一种基于服务器查找数据流分割点的方法及服务器

技术领域

本发明涉及信息技术领域，尤其涉及一种基于服务器查找数据流分割点的方法及服务器。

背景技术

数据量的不断增长，使得提供充足的数据存储成为当前存储领域面临的严峻挑战。目前应对这一挑战的一种方式为利用需要存储的数据的冗余特性，使用重复数据删除技术，从而减少存储的数据量。

现有技术中，基于内容分块(Content Defined Chunk, CDC)的重复数据删除算法，首先要将待存储的数据流分成很多数据块，而将数据流分成数据块就需要在数据流中查找合适的分割点，两个相邻数据流分割点之间的数据构成一个数据块。计算数据块的特征值，从而查找是否存在相同特征值的数据块，如果查找到相同特征指的数据块，则认为存在重复数据。具体的，基于内容分块的重复数据删除技术是应用滑动窗口技术（Sliding Window Technique）基于文件的内容来查找分块的分割点，即通过计算窗口内数据的Rabin指纹来确定数据流分割点。假设从数据流的左边向右边查找分割点，每次计算滑动窗口内数据的指纹，并且将指纹值基于给定的整数K取模后，与给定的余数R进行比对；若相等则窗口的右端为数据流分割点，否则将窗口继续往右滑动一个字节，依次循环地进行计算和比对，直到到达数据流末尾。在基于内容分块的重复数据删除过程中，查找数据流分割点，需要消耗大量的计算资源，从而成为提升重复数据删除性能的瓶颈。

发明内容

第一方面，本发明实施例提供了一种基于服务器查找数据流分割点的方法，在所述服务器上预设有规则，所述规则为：为潜在分割点

k确定M个点p_x、点p_x对应的窗口$W_x[p_x-A_x,p_x+B_x]$和窗口$W_x[p_x-A_x,p_x+B_x]$对应的预定条件C_x，其中， x为1到M连续的自然数，M≥2，A_x和B_x为整数；所述方法包括:

a)依据所述规则为当前潜在分割点k_i确定点p_{iz}及所述点p_{iz}对应的窗口$W_{iz}[p_{iz}-A_z,p_{iz}+B_z]$，i和z为整数，并且1≤z≤M;

b) 判断所述窗口$W_{iz}[p_{iz}-A_z,p_{iz}+B_z]$中至少部分数据是否满足预定条件$C_z$;

当所述窗口$W_{iz}[p_{iz}-A_z,p_{iz}+B_z]$中至少部分数据不满足所述预定条件$C_z$，从所述点$p_{iz}$沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，N*U不大于$\|B_z\|+\max_x(\|A_x\|+\|(k_i-p_{ix})\|)$，获得新的潜在分割点，执行步骤a)；

c)当所述当前潜在分割点k_i的M个窗口中的每一个窗口$W_{ix}[p_{ix}-A_x,p_{ix}+B_x]$中至少部分数据满足预定条件$C_x$，则所述当前潜在分割点$k_i$为数据流分割点。

结合第一方面，第一种可能的实现方式中，所述规则还包括：至少两个点p_e和p_f,满足条件$A_e=A_f$， $B_e=B_f$， $C_e=C_f$。

结合第一方面的第一种可能的实现方式，第二种可能的实现方式中，所述规则还包括：所述至少两个点p_e和p_f,相对于所述潜在分割点k,在所述数据流分割点查找反方向上。

结合第一方面的第一种可能的实现方式或第二种可能的实现方式，第三种可能的实现方式中，所述规则还包括：所述至少两个点p_e和p_f之间的距离为1个U。

结合第一方面，或第一方面第一至第三种任一可能的实现方式，第四种可能的实现方式中，判断所述窗口$W_{iz}[p_{iz}-A_z,p_{iz}+B_z]$中至少部

分数据是否满足所述预定条件C_z，具体包括:

使用随机函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$。

结合第一方面的第四种可能的实现方式，第五种可能的实现方式中，所述使用随机函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$，具体为使用hash函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$。

结合第一方面，或第一方面第一至第五种任一可能的实现方式，第六种可能的实现方式中，当所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据不满足所述预定条件$C_z$，从所述点$p_{iz}$沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，获得所述新的潜在分割点，根据所述规则，为所述新的潜在分割点确定的点p_{ic}对应的窗口$W_{ic}[p_{ic}-A_c, p_{ic}+B_c]$的左边界与所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$的右边界重合或者为所述新的潜在分割点确定的所述点$p_{ic}$对应的所述窗口$W_{ic}[p_{ic}-A_c, p_{ic}+B_c]$的左边界位于所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$范围之内；其中，为所述新的潜在分割点确定的所述点$p_{ic}$是根据所述规则，为所述新的潜在分割点确定的M个点按照数据流查找方向获得的序列中排序第一的点。

结合第一方面的第四种可能的实现方式，第七种可能的实现方式中，使用随机函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$，具体包括:

在所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中选择F个字节，将所述F个字节反复利用H次，共获得F*H个字节，其中每个字节由8位组成，记为$a_{m,1}\ldots a_{m,8}$，表示所述F*H个字节中第m个字节的第1到第8位，所述F*H个字

节对应的位可以表示为：$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,8} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ a_{F*H,1} & a_{F*H,2} & \cdots & a_{F*H,8} \end{pmatrix}$，当 $a_{m,n}$=1时，$V_{am,n}$ =1，当 $a_{m,n}$=0时，$V_{am,n}$=-1，其中 $a_{m,n}$ 表示 $a_{m,1}\ldots a_{m,8}$ 中的任一个，所述F*H个字节对应的位按照 $a_{m,n}$ 与 $V_{am,n}$ 的转换关系得到矩阵 V_a，所述矩阵 V_a 表示为：$\begin{pmatrix} V_{a1,1} & V_{a1,2} & \cdots & V_{a1,8} \\ V_{a2,1} & V_{a2,2} & \cdots & V_{a2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{aF*H,1} & V_{aF*H,2} & \cdots & V_{aF*H,8} \end{pmatrix}$，从服务正态分布的随机数中选择F*H*8个随机数组成矩阵R，所述矩阵R表示为：

$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,8} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{F*H,1} & h_{F*H,2} & \cdots & h_{F*H,8} \end{pmatrix}$，将所述矩阵 V_a 的第m行与所述矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为 $S_{am}=V_{am,1}*h_{m,1}+V_{am,2}*h_{m,2}+\ldots+V_{am,8}*h_{m,8}$，同理，获得 S_{a1}、S_{a2} …到 S_{aF*H}，统计 S_{a1}、S_{a2} …到 S_{aF*H} 中满足大于0的值的个数K，当K为偶数，则所述窗口 $W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$ 中至少部分数据满足所述预定条件 C_z。

第二方面，本发明实施例提供了一种基于服务器查找数据流分割点的方法，在所述服务器上预设有规则，所述规则为：为潜在分割点 k 确定M个窗口 $W_x[k-A_x, k+B_x]$ 和窗口 $W_x[k-A_x, k+B_x]$ 对应的预定条件 C_x，其中，x为1到M连续的自然数，M≥2，A_x 和 B_x 为整数；

