录入数据：录入科技项目数据；

S1.2、提取名称：提取录入的科技项目名称数据；

S1.3、提取关键字：提取科技项目名称数据中的关键词；

S1.4、数据分类：根据提取关键字的相似度对录入科技项目数据进行分类。

[0006] 作为优选，所述S1.2中，提取名称选用边缘文字检测算法，其算法流程如下：

S1.2.1、使用边缘检测算子检测出名称文字边缘特征；

S1.2.2、对边缘特征进行滤波处理；

S1.2.3、通过形态学操作将边缘合并呈区域；

S1.2.4、根据水平投影算法提取文字区域。

[0007] 作为优选，所述边缘检测算子采用Sobel算子检测文字边缘特征，其算子公式为：

$$S_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad (1)$$

$$S_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \quad (2)$$

$$K = \begin{bmatrix} a_0 & a_1 & a_2 \\ a_7 & (i,\ j) & a_3 \\ a_6 & a_5 & a_4 \end{bmatrix} \quad (3)$$

K代表邻域点标记矩阵模板,以(i,j)为中心3×3 邻域矩阵,a为是条件中的控制因子,取值范围为0至1,通过多个的a取值来控制边缘的宽度;

矩阵(1)、(2)和(3)分别为该算子的x向卷积模板、y向卷积模板以及待处理点的邻域点标记矩阵。

[0008] 作为优选,所述边缘特征进行滤波处理采用高斯滤波处理,其公式如下:

$$\mathrm{K} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma*\sigma}} \quad (4)$$

其中,σ高斯滤波器宽度,σ决定着平滑程度,x为坐标,控制高斯核形状。

[0009] 作为优选,所述水平投影算法的公式如下:

$$HP(i) = \sum_{i-1}^{h} E(i,j) \quad (5)$$

其中,E表示文本区域的边缘图,i,j是图像中像素点的坐标,h为图像的高度,$HP(i)$为横坐标为i的水平投影。

[0010] 作为优选,所述S1.3中,提取关键字采用TFIDF算法,其算法流程如下:

S1.3.1、先给本聚类内的所有文档进行分词,然后用一个字典保存每个词出现的次数;

S1.3.2、遍历每个词,得到每个词在所有文档里的IDF值以及在本聚类内出现的次数TF相乘的值;

S1.3.3、用一个字典来保存所有的词信息,然后按value对字典排序,最后取权重排名靠前的几个词作为关键词。

[0011] 作为优选,所述关键字的相似度采用汉明距离的文本相似度计算方法,其计算方法公式如下:

$$D(x,y)=\sum_{k-1}^{n} x_k \oplus y_k \quad (6)$$

其中,$\oplus$表示模2加运算,$x_k \in \{0,1\}$, $y_k \in \{0,1\}$, $D(x,y)$表示两码字在相同位置上不同码符号的数目的总和,n为两个长码字之间的距离,k为码字个数。

[0012] 作为优选,所述数据分类采用K-means聚类算法,其方法步骤如下:

S1.4.1、对于等待聚类的文本集D,确定要生成的簇的数目k;

S1.4.2、生成k个聚类中心作为聚类的初始中心点,$S = \{s_1, s_2, \ldots\ldots s_j, \ldots\ldots, s_k\}$;

S1.4.3、对D中的每一个文本d_j，依次计算它与各个中心点s_j的相似度$sim(d_j, s_j)$；

S1.4.4、选取具有最大的相似度的中心点$\arg\ max\ sim(d_j, s_j)$，将d_i归入以s_j为聚类中心的簇C_j，从而得到D一个聚类$C = \{C_1, \ldots, C_k\}$；

S1.4.5、重新确定每个簇的中心点；

S1.4.6、反复执行S1.4.3-S1.4.5，到中心点不再改变，文本不再重新被分配为止。

[0013] 与现有技术相比，本发明的有益效果：

1、该科技项目申报线上服务终端中，基于边缘文字检测算法提取录入的科技项目名称信息，并对录入的科技项目名称信息的关键字进行提取，按照关键字的相似度对科技项目进行分类，完成科技项目的录入，便于后期分类处理，提高处理效率。

[0014] 2、该科技项目申报线上服务终端中，采用资料预检查单元对申报的科技项目数据资料进行预处理检查，提高申报的科技项目数据的完整性。

附图说明

[0015] 图1为本发明的整体流程框图；

图2为本发明的边缘文字检测算法流程框图；

图3为本发明的提取关键字流程框图；

图4为本发明的数据分类流程框图；

图5为本发明的膨胀单元原理图；

图6为本发明的腐蚀单元原理图。

具体实施方式

[0016] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，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做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0017] 请参阅图1-图6所示，本发明提供一种技术方案：

本发明提供一种科技项目申报线上服务终端，包括资料收集单元、资料预检查单元和信息查询单元，资料收集单元用于对申报的科技项目数据资料进行收集和归类，资料预检查单元用于对申报的科技项目数据资料进行预处理检查，信息查询单元用于对申报的科技项目数据资料处理流程溯源信息进行查询。

[0018] 本实施例中，服务终端采用J2EE方式是实现，主要采用servlet技术实现与使用者的移动终端产生交互，采用J2EE具有平台无关性，易移植，性能高，容易部署等特点，在该系统实现中，只需要在机房安装一台小型机做为服务器硬件，同时申请一个域名，即可实现信息共享，使用者只需与服务终端进行交互就可实现资料收集单元、资料预检查单元和信息查询单元的使用。

[0019] 进一步的，资料收集单元包括如下流程步骤：

S1.1、录入数据：录入科技项目数据；

S1.2、提取名称：提取录入的科技项目名称数据；

S1.3、提取关键字:提取科技项目名称数据中的关键词;

S1.4、数据分类:根据提取关键字的相似度对录入科技项目数据进行分类。

[0020] 其中,S1.2中,提取名称选用边缘文字检测算法,其算法流程如下:

S1.2.1、使用边缘检测算子检测出名称文字边缘特征;

S1.2.2、对边缘特征进行滤波处理;

S1.2.3、通过形态学操作将边缘合并呈区域;

S1.2.4、根据水平投影算法提取文字区域。

[0021] 进一步的,边缘检测算子采用Sobel算子检测文字边缘特征,其算子公式为:

$$S_x=\begin{bmatrix}-1 & 0 & 1\\ -2 & 0 & 2\\ -1 & 0 & 1\end{bmatrix} \quad (1)$$

$$S_y=\begin{bmatrix}1 & 2 & 1\\ 0 & 0 & 0\\ -1 & -2 & -1\end{bmatrix} \quad (2)$$

$$K=\begin{bmatrix}a_0 & a_1 & a_2\\ a_7 & (i,\ j) & a_3\\ a_6 & a_5 & a_4\end{bmatrix} \quad (3)$$

K代表邻域点标记矩阵模板,以(i,j)为中心3×3 邻域矩阵,a为是条件中的控制因子,取值范围为0至1,通过多个的a取值来控制边缘的宽度;

矩阵(1)、(2)和(3)分别为该算子的x向卷积模板、y向卷积模板以及待处理点的邻域点标记矩阵,据此可用数学公式表达其每个点的梯度幅值为:

$$G(i,\ j)=\sqrt{s_x^2+s_y^2} \quad (7)$$

$$S_x=(a_2+2a_3+a_4)-(a_0+2a_7+a_6) \quad (8)$$

$$S_y=(a_0+2a_1+a_2)-(a_6+2a_5+a_4) \quad (9)\ 。$$

具体的,边缘特征进行滤波处理采用高斯滤波处理,高斯滤波的实现可以用两个一维高斯核分别两次加权实现,高斯核实现公式如下:

$$K=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma*\sigma}} \quad (4)$$

其中,σ高斯滤波器宽度, σ决定着平滑程度,x为坐标,控制高斯核形状。

[0022] 式(4)为离散化的一维高斯函数,确定参数就可以得到一维核向量,其公式为:

$$K=\frac{1}{2\pi\sigma*\sigma}e^{-\frac{x*x+y*y}{2\sigma*\sigma}} \quad (4.1)$$

式(4.1)为离散化的二维高斯函数,确定参数就可以得到二维核向量。

[0023] 值得说明的是,水平投影算法的公式如下:

$$HP(i)=\sum_{i-1}^{h}E(i,j)\ (5)$$

其中，E表示文本区域的边缘图，i,j是图像中像素点的坐标，h为图像的高度，$HP(i)$为横坐标为i的水平投影。

[0024] 值得说明的是，形态学操作包括膨胀单元、腐蚀单元、开运算单元和闭运算单元。

[0025] 其中，膨胀单元定义为：把结构元素B平移a后得到B_a，若B_a击中X，我们记下这个a点。所有满足上述条件的a点组成的集合称做X被B膨胀的结果。用公式表示为：$D(X)=\{a|B_a\uparrow X\}=X\oplus B$，($B_a\uparrow X$表示$B_a$击中X，$\oplus$表示异或运算，其运算法则为$a\oplus b=(\neg a\wedge b)\vee(a\wedge\neg b)$)，如图5所示，X是被处理的对象，B是结构元素，对于任意一个在阴影部分的点a，Ba击中X，X被B膨胀的结果就是图5中阴影部分。

[0026] 其中，腐蚀单元定义为：结构元素B平移a后得到B_a，若B_a包含于X，我们记下这个a点，所有满足上述条件的a点组成的集合称做X被B腐蚀的结果，用公式表示为：$E(X)=\{a|B_a\subset X\}=X\ominus B$，其中$X\ominus B$表示X被B腐蚀的结果，如图6所示，图中，X是被处理的对象，B是结构元素，对于任意一个在阴影部分的点a，Ba包含于X，X被B腐蚀的结果为图6中阴影部分。

[0027] 其中，开运算单元是结构元素B对输入图像A的开运算记为$A\bigcirc B$，定义为$A\bigcirc B=(A\ominus B)\oplus B=U\{B+x:B+x\subset A\}$。开运算可以通过计算所有可以填入图像内部的结构元素平移的并求得，即是对A先腐蚀后膨胀运算的结果，开运算具有平滑功能，能清除图像的某些微小连接、边缘毛刺和孤立斑点。

[0028] 其中，闭运算单元是结构元素B对输入图像A的闭运算记为$A\bullet B$，定义为$A\bullet B=(A\oplus B)\ominus B$。闭运算是开运算的对偶运算，即是对A先膨胀后腐蚀运算的结果，闭运算具有过滤功能，可填平图像内部小沟、孔洞和裂缝，使断线相连。

[0029] 进一步的，S1.3中，提取关键字采用TFIDF算法，其算法流程如下：

S1.3.1、先给本聚类内的所有文档进行分词，然后用一个字典保存每个词出现的次数；

S1.3.2、遍历每个词，得到每个词在所有文档里的IDF值以及在本聚类内出现的次数TF相乘的值；

S1.3.3、用一个字典来保存所有的词信息，然后按value对字典排序，最后取权重排名靠前的几个词作为关键词。

[0030] 具体的，关键字的相似度采用汉明距离的文本相似度计算方法，其计算方法公式如下：

$$D(x,y)=\sum_{k-1}^{n}x_k\oplus y_k\cdots\cdots\cdots\cdots\ (6)$$

其中，n为两个长码字之间的距离，k为码字个数，$\oplus$表示模2加运算，$x_k\in\{0,1\}$，$y_k\in\{0,1\}$，$D(x,y)$表示两码字在相同位置上不同码符号的数目的总和，它能够反映两码字之间的差异，可作为提供码字之间的相似程度的客观依据。该方法将文本中的关键词、文摘等信息排列成一个有n个位序列的码字，文本信息就用这些码字表示，使文本与码字建立1-1对应的关系。

[0031] 具体的，若文本w_1对应的码字为M_1，查询式对应的码字为M_2，对于$D(M_1,\ M_2)$来说，它们之间的距离介于0和n之间，当文本与查询式用n位码字表示完全不同时，距离为n，当文本与查询式码字完全相同时，距离为0，相似度计算时，先确定文本集对应的码字集，对于不同的文本或文本与查询式之间，设