所述方法包括：

a) 依据所述规则为当前潜在分割点 k_i 确定对应的窗口 $W_{iz}[k_i-A_z, k_i+B_z]$，i和z为整数，并且1≤z≤M；

b) 判断所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 中至少部分数据是否满足预定条件 C_z；

当所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据不满足所述预定条件C_z，从所述当前潜在分割点k_i沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，N*U不大于$\|B_z\|+\max_x(\|A_x\|)$，获得新的潜在分割点，执行步骤a)；

c)当所述当前潜在分割点k_i的M个窗口中的每一个窗口$W_{ix}[k_i-A_x,k_i+B_x]$中至少部分数据满足预定条件C_x，则所述当前潜在分割点k_i为数据流分割点。

结合第二方面，第一种可能的实现方式中，所述规则还包括：至少两个窗口$W_{ie}[k_i-A_e,k_i+B_e]$与$W_{if}[k_i-A_f,k_i+B_f]$，满足条件：$|A_e+B_e|=|A_f+B_f|$，$C_e=C_f$。

结合第二方面的第一种可能的实现方式，第二种可能的实现方式中，所述规则还包括：A_e和A_f为正整数。

结合第二方面的第一种可能的实现方式或第二种可能的实现方式，在第三种可能的实现方式中，所述规则还包括：$A_e-1=A_f$，$B_e+1=B_f$。

结合第二方面，或第二方面第一至第三任一可能的实现方式，第四种可能的实现方式中，判断所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据是否所述满足预定条件C_z，具体包括：

使用随机函数判断所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据是否满足所述预定条件C_z。

结合第二方面的第四种可能的实现方式，第五种可能的实现方式中，所述使用随机函数判断所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据是否满足所述预定条件C_z，具体为使用hash函数判断所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据是否满足所述预定条件C_z。

结合第二方面，或第二方面第一至第五任一可能的实现方式，第

六种可能的实现方式中，当所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 中至少部分数据不满足所述预定条件 C_z，从所述当前潜在分割点 k_i 沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，获得所述新的潜在分割点，根据所述规则，为所述新的潜在分割点确定的窗口 $W_{ic}[k_i-A_c, k_i+B_c]$ 的左边界与所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 的右边界重合或者为所述新的潜在分割点确定的所述窗口 $W_{ic}[k_i-A_c, k_i+B_c]$ 的左边界位于所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 范围之内；其中，为所述新的潜在分割点确定的所述窗口 $W_{ic}[k_i-A_c, k_i+B_c]$ 是根据所述规则，为所述新的潜在分割点确定的M个窗口按照数据流查找方向获得的序列中排序第一的窗口。

结合第二方面的第四种可能的实现方式，第七种可能的实现方式中，使用随机函数判断所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 中至少部分数据是否满足所述预定条件 C_z，具体包括：

在所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 中选择F个字节，将所述F个字节反复利用H次，共获得F*H个字节，其中每个字节由8位组成，记为 $a_{m,1}\ldots a_{m,8}$，表示所述F*H个字节中第m个字节的第1到第8位，所述F*H个字节对应的位可以表示为：$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,8} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ a_{F*H,1} & a_{F*H,2} & \cdots & a_{F*H,8} \end{pmatrix}$，当 $a_{m,n}$=1时，$V_{am,n}$=1，当 $a_{m,n}$=0时，$V_{am,n}$=-1，其中 $a_{m,n}$ 表示 $a_{m,1}\ldots a_{m,8}$ 中的任一个，所述F*H个字节对应的位按照 $a_{m,n}$ 与 $V_{am,n}$ 的转换关系得到矩阵 V_a，所述矩阵 V_a 表示为：$\begin{pmatrix} V_{a1,1} & V_{a1,2} & \cdots & V_{a1,8} \\ V_{a2,1} & V_{a2,2} & \cdots & V_{a2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{aF*H,1} & V_{aF*H,2} & \cdots & V_{aF*H,8} \end{pmatrix}$，从服务正态分布的随机数中选

择F*H*8个随机数组成矩阵R，所述矩阵R表示为：

$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,8} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{F*H,1} & h_{F*H,2} & \cdots & h_{F*H,8} \end{pmatrix}$，将所述矩阵 V_a 的第m行与所述矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为 $S_{am}=V_{am,1}*h_{m,1}+V_{am,2}*h_{m,2}+\ldots+V_{am,8}*h_{m,8}$，同理，获得 S_{a1}、S_{a2} …到 S_{aF*H}，统计 S_{a1}、S_{a2} …到 S_{aF*H} 中满足大于0的值的个数K，当K为偶数，则所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 中至少部分数据满足所述预定条件 C_z。

第三方面，本发明实施例提供了一种用于查找数据流分割点的服务器，所述服务器包括中央处理单元和主存储器，所述中央处理单元与所述主存储器通信，在所述服务器上预设有规则，所述规则为：为潜在分割点 k 确定M个点 p_x、点 p_x 对应的窗口 $W_x[p_x-A_x, p_x+B_x]$ 和窗口 $W_x[p_x-A_x, p_x+B_x]$ 对应的预定条件 C_x，其中，x为1到M连续的自然数，M≥2，A_x 和 B_x 为整数；

所述主存储器用于存储可执行指令，所述中央处理单元执行所述可执行指令，以执行如下步骤：

a)依据所述规则为当前潜在分割点 k_i 确定点 p_{iz} 及所述点 p_{iz} 对应的窗口 $W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$，i和z为整数，并且1≤z≤M；

b) 判断所述窗口 $W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$ 中至少部分数据是否满足预定条件 C_z；

当所述窗口 $W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$ 中至少部分数据不满足所述预定条件 C_z，从所述点 p_{iz} 沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，N*U不大于 $\|B_z\|+\max_x(\|A_x\|+\|(k_i-p_{ix})\|)$，获得新的潜在分割点，执行步骤a)；

c)当所述当前潜在分割点 k_i 的M个窗口中的每一个窗口 $W_{ix}[p_{ix}-$

$A_x, p_{ix}+B_x]$中至少部分数据满足预定条件C_x，则所述当前潜在分割点k_i为数据流分割点。

结合第三方面，第一种可能的实现方式中，所述规则还包括：至少两个点p_e和p_f，满足条件$A_e=A_f$，$B_e=B_f$，$C_e=C_f$。

结合第三方面的第一种可能的实现方式，第二种可能的实现方式中，所述规则还包括：所述至少两个点p_e和p_f，相对于所述潜在分割点k,在所述数据流分割点查找反方向上。

结合第三方面的第一种可能的实现方式或第二种可能的实现方式，第三种可能的实现方式中，所述规则还包括：所述至少两个点p_e和p_f之间的距离为1个U。

结合第三方面，或第一至第三任一可能的实现方式，第四种可能的实现方式中，所述中央处理单元具体用于

使用随机函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$。

结合第三方面的第四种可能的实现方式，第五种可能的实现方式中，所述中央处理单元具体用于使用hash函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$。

结合第三方面，或第一至第五任一可能的实现方式，第六种可能的实现方式中，当所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据不满足所述预定条件$C_z$，从所述点$p_{iz}$沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，获得所述新的潜在分割点，根据所述规则，为所述新的潜在分割点确定的点p_{ic}对应的窗口$W_{ic}[p_{ic}-A_c, p_{ic}+B_c]$的左边界与所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$的右边界重合或者为所述新的潜在分割点确定的所述点$p_{ic}$对应的所述窗口$W_{ic}[p_{ic}-A_c, p_{ic}+B_c]$

的左边界位于所述窗口 $W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$ 范围之内；其中，为所述新的潜在分割点确定的所述点 p_{ic} 是根据所述规则，为所述新的潜在分割点确定的M个点按照数据流查找方向获得的序列中排序第一的点。

结合第三方面的第四种可能的实现方式，第七种可能的实现方式中，所述中央处理单元使用随机函数判断所述窗口 $W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$ 中至少部分数据是否满足所述预定条件 C_z，具体包括：

在所述窗口 $W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$ 中选择F个字节，将所述F个字节反复利用H次，共获得F*H个字节，其中每个字节由8位组成，记为 $a_{m,1}\ldots a_{m,8}$，表示所述F*H个字节中第m个字节的第1到第8位，所述F*H个字节对应的位可以表示为：$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,8} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ a_{F*H,1} & a_{F*H,2} & \cdots & a_{F*H,8} \end{pmatrix}$，当 $a_{m,n}$=1时，$V_{am,n}$=1，当 $a_{m,n}$=0时，$V_{am,n}$=-1，其中 $a_{m,n}$ 表示 $a_{m,1}\ldots a_{m,8}$ 中的任一个，所述F*H个字节对应的位按照 $a_{m,n}$ 与 $V_{am,n}$ 的转换关系得到矩阵 V_a，所述矩阵 V_a 表示为：$\begin{pmatrix} V_{a1,1} & V_{a1,2} & \cdots & V_{a1,8} \\ V_{a2,1} & V_{a2,2} & \cdots & V_{a2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{aF*H,1} & V_{aF*H,2} & \cdots & V_{aF*H,8} \end{pmatrix}$，从服务正态分布的随机数中选择F*H*8个随机数组成矩阵R，所述矩阵R表示为：

$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,8} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{F*H,1} & h_{F*H,2} & \cdots & h_{F*H,8} \end{pmatrix}$，将所述矩阵 V_a 的第m行与所述矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为 $S_{am}=V_{am,1}*h_{m,1}+V_{am,2}*h_{m,2}+\ldots+V_{am,8}*h_{m,8}$，同理，获得 S_{a1}、$S_{a2}\ldots$到 S_{aF*H}，统计 S_{a1}、$S_{a2}\ldots$到 S_{aF*H} 中满足大于0的值的个数K，当K为偶数，则所述窗口 $W_{iz}[p_{iz}-A_z,$

$p_{iz}+B_z]$中至少部分数据满足所述预定条件C_z。第四方面，本发明实施例提供了一种用于查找数据流分割点的服务器，所述服务器包括中央处理单元和主存储器，所述中央处理单元与所述主存储器通信，在所述服务器上预设有规则，所述规则为：为潜在分割点k确定M个窗口$W_x[k-A_x, k+B_x]$和窗口$W_x[k-A_x, k+B_x]$对应的预定条件C_x，其中，x为1到M连续的自然数，M≥2，A_x和B_x为整数；

所述主存储器用于存储可执行指令，所述中央处理单元执行所述可执行指令，以执行以下步骤：

a)依据所述规则为当前潜在分割点k_i确定对应的窗口$W_{iz}[k_i-A_z, k_i+B_z]$，i和z为整数，并且1≤z≤M；

b)判断所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据是否满足预定条件C_z；

当所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据不满足所述预定条件C_z，从所述当前潜在分割点k_i沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，N*U不大于$\|B_z\|+\max_x(\|A_x\|)$，获得新的潜在分割点，执行步骤a)；

c)当所述当前潜在分割点k_i的M个窗口中的每一个窗口$W_{ix}[k_i-A_x, k_i+B_x]$中至少部分数据满足预定条件C_x，则所述当前潜在分割点k_i为数据流分割点。

结合第四方面，第一种可能的实现方式中，所述规则还包括：至少两个窗口$W_{ie}[k_i-A_e, k_i+B_e]$与$W_{if}[k_i-A_f, k_i+B_f]$，满足条件：$|A_e+B_e|=|A_f+B_f|$，$C_e=C_f$。

结合第四方面的第一种可能的实现方式，第二种可能的实现方式中，所述规则还包括：A_e和A_f为正整数。

结合第四方面的第一种可能的实现方式或第二种可能的实现方

式，在第三种可能的实现方式中，所述规则还包括：A_e-1=A_f，B_e+1=B_f。

结合第四方面，或第一至第三任一可能的实现方式，第四种可能的实现方式中，所述中央处理单元具体用于

使用随机函数判断所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据是否满足所述预定条件C_z。

结合第四方面的第四种可能的实现方式，第五种可能的实现方式中，所述中央处理单元具体用于使用hash函数判断所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据是否满足所述预定条件C_z。

结合第四方面，或第一至第五任一可能的实现方式，第六种可能的实现方式中，当所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据不满足所述预定条件C_z，从所述当前潜在分割点k_i沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，获得所述新的潜在分割点，根据所述规则，为所述新的潜在分割点确定的窗口$W_{ic}[k_i-A_c,k_i+B_c]$的左边界与所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$的右边界重合或者为所述新的潜在分割点确定的所述窗口$W_{ic}[k_i-A_c,k_i+B_c]$的左边界位于所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$范围之内；其中，为所述新的潜在分割点确定的所述窗口$W_{ic}[k_i-A_c,k_i+B_c]$是根据所述规则，为所述新的潜在分割点确定的M个窗口按照数据流查找方向获得的序列中排序第一的窗口。

结合第四方面的第四种可能的实现方式，第七种可能的实现方式中，所述中央处理单元使用随机函数判断所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据是否满足所述预定条件C_z，具体包括：

在所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中选择F个字节，将所述F个字节反复利用H次，共获得F*H个字节，其中每个字节由8位组成，记为$a_{m,1}\ldots a_{m,8}$，表示所述F*H个字节中第m个字节的第1到第8位，所述F*H个字

节对应的位可以表示为：$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,8} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ a_{F*H,1} & a_{F*H,2} & \cdots & a_{F*H,8} \end{pmatrix}$，当 $a_{m,n}$=1时，$V_{am,n}$=1，当 $a_{m,n}$=0时，$V_{am,n}$=-1，其中 $a_{m,n}$ 表示 $a_{m,1}\ldots a_{m,8}$ 中的任一个，所述F*H个字节对应的位按照 $a_{m,n}$ 与 $V_{am,n}$ 的转换关系得到矩阵 V_a，所述矩阵 V_a 表示为：$\begin{pmatrix} V_{a1,1} & V_{a1,2} & \cdots & V_{a1,8} \\ V_{a2,1} & V_{a2,2} & \cdots & V_{a2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{aF*H,1} & V_{aF*H,2} & \cdots & V_{aF*H,8} \end{pmatrix}$，从服务正态分布的随机数中选择F*H*8个随机数组成矩阵R，所述矩阵R表示为：

$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,8} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{F*H,1} & h_{F*H,2} & \cdots & h_{F*H,8} \end{pmatrix}$，将所述矩阵 V_a 的第m行与所述矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为 $S_{am}=V_{am,1}*h_{m,1}+V_{am,2}*h_{m,2}+\ldots+V_{am,8}*h_{m,8}$，同理，获得 S_{a1}、S_{a2} …到 S_{aF*H}，统计 S_{a1}、S_{a2} …到 S_{aF*H} 中满足大于0的值的个数K，当K为偶数，则所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 中至少部分数据满足所述预定条件 C_z。