$M_1=(x_1, x_2, x_3, \ldots, x_k, \ldots, x_n)$，$M_2=(y_1, y_2, y_3, \ldots, y_k, \ldots, y_n)$，基于汉明距离的相似度计算如公式所示：

$$Sim(M_1, M_2) = 1 - \sum_{k-1}^{n} x_k \oplus y_k \ldots\ldots\ldots\ldots \quad (6.1)$$

其中，x_k、y_k分别表示文本w_1对应的码字M_1和查询式w_2对应的码字M_2中第k位的分量，或者为0或者为1，$\oplus$就是模2加运算。

[0032] 值得说明的是，数据分类采用K-means聚类算法，其方法步骤如下：

S1.4.1、对于等待聚类的文本集D，确定要生成的簇的数目k；

S1.4.2、生成k个聚类中心作为聚类的初始中心点，$S = \{s_1, s_2, \ldots\ldots s_j, \ldots\ldots, s_k\}$；

S1.4.3、对D中的每一个文本d_j，依次计算它与各个中心点s_j的相似度$sim(d_j, s_j)$；

S1.4.4、选取具有最大的相似度的中心点 $\arg\ max\ sim(d_j, s_j)$，将d_i归入以s_j为聚类中心的簇C_j，从而得到D一个聚类$C = \{C_1, \ldots, C_k\}$；

S1.4.5、重新确定每个簇的中心点；

S1.4.6、反复执行S1.4.3-S1.4.5，到中心点不再改变，文本不再重新被分配为止。

[0033] 值得说明的是，资料预检查单元包括纠正错误模块、删除重复项模块、统一规格模块、修正逻辑模块、转换构造模块、数据压缩模块、数据补缺模块和数据丢弃模块。

[0034] 本实施例中，纠正错误模块用于纠正数据错误形式，纠正错误模块用于数据值错误的纠正、数据类型错误的纠正、数据编码错误的纠正、数据格式错误的纠正、数据异常错误的纠正、依赖冲突的纠正和多值错误的纠正。

[0035] 进一步的，由于各种原因，数据中可能存在重复记录或重复字段(列)，对于这些重复项目(行和列)需要删除重复项模块进行处理，删除重复项模块用于删除数据中存在的重复记录或重复字段，对于重复项的判断，基本思想是“排序和合并”，先将数据库中的记录按一定规则排序，然后通过比较邻近记录是否相似来检测记录是否重复。

[0036] 具体的，由于数据源系统分散在各个业务线，不同业务线对于数据的要求、理解和规格不同，导致对于同一数据对象描述规格完全不同，因此在清洗过程中需要通过统一规格模块统一数据规格并将一致性的内容抽象出来。

[0037] 此外，修正逻辑模块用于明确各个源系统的逻辑、条件、口径，并对异常源系统的采集逻辑进行修正。

[0038] 除此之外，转换构造模块用于对数据进行标准化处理，转换构造模块包括数据类型转换、数据语义转换、数据粒度转换、表/数据拆分、行列转换、数据离散化、数据标准化、提炼新字段和属性构造。

[0039] 其中，数据类型转换：当数据来自不同数据源时，不同类型的数据源数据类型不兼容可能导致系统报错，这时需要将不同数据源的数据类型进行统一转换为一种兼容的数据类型。

[0040] 其中，数据语义转换：传统数据仓库中基于第三范式可能存在维度表、事实表等，此时在事实表中会有很多字段需要结合维度表才能进行语义上的解析。

[0041] 其中，数据粒度转换：将数据按照数据仓库中不同的粒度需求进行聚合。

[0042] 其中，表/数据拆分：某些字段可能存储多中数据信息，例如时间戳中包含了年、月、日、小时、分、秒等信息，有些规则中需要将其中部分或者全部时间属性进行拆分，以此来满足多粒度下的数据聚合需求。

[0043] 其中，行列转换：对表内的行列数据进行转换。

[0044] 其中，数据离散化：将连续取值的属性离散化成若干区间，来帮助消减一个连续属性的取值个数。

[0045] 其中，数据标准化：不同字段间由于字段本身的业务含义不同，需要消除变量之间不同数量级造成的数值之间的悬殊差异。

[0046] 其中，提炼新字段：很多情况下，需要基于业务规则提取新的字段，这些字段也称为复合字段。

[0047] 其中，属性构造：在建模过程中，根据已有的属性集构造新的属性。

[0048] 进一步的，数据压缩模块用于保持原有数据集的完整性和准确性，不丢失有用信息的前提下，按照一定的算法和方式对数据进行重新组织，大规模的数据进行复杂的数据分析与数据计算通常需要耗费大量时间，所以在这之前需要进行数据的约减和压缩，减小数据规模，而且还可能面临交互式的数据挖掘，根据数据挖掘前后对比对数据进行信息反馈。这样在精简数据集上进行数据挖掘显然效率更高，并且挖掘出来的结果与使用原有数据集所获得结果基本相同。

[0049] 此外，数据补缺模块用于对残缺数据的数据进行补充，数据补充包括补充缺失值和补充空值，缺失值指的是的数据原本是必须存在的，但实际上没有数据，空值指的是实际存在可能为空的情况。

[0050] 除此之外，数据丢弃模块对于数据中的异常数据进行删除，丢弃数据的类型包含整条删除和变量删除，整条删除指的是删除含有缺失值的样本，变量删除，如果某一变量的无效值和缺失值很多，而且该变量对于所研究的问题不是特别重要，则可以考虑将该变量删除，这种做法减少了供分析用的变量数目，但没有改变样本量。

[0051] 以上显示和描述了本发明的基本原理、主要特征和本发明的优点。本行业的技术人员应该了解，本发明不受上述实施例的限制，上述实施例和说明书中描述的仅为本发明的优选例，并不用来限制本发明，在不脱离本发明精神和范围的前提下，本发明还会有各种变化和改进，这些变化和改进都落入要求保护的本发明范围内。本发明要求保护范围由所附的权利要求书及其等效物界定。