第五方面，本发明实施例提供了一种用于查找数据流分割点的服务器，在所述服务器上预设有规则，所述规则为：为潜在分割点 k 确定M个点 p_x、点 p_x 对应的窗口 $W_x[p_x-A_x, p_x+B_x]$ 和窗口 $W_x[p_x-A_x, p_x+B_x]$ 对应的预定条件 C_x，其中，x为1到M连续的自然数，M≥2，A_x 和 B_x 为整数；

所述服务器包括：处理单元，用于执行步骤a）：

a)依据所述规则为当前潜在分割点 k_i 确定点 p_{iz} 及所述点 p_{iz} 对应的窗口 $W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$，i和z为整数，并且1≤z≤M；

判断处理单元，用于判断所述窗口 $W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$ 中至少部分

数据是否满足预定条件C_z；

当所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据不满足所述预定条件$C_z$，从所述点$p_{iz}$沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，N*U不大于$\| B_z \| + \max_x(\| A_x \| + \| (k_i - p_{ix}) \|)$，获得新的潜在分割点，则所述确定单元为所述新的潜在分割点执行步骤a)；

当所述当前潜在分割点k_i的M个窗口中的每一个窗口$W_{ix}[p_{ix}-A_x, p_{ix}+B_x]$中至少部分数据满足预定条件$C_x$，则所述当前潜在分割点$k_i$为数据流分割点。

结合第五方面，第一种可能的实现方式中，所述规则还包括：至少两个点p_e和p_f，满足条件$A_e=A_f$，$B_e=B_f$，$C_e=C_f$。

结合第五方面的第一种可能的实现方式，第二种可能的实现方式中，所述规则还包括：所述至少两个点p_e和p_f，相对于所述潜在分割点k,在所述数据流分割点查找反方向上。

结合第五方面的第一种可能的实现方式或第二种可能的实现方式，在第三种可能的实现方式中，所述规则还包括：所述至少两个点p_e和p_f之间距离为1个U。

结合第五方面，或第一至第三任一可能的实现方式，第四种可能的实现方式中，所述判断处理单元具体使用随机函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$。

结合第五方面的第四种可能的实现方式，第五种可能的实现方式中，所述判决处理单元具体用于使用hash函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$。

结合第五方面，或第一至第五任一可能的实现方式，第六种可能的实现方式中，所述判断处理单元用于当所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$

中至少部分数据不满足所述预定条件C_z，从所述点p_{iz}沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，获得所述新的潜在分割点，所述确定单元为所述新的潜在分割点执行步骤a），根据所述规则，为所述新的潜在分割点确定的点p_{ic}对应的窗口$W_{ic}[p_{ic}-A_c, p_{ic}+B_c]$的左边界与所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$的右边界重合或者为所述新的潜在分割点确定的所述窗口$W_{ic}[p_{ic}-A_c, p_{ic}+B_c]$的左边界位于所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$范围之内；其中，为所述新的潜在分割点确定的所述窗口$W_{ic}[p_{ic}-A_c, p_{ic}+B_c]$是根据所述规则，为所述新的潜在分割点确定的M个点按照数据流查找方向获得的序列中排序第一的点。

结合第五方面的第四种可能的实现方式，第七种可能的实现方式中，所述判断处理单元具体用于使用随机函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$，具体包括：

在所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中选择F个字节，将所述F个字节反复利用H次，共获得F*H个字节，其中每个字节由8位组成，记为$a_{m,1}\ldots a_{m,8}$，表示所述F*H个字节中第m个字节的第1到第8位，所述F*H个字节对应的位可以表示为：$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,8} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ a_{F*H,1} & a_{F*H,2} & \cdots & a_{F*H,8} \end{pmatrix}$，当$a_{m,n}$=1时，$V_{am,n}$=1，当$a_{m,n}$=0时，$V_{am,n}$=-1，其中$a_{m,n}$表示$a_{m,1}\ldots a_{m,8}$中的任一个，所述F*H个字节对应的位按照$a_{m,n}$与$V_{am,n}$的转换关系得到矩阵$V_a$，所述矩阵$V_a$表示为：$\begin{pmatrix} V_{a1,1} & V_{a1,2} & \cdots & V_{a1,8} \\ V_{a2,1} & V_{a2,2} & \cdots & V_{a2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{aF*H,1} & V_{aF*H,2} & \cdots & V_{aF*H,8} \end{pmatrix}$，从服务正态分布的随机数中选择F*H*8个随机数组成矩阵R，所述矩阵R表示为：

$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,8} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{F*H,1} & h_{F*H,2} & \cdots & h_{F*H,8} \end{pmatrix}$，将所述矩阵$V_a$的第m行与所述矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为$S_{am}=V_{am,1}*h_{m,1}+V_{am,2}*h_{m,2}+\ldots+V_{am,8}*h_{m,8}$，同理，获得$S_{a1}$、$S_{a2}$...到$S_{aF*H}$，统计$S_{a1}$、$S_{a2}$...到$S_{aF*H}$中满足大于0的值的个数K，当K为偶数，则所述窗口$W_{iz}$[$p_{iz}-A_z$，$p_{iz}+B_z$]中至少部分数据满足所述预定条件$C_z$。

第六方面，本发明实施例提供了一种用于查找数据流分割点的服务器，在所述服务器上预设有规则，所述规则为：为潜在分割点k确定M个窗口$W_x[k-A_x,k+B_x]$和窗口$W_x[k-A_x,k+B_x]$对应的预定条件C_x，其中，x为1到M连续的自然数，M≥2，A_x和B_x为整数；

所述服务器包括：确定单元，用于执行步骤a：

a)依据所述规则为当前潜在分割点k_i确定对应的窗口$W_{iz}[k_i-A_z,k_i+B_z]$，i和z为整数，并且1≤z≤M；

判断处理单元，用于判断所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据是否满足预定条件C_z；

当所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据不满足所述预定条件C_z，从所述当前潜在分割点k_i沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，N*U不大于$\|B_z\|+\max_x(\|A_x\|)$，获得新的潜在分割点，执行步骤a)；

c当所述当前潜在分割点k_i的M个窗口中的每一个窗口$W_{ix}[k_i-A_x,k_i+B_x]$中至少部分数据满足预定条件C_x，则所述当前潜在分割点k_i为数据流分割点。

结合第六方面，第一种可能的实现方式中，所述规则还包括：至

少两个窗口$W_{ie}[k_i-A_e,k_i+B_e]$与$W_{if}[k_i-A_f,k_i+B_f]$，满足条件：$|A_e+B_e|=|A_f+B_f|$，$C_e=C_f$。

结合第六方面的第一种可能的实现方式，第二种可能的实现方式中，所述规则还包括：A_e和A_f为正整数。

结合第六方面的第一种可能的实现方式或第二种可能的实现方式，在第三种可能的实现方式中，所述规则还包括：$A_e-1=A_f$，$B_e+1=B_f$。

结合第六方面，或第一至第三任一可能的实现方式，第四种可能的实现方式中，所述判断处理单元具体用于

使用随机函数判断所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据是否满足所述预定条件C_z。

结合第六方面的第四种可能的实现方式，第五种可能的实现方式中，所述判断处理单元具体用于使用hash函数判断所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据是否满足所述预定条件C_z。

结合第六方面，或第一至第五任一可能的实现方式，第六种可能的实现方式中，所述判断处理单元用于当所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据不满足所述预定条件C_z，从所述当前潜在分割点k_i沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，获得所述新的潜在分割点，所述确定单元为所述新的潜在分割点执行步骤a），根据所述规则，为所述新的潜在分割点确定的窗口$W_{ic}[k_i-A_c,k_i+B_c]$的左边界与所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$的右边界重合或者为所述新的潜在分割点确定的所述窗口$W_{ic}[k_i-A_c,k_i+B_c]$的左边界位于所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$范围之内；其中，为所述新的潜在分割点确定的所述窗口$W_{ic}[k_i-A_c,k_i+B_c]$是根据所述规则，为所述新的潜在分割点确定的M个窗口按照数据流查找方向获得的序列中排序第一

的窗口。

结合第六方面的第四种可能的实现方式，第七种可能的实现方式中，所述判断处理单元使用随机函数判断所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 中至少部分数据是否满足所述预定条件 C_z，具体包括:

在所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 中选择F个字节，将所述F个字节反复利用H次，共获得F*H个字节，其中每个字节由8位组成，记为 $a_{m,1}\ldots a_{m,8}$，表示所述F*H个字节中第m个字节的第1到第8位，所述F*H个字节对应的位可以表示为: $\begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,8} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ a_{\mathrm{F}*H,1} & a_{F*H,2} & \ldots & a_{F*H,8} \end{pmatrix}$，当 $a_{m,n}$=1时，$V_{am,n}$=1，当 $a_{m,n}$=0时，$V_{am,n}$=-1，其中 $a_{m,n}$ 表示 $a_{m,1}\ldots a_{m,8}$ 中的任一个，所述F*H个字节对应的位按照 $a_{m,n}$ 与 $V_{am,n}$ 的转换关系得到矩阵 V_a，所述矩阵 V_a 表示为: $\begin{pmatrix} V_{a1,1} & V_{a1,2} & \cdots & V_{a1,8} \\ V_{a2,1} & V_{a2,2} & \cdots & V_{a2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{a\mathrm{F}*H,1} & V_{aF*H,2} & \cdots & V_{aF*H,8} \end{pmatrix}$，从服务正态分布的随机数中选择F*H*8个随机数组成矩阵R，所述矩阵R表示为:

$\begin{pmatrix} \mathrm{h}_{1,1} & h_{1,2} & \ldots & h_{1,8} \\ h_{2,1} & h_{2,2} & \ldots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{\mathrm{F}*H,1} & h_{F*H,2} & \ldots & h_{F*H,8} \end{pmatrix}$，将所述矩阵 V_a 的第m行与所述矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为 $S_{am}=V_{am,1}*h_{m,1}+V_{am,2}*h_{m,2}+\ldots+V_{am,8}*h_{m,8}$，同理，获得 S_{a1}、$S_{a2}\ldots$到 S_{aF*H}，统计 S_{a1}、$S_{a2}\ldots$到 S_{aF*H} 中满足大于0的值的个数K，当K为偶数，则所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 中至少部分数据满足所述预定条件 C_z。

第七方面，本发明实施例提供了一种计算机可读存储介质，所述计算机可读存储介质用于存储可执行指令，服务器执行所述可执行指

令以查找数据流分割点，在所述服务器上预设有规则，所述规则为：为潜在分割点k确定M个点p_x、点p_x对应的窗口$W_x[p_x-A_x, p_x+B_x]$和窗口$W_x[p_x-A_x, p_x+B_x]$对应的预定条件C_x，其中，x为1到M连续的自然数，M≥2，A_x和B_x为整数；

当所述服务器执行所述可执行指令，以执行以下步骤：

a)依据所述规则为当前潜在分割点k_i确定点p_{iz}及所述点p_{iz}对应的窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$，i和z为整数，并且1≤z≤M；

b) 判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足预定条件$C_z$；

当所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据不满足所述预定条件$C_z$，从所述点$p_{iz}$沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，N*U不大于$\| B_z \| + \max_x(\| A_x \| + \| (k_i - p_{ix}) \|)$，获得新的潜在分割点，执行步骤a)；

c)当所述当前潜在分割点k_i的M个窗口中的每一个窗口$W_{ix}[p_{ix}-A_x, p_{ix}+B_x]$中至少部分数据满足预定条件$C_x$，则所述当前潜在分割点$k_i$为数据流分割点。

结合第七方面，第一种可能的实现方式中，所述规则还包括：至少两个点p_e和p_f，满足条件$A_e=A_f$，$B_e=B_f$，$C_e=C_f$。

结合第七方面的第一种可能的实现方式，第二种可能的实现方式中，所述规则还包括：所述至少两个点p_e和p_f，相对于所述潜在分割点k,在所述数据流分割点查找反方向上。

结合第七方面的第一种可能的实现方式或第二种可能的实现方式，在第三种可能的实现方式中，所述规则还包括：所述至少两个点p_e和p_f之间的距离为1个U。

结合第七方面，或第七方面第一至第三任一可能的实现方式,第四种可能的实现方式中，所述服务器判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$，具体包括:

所述服务器使用随机函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$。

结合第七方面的第四种可能的实现方式，第五种可能的实现方式中，所述服务器使用随机函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$，具体包括:

所述服务器使用hash函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$。

结合第七方面，或第七方面第一至第五任一可能的实现方式，第六种可能的实现方式中，当所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据不满足所述预定条件$C_z$，从所述点$p_{iz}$沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，获得所述新的潜在分割点，根据所述规则，为所述新的潜在分割点确定的点p_{ic}对应的窗口$W_{ic}[p_{ic}-A_c, p_{ic}+B_c]$的左边界与所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$的右边界重合或者为所述新的潜在分割点确定的所述点$p_{ic}$对应的所述窗口$W_{ic}[p_{ic}-A_c, p_{ic}+B_c]$的左边界位于所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$范围之内；其中，为所述新的潜在分割点确定的所述点$p_{ic}$是根据所述规则，为所述新的潜在分割点确定的M个点按照数据流查找方向获得的序列中排序第一的点。

结合第七方面的第四种可能的实现方式，第七种可能的实现方式中，使用随机函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$，具体包括:

在所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中选择F个字节，将所述F个字节

反复利用H次，共获得F*H个字节，其中每个字节由8位组成，记为 $a_{m,1}\ldots a_{m,8}$，表示所述F*H个字节中第m个字节的第1到第8位，所述F*H个字节对应的位可以表示为：$\begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,8} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ a_{F*H,1} & a_{F*H,2} & \ldots & a_{F*H,8} \end{pmatrix}$，当 $a_{m,n}$=1时，$V_{am,n}$=1，当 $a_{m,n}$=0时，$V_{am,n}$=-1，其中 $a_{m,n}$ 表示 $a_{m,1}\ldots a_{m,8}$ 中的任一个，所述F*H个字节对应的位按照 $a_{m,n}$ 与 $V_{am,n}$ 的转换关系得到矩阵 V_a，所述矩阵 V_a 表示为：$\begin{pmatrix} V_{a1,1} & V_{a1,2} & \cdots & V_{a1,8} \\ V_{a2,1} & V_{a2,2} & \cdots & V_{a2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{aF*H,1} & V_{aF*H,2} & \cdots & V_{aF*H,8} \end{pmatrix}$，从服务正态分布的随机数中选择F*H*8个随机数组成矩阵R，所述矩阵R表示为：