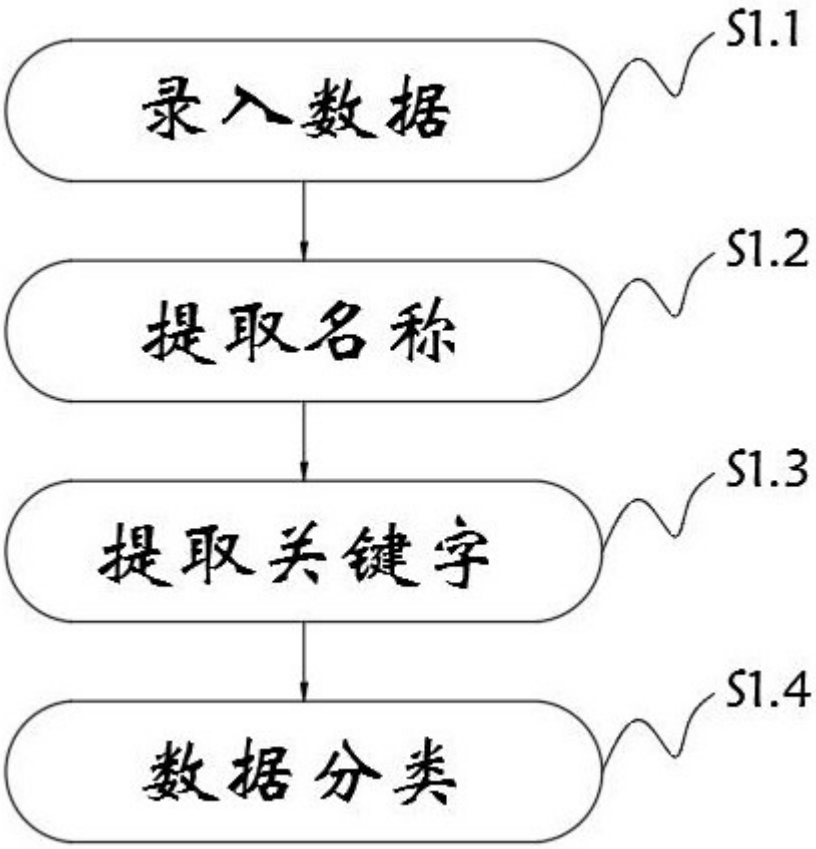


图1

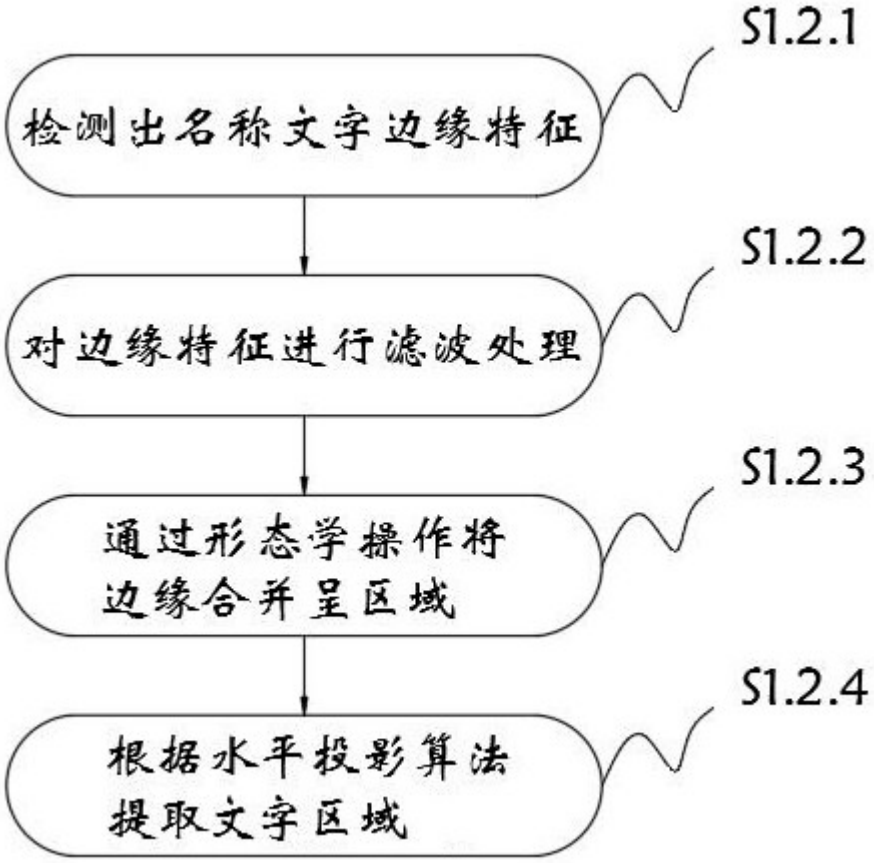

S1.2.1
检测出名称文字边缘特征
S1.2.2
对边缘特征进行滤波处理
S1.2.3
通过形态学操作将
边缘合并呈区域
S1.2.4
根据水平投影算法
提取文字区域