$\begin{pmatrix} h_{1,1} & h_{1,2} & \ldots & h_{1,8} \\ h_{2,1} & h_{2,2} & \ldots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{F*H,1} & h_{F*H,2} & \ldots & h_{F*H,8} \end{pmatrix}$，将所述矩阵 V_a 的第m行与所述矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为 $S_{am}=V_{am,1}*h_{m,1}+V_{am,2}*h_{m,2}+\ldots+V_{am,8}*h_{m,8}$，同理，获得 S_{a1}、S_{a2} …到 S_{aF*H}，统计 S_{a1}、S_{a2} …到 S_{aF*H} 中满足大于0的值的个数K，当K为偶数，则所述窗口 $W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$ 中至少部分数据满足所述预定条件 C_z。

第八方面，本发明实施例提供了一种计算机可读存储介质，所述计算机可读存储介质用于存储可执行指令，服务器执行所述可执行指令以查找数据流分割点，在所述服务器上预设有规则，所述规则为：为潜在分割点 k 确定M个窗口 $W_x[k-A_x, k+B_x]$ 和窗口 $W_x[k-A_x, k+B_x]$ 对应的预定条件 C_x，其中，x为1到M连续的自然数，M≥2，A_x 和 B_x 为整数；当所述服务器执行所述可执行指令，以执行以下步骤：

a)依据所述规则为当前潜在分割点k_i确定对应的窗口$W_{iz}[k_i-A_z,k_i+B_z]$，i和z为整数，并且1≤z≤M；

b)判断所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据是否满足预定条件C_z；

当所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据不满足所述预定条件C_z，从所述当前潜在分割点k_i沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，N*U不大于$\| B_z \|+\max_x(\| A_x \|)$，获得新的潜在分割点，执行步骤a)；

c)当所述当前潜在分割点k_i的M个窗口中的每一个窗口$W_{ix}[k_i-A_x,k_i+B_x]$中至少部分数据满足预定条件C_x，则所述当前潜在分割点k_i为数据流分割点。

结合第八方面，第一种可能的实现方式中，所述规则还包括：至少两个窗口$W_{ie}[k_i-A_e,k_i+B_e]$与$W_{if}[k_i-A_f,k_i+B_f]$，满足条件：$|A_e+B_e|=|A_f+B_f|$，$C_e=C_f$。

结合第八方面的第一种可能的实现方式，第二种可能的实现方式中，所述规则还包括：A_e和A_f为正整数。

结合第八方面的第一种可能的实现方式或第二种可能的实现方式，在第三种可能的实现方式中，所述规则还包括：$A_e-1=A_f$，$B_e+1=B_f$。

结合第八方面，或第八方面第一到第三任一可能的实现方式，第四种可能的实现方式中，所述服务器判断所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据是否满足所述预定条件C_z，具体包括：

使用随机函数判断所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据是否满足所述预定条件C_z。

结合第八方面的第四种可能的实现方式，第五种可能的实现方式

中，所述服务器使用随机函数判断所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据是否满足所述预定条件C_z，具体为所述服务器使用hash函数判断所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据是否满足所述预定条件C_z。

结合第八方面，或第八方面第一到第五任一可能的实现方式，第六种可能的实现方式中，当所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据不满足所述预定条件C_z，从所述当前潜在分割点k_i沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，获得所述新的潜在分割点，根据所述规则，为所述新的潜在分割点确定的窗口$W_{ic}[k_i-A_c,k_i+B_c]$的左边界与所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$的右边界重合或者为所述新的潜在分割点确定的所述窗口$W_{ic}[k_i-A_c,k_i+B_c]$的左边界位于所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$范围之内；其中，为所述新的潜在分割点确定的所述窗口$W_{ic}[k_i-A_c,k_i+B_c]$是根据所述规则，为所述新的潜在分割点确定的M个窗口按照数据流查找方向获得的序列中排序第一的窗口。

结合第八方面的第四种可能的实现方式，第七种可能的实现方式中，使用随机函数判断所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据是否满足所述预定条件C_z，具体包括：

在所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中选择F个字节，将所述F个字节反复利用H次，共获得F*H个字节，其中每个字节由8位组成，记为$a_{m,1}\ldots a_{m,8}$，表示所述F*H个字节中第m个字节的第1到第8位，所述F*H个字节对应的位可以表示为：$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,8} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ a_{F*H,1} & a_{F*H,2} & \cdots & a_{F*H,8} \end{pmatrix}$，当$a_{m,n}$=1时，$V_{am,n}$

=1，当 $a_{m,n}$=0时， $V_{am,n}$=-1，其中 $a_{m,n}$ 表示 $a_{m,1}$... $a_{m,8}$ 中的任一个，所述F*H个字节对应的位按照 $a_{m,n}$ 与 $V_{am,n}$ 的转换关系得到矩阵 V_a，所述矩阵 V_a 表示为：$\begin{pmatrix} V_{a1,1} & V_{a1,2} & \cdots & V_{a1,8} \\ V_{a2,1} & V_{a2,2} & \cdots & V_{a2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{aF*H,1} & V_{aF*H,2} & \cdots & V_{aF*H,8} \end{pmatrix}$，从服务正态分布的随机数中选择F*H*8个随机数组成矩阵R，所述矩阵R表示为：

$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,8} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{F*H,1} & h_{F*H,2} & \cdots & h_{F*H,8} \end{pmatrix}$，将所述矩阵 V_a 的第m行与所述矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为 $S_{am}=V_{am,1}*h_{m,1}+V_{am,2}*h_{m,2}+\ldots+V_{am,8}*h_{m,8}$，同理，获得 S_{a1}、 S_{a2} ...到 S_{aF*H}，统计 S_{a1}、 S_{a2} ...到 S_{aF*H} 中满足大于0的值的个数K，当K为偶数，则所述窗口 W_{iz} $[k_i-A_z, k_i+B_z]$ 中至少部分数据满足所述预定条件 C_z。

本发明实施例中通过判断M个窗口中某一个窗口中至少部分数据是否满足预定条件，来查找数据流分割点，当某一个窗口中至少部分数据不满足预定条件，则跳过N*U个长度，获得下一个潜在分割点，提高了数据流分割点查找效率。

附图说明

图1为本发明实施例一种应用场景示意图；

图2为数据流分割点示意图；

图3为查找数据流分割点示意图；

图4为本发明实施例方法示意图；

图5和图6为查找数据流分割点实施方式示意图；

图7和图8为查找数据流分割点实施方式示意图；

图9和图10为找数据流分割点实施方式示意图；

图11和图12和图13为找数据流分割点实施方式示意图;

图14和图15为找数据流分割点实施方式示意图;

图16和图17为判断窗口中至少部分数据是否满足预定条件示意图;

图18为去重服务器结构图;

图19为去重服务器结构图;

图20为本发明实施例方法示意图;

图21和图22为查找数据流分割点实施方式示意图;

图23和图24为查找数据流分割点实施方式示意图;

图25和图26为找数据流分割点实施方式示意图;

图27和图28和图29为找数据流分割点实施方式示意图;

图30和图31为找数据流分割点实施方式示意图;

图32和图33为判断窗口中至少部分数据是否满足预定条件示意图。

具体实施例

随着存储技术的不断进步，数据产生量也在不断增加，大量的数据对存储容量提出了最高的要求。存储容量增加的同时，也增加了IT设备采购成本，为了缓解数据量与存储容量之间的需求矛盾，节约IT设备采购成本，在数据存储领域引入了重复数据删除技术。

本发明实施例一种使用场景为数据备份场景。数据备份是为防止各种原因导致的数据丢失，通过备份服务器将数据备份到其他存储介质的过程。如图1所示的数据备份系统架构。数据备份系统包括客户端（101a、101b...101n）、备份服务器102、重复数据删除服务器(简称去重服务器或重删服务器)103和存储设备（104a、104b...104n）。其中客户端（101a、101b...101n）可以为应用服务器、工作站等；备份服务器102用于备份客户端生成的数据；去重服务器103用于执行备份数据的重复数据删除任务；存储设备（104a、104b...104n）作为存储重复数据删除后的数据的存储介质，可以为磁盘阵列、磁带库等存

储介质。客户端(101a、101b...101n）、备份服务器102、重复数据删除服务器103和存储设备（104a、104b...104n）可以通过交换机、局域网、互联网、光纤等方式连接，上述设备可以位于同一地点，也可以位于不同地点。备份服务器102、重删服务器103、存储设备（104a、104b...104n）可以为独立的物理设备，或者在具体实现中物理上集成为一体，或者备份服务器102与重删服务器103集成为一体，或者重删服务器103与存储设备（104a、104b...104n）集成为一体等。

去重服务器103对备份数据的数据流执行重复数据删除操作，一般包括以下步骤:

1）数据流分割点查找：根据特定算法在数据流中查找数据流分割点;

2）根据查找到的数据流分割点划分数据块;

3）计算数据块的特征值：计算数据块的特征值作为标识该数据块的特征;将计算得到的特征值添加到该数据流对应的文件的数据块的特征列表中；一般利用SHA-1或MD5算法计算数据块的特征值;

4）相同数据块检测：将计算得到的数据块的特征值与数据块特征列表中已存在的特征值进行比对以确定是否存在相同数据块;

5）删除重复数据块：通过相同数据块检测，如果发现数据块特征列表中存在与该数据块相同的特征值，则不需要再存储该数据块或者根据备份策略确定的重复数据块存储数量决定是否存储该数据块。

通过去重服务器103对备份数据的数据流执行重复数据删除操作的步骤可知，数据流分割点查找作为重复数据删除操作的关键步骤，直接决定了重复数据删除的性能。

本发明实施例中，去重服务器103接收备份服务器102发送的备份文件，对该文件执行重复数据删除处理。通常待处理备份文件在去重服务器103中以数据流形式呈现，去重服务器103查找数据流中的分割

点时，通常要确定数据流分割点最小查找单位，具体如图2所示，如潜在分割点k_1位于序号分别为1和2的连续两个数据流分割点最小查找单位之间，潜在分割点是指需要进行判断是否可以作为数据流分割点的点；当点k_1为一个数据流分割点，数据流分割点查找方向如图2中箭头所示，查找下一个潜在分割点为k_7，即位于序号分别为7和8的连续两个数据流分割点最小查找单位之间，当潜在分割点k_7为数据流分割点，则相邻的两个数据流分割点k_1、k_7之间的数据为1个数据块。数据流分割点最小查找单位具体可以根据实际需要确定，这里以1个字节（Byte）为例，即序号为1、2、7和8的数据流分割点最小查找单位大小均为1个字节。如图2所示的数据流分割点查找方向通常表示由文件头向文件尾方向查找，或者由文件尾向文件头方向，本实施例中以从文件头向文件尾方向查找为例。

在重复数据删除场景，通常数据块越小，重复数据删除率越高，越容易查找到重复数据块，但是由此生成的元数据数量越大，而且数据块小到一定程度之后，重复数据删除率就不会增加了，但是元数据数量却会急剧增加。因此，必须控制数据块大小，实际应用中，通常会设定数据块的最小值，例如4KB（4096个字节），同时考虑到重复数据删除率，也会设定数据块的最大值，即数据块大小不能超过最大值，例如12KB（12288个字节）。一种具体实现方式如图3所示，去重服务器103在沿着箭头所示方向查找数据流分割点，k_a为当前查找到的数据流分割点，从k_a向数据流分割点查找方向查找下一个潜在分割点，为满足最小数据块要求，通常会从数据流分割点开始沿着数据流分割点查找方向跳过最小数据块大小，从最小数据块结束位置开始查找，也就是将最小数据块结束位置作为下一个潜在分割点k_i。在本发明实施例中，可以先从k_a点沿数据流分割点查找方向跳跃最小数据

块4KB，即4*1024=4096字节。从k_a点沿数据流分割点查找方向跳跃4096个字节，在第4096个字节的结束位置获得点k_i，作为潜在分割点，例如k_i位于序号分别为4096和4097的连续两个数据流分割点最小查找单位之间。仍然以图3为例， k_a为当前查找到的数据流分割点，沿如图3所示方向查找下一个数据流分割点，如果超过数据块最大值仍然没有找到下一个数据流分割点，则在从k_a点开始向数据流分割点查找方向达到数据块最大值的点k_z作为下一个数据流分割点，进行强制分割。

本发明实施例提供一种基于去重服务器查找数据流分割点的方法，如图4所示，包括:

在去重服务器103上预设有规则，所述规则为：为潜在分割点k确定M个点p_x、点p_x对应的窗口$W_x[p_x-A_x, p_x+B_x]$和窗口$W_x[p_x-A_x, p_x+B_x]$对应的预定条件C_x，其中， x为1到M连续的自然数，M≥2，A_x和B_x为整数；其中，p_x与潜在分割点k之间距离d_x个数据流分割点最小查找单位，数据流分割点最小查找单位以U表示，本实施例中U=1个字节，。在图3所示的实现方式中，关于M的取值，其中一种实现方式，M*U取值不大于预设的两个相邻的数据流分割点之间的最大距离，即预设的数据块最大长度。判断点p_z对应的窗口$W_z[p_z-A_z, p_z+B_z]$中至少部分数据是否满足预定条件C_z，其中，z为整数，1≤z≤M，（p_z-A_z）与（p_z+B_z）分别表示窗口W_z的两个边界。当判断任意一个点p_z的窗口$W_z[p_z-A_z, p_z+B_z]$中至少部分数据不满足预定条件C_z，则从不满足预定条件的窗口$W_z[p_z-A_z, p_z+B_z]$对应的点p_z沿数据流分割点查找方向跳跃N个字节，$N \le \| B_z \| + \max_x(\| A_x \| + \| (k-p_x) \|)$。其中，$\| (k-p_x) \|$表示M个点$p_x$中任一个点与潜在分割点$k$之间的距离，$\max_x(\| A_x \| + \| (k-p_x) \|)$表示M个点$p_x$中任一个点与潜在分割点$k$

之间的距离及该点对应的A_x的绝对值之和的最大值；$\| B_z \|$表示W_z[$p_z - A_z$，$p_z + B_z$]中B_z的绝对值，将在下面实施例中具体介绍N取值的原理。当判断M个窗口中的每一个窗口W_x[$p_x - A_x$，$p_x + B_x$]中至少部分数据满足预定条件C_x，则潜在分割点k为数据流分割点。