图2

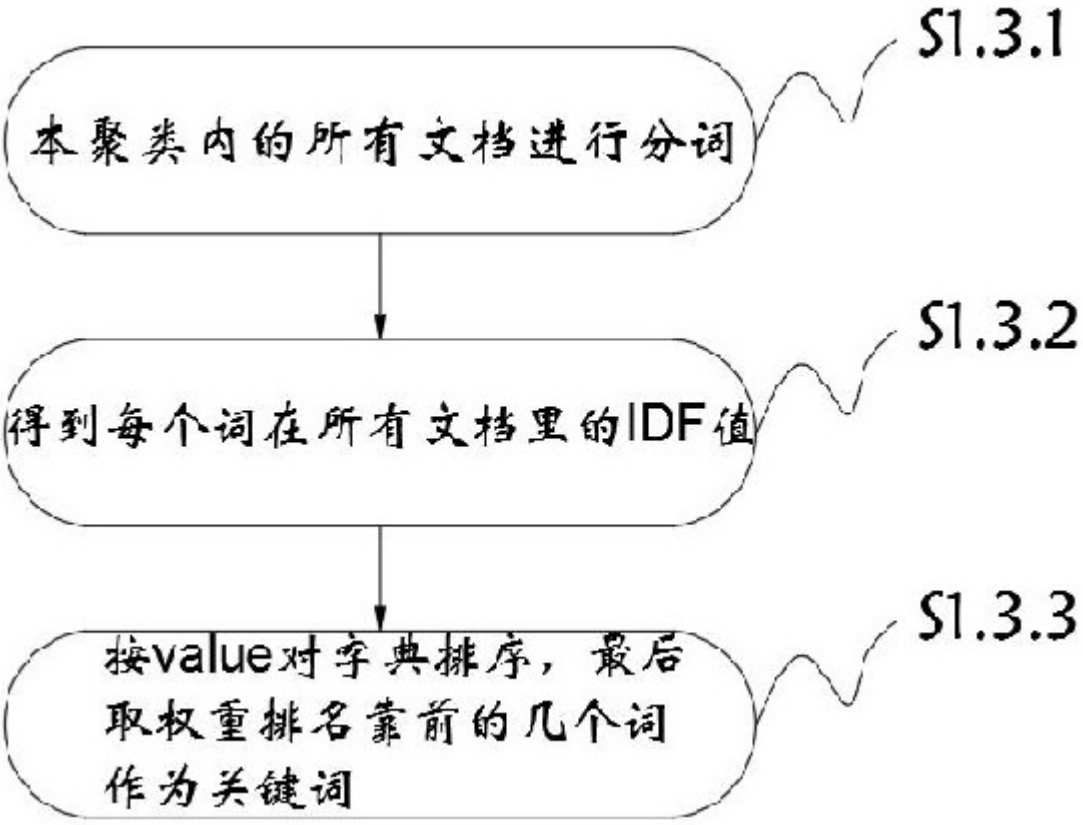

S1.3.1
本聚类内的所有文档进行分词
S1.3.2
得到每个词在所有文档里的IDF值
S1.3.3
按value对字典排序，最后取权重排名靠前的几个词作为关键词

图3

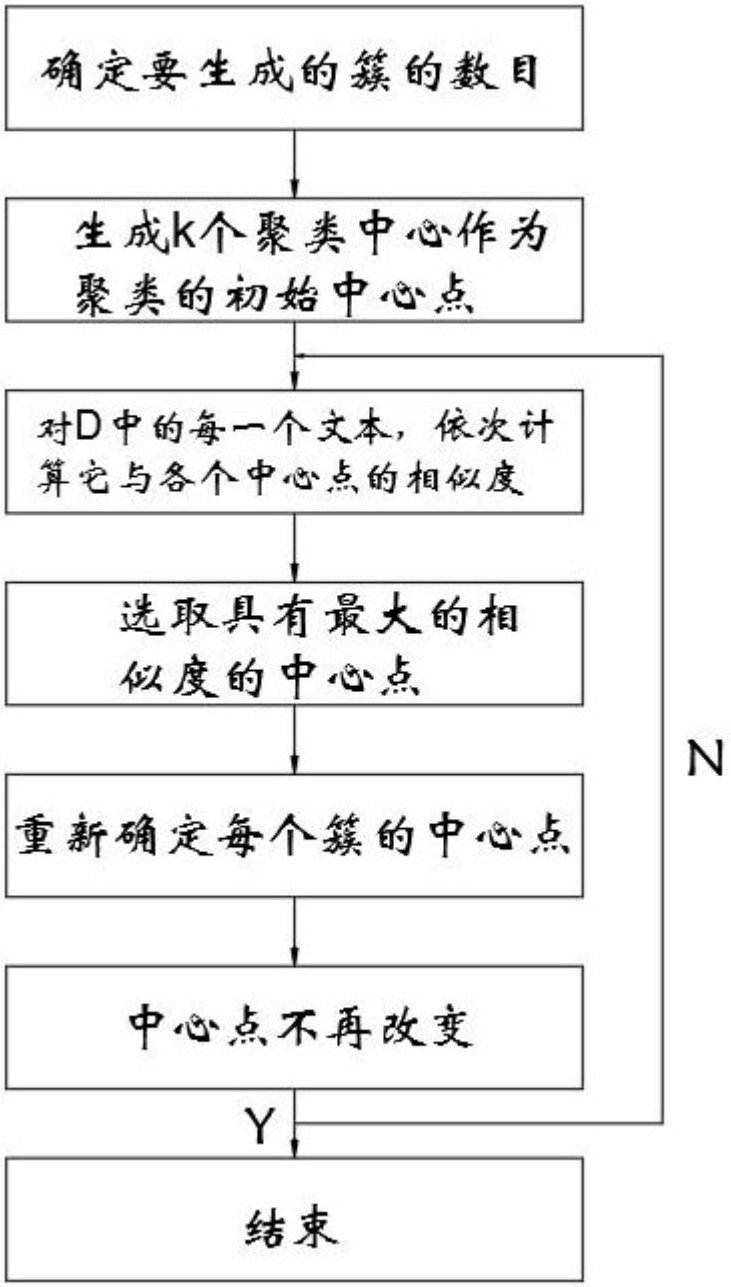

确定要生成的簇的数目
生成k个聚类中心作为聚类的初始中心点
对D中的每一个文本，依次计算它与各个中心点的相似度
选取具有最大的相似度的中心点
重新确定每个簇的中心点
中心点不再改变
N
Y
结束

图4

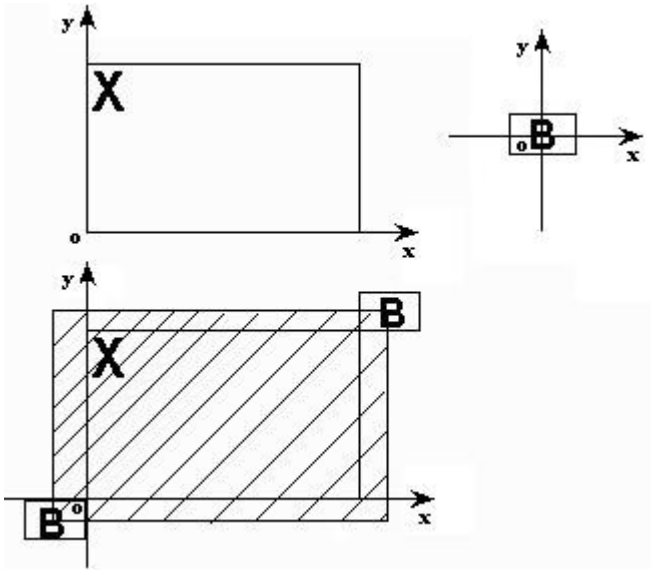


图5

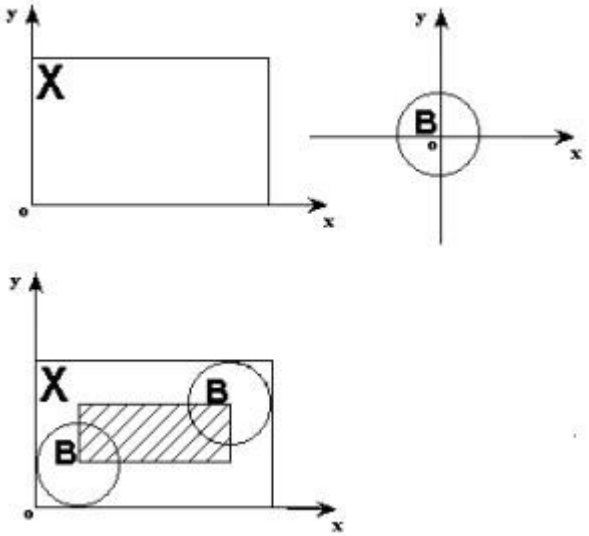


图6