具体的，对当前潜在分割点k_i，依据所述规则，执行以下步骤：

步骤401：依据所述规则为当前潜在分割点k_i确定点p_{iz}及所述点p_{iz}对应的窗口W_{iz}[$p_{iz} - A_z, p_{iz} + B_z$]，i和z为整数，并且1≤z≤M；

步骤402：判断所述窗口W_{iz}[$p_{iz} - A_z, p_{iz} + B_z$]中至少部分数据是否满足预定条件$C_z$；

当所述窗口W_{iz}[$p_{iz} - A_z, p_{iz} + B_z$]中至少部分数据不满足所述预定条件$C_z$，从所述点$p_{iz}$沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，N*U不大于$\| B_z \| + \max_x(\| A_x \| + \| (k_i - p_{ix}) \|)$，获得新的潜在分割点，执行步骤401；

当所述当前潜在分割点k_i的M个窗口中的每一个窗口W_{ix}[$p_{ix} - A_x$，$p_{ix} + B_x$]中至少部分数据满足预定条件C_x，则所述当前潜在分割点k_i为数据流分割点。

进一步地，所述规则还包括：至少两个点p_e和p_f，满足条件$A_e = A_f$，$B_e = B_f$，$C_e = C_f$；

所述规则还包括：所述至少两个点p_e和p_f，相对于所述潜在分割点k，在所述数据流分割点查找反方向上。

所述规则还包括：所述至少两个点p_e和p_f之间的距离为1个U。

判断所述窗口W_{iz}[$p_{iz} - A_z, p_{iz} + B_z$]中至少部分数据是否满足所述预定条件$C_z$，具体包括：

使用随机函数判断所述窗口W_{iz}[$p_{iz} - A_z, p_{iz} + B_z$]中至少部分数据是

否满足所述预定条件C_z。

所述使用随机函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$，具体为使用hash函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$。

当所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据不满足所述预定条件$C_z$，从所述点$p_{iz}$沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，获得所述新的潜在分割点，根据所述规则，为所述新的潜在分割点确定的点p_{ic}对应的窗口$W_{ic}[p_{ic}-A_c, p_{ic}+B_c]$的左边界与所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$的右边界重合或者为所述新的潜在分割点确定的所述点$p_{ic}$对应的所述窗口$W_{ic}[p_{ic}-A_c, p_{ic}+B_c]$的左边界位于所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$范围之内；其中，为所述新的潜在分割点确定的所述点$p_{ic}$是根据所述规则，为所述新的潜在分割点确定的M个点按照数据流查找方向获得的序列中排序第一的点。

本发明实施例中通过判断M个窗口中某一个窗口中至少部分数据是否满足预定条件，来查找数据流分割点，当某一个窗口中至少部分数据不满足预定条件，则跳过N*U个长度，其中，N*U不大于$\| B_z \| + \max_x(\| A_x \| + \| (k_i - p_{ix}) \|)$，获得下一个潜在分割点，提高了数据流分割点查找效率。

在重复数据删除过程中，为保证数据块大小均匀，会考虑平均数据块（也称为平均分块）大小，即在满足最小数据块大小和最大数据块大小限定的同时，会确定平均数据块大小，以保证获得的数据块大小均匀。点p_x个数M与点p_x对应的窗口$W_x[p_x-A_x, p_x+B_x]$中至少部分数据满足预定条件C_x的概率，这两个因素决定了找到数据流分割点的概率（以P(n)表示）。前者影响跳跃的长度，后者影响跳跃的概率，二者共同影响平均分块大小。一般而言，在平均分块大小固定时，点

p_x个数M增加，则单个点p_x对应的窗口$W_x[p_x-A_x，p_x+B_x]$中至少部分数据满足预定条件C_x的概率也增加，例如在去重服务器103上预设的规则为：为潜在分割点k确定11个点p_x，x分别为1到11连续的自然数，11个点中任一个点p_x对应的窗口$W_x[p_x-A_x，p_x+B_x]$中至少部分数据满足预定条件C_x的概率为1/2。而在去重服务器103上预设的另一组规则为：为潜在分割点k选择的24个点p_x，x分别为1到24连续的自然数，24个点中任一个点p_x对应的窗口$W_x[p_x-A_x，p_x+B_x]$中至少部分数据满足预定条件C_x的概率3/4。具体窗口$W_x[p_x-A_x，p_x+B_x]$中至少部分数据满足预定条件C_x的概率设定可参见判断窗口$W_x[p_x-A_x，p_x+B_x]$中至少部分数据是否满足预定条件C_x部分的描述。点p_x个数M与点p_x对应的窗口$W_x[p_x-A_x，p_x+B_x]$中至少部分数据满足预定条件C_x的概率这两个因素决定P(n)，P（n）表示：从数据流起始位置/上一数据流分割点查找n个数据流分割点最小查找单位后没找到数据流分割点的概率。关于这两个因素决定P（n）的计算过程，实际上是一个多步长Fibonacci数列，后面将具体描述。得到P（n）后，1-P(n)即为数据流分割点的分布函数，(1-P(n))-(1-P(n-1))=P(n-1)-P(n)，即为在第n个点找到数据流分割点的概率，也就是数据流分割点的密度函数，根据数据流分割点的密度函数就可以积分$\sum_{n=4*1024}^{12*1024} n\times(P(n-1)-P(n))$，从而求得数据流分割点的期望长度，即平均分块大小，其中，4*1024（字节）表示最小数据块长度，12*1024（字节）表示最大数据块长度。

如图3所示的数据流分割点查找的基础上，在图5所示的实施方式中，在去重服务器103上预设有规则，所述规则为：为潜在分割点k确定11个点p_x、点p_x对应的窗口$W_x[p_x-A_x,p_x+B_x]$（简称窗口W_x）和窗口$W_x[p_x-A_x,p_x+B_x]$对应的预定条件C_x，其中，$A_1=A_2=A_3=A_4=A_5=A_6=A_7$

$=A_8=A_9=A_{10}=A_{11}$=169，$B_1=B_2=B_3=B_4=B_5=B_6=B_7=B_8=B_9=B_{10}=B_{11}$=0，并且$C_1=C_2=C_3=C_4=C_5=C_6=C_7=C_8=C_9=C_{10}=C_{11}$。其中，点$p_x$与潜在分割点$k$之间距离$d_x$个字节，具体的，点$p_1$与潜在分割点$k$之间距离0个字节，点$p_2$与潜在分割点$k$之间距离1个字节，点$p_3$与潜在分割点$k$之间距离2个字节，点$p_4$与潜在分割点$k$之间距离3个字节，点$p_5$与潜在分割点$k$之间距离4个字节，点$p_6$与潜在分割点$k$之间距离5个字节，点$p_7$与潜在分割点$k$之间距离6个字节，点$p_8$与潜在分割点$k$之间距离7个字节，点$p_9$与潜在分割点$k$之间距离8个字节，点$p_{10}$与潜在分割点$k$之间距离9个字节，点$p_{11}$与潜在分割点$k$之间距离10个字节，并且点$p_2$、$p_3$、$p_4$、$p_5$、$p_6$、$p_7$、$p_8$、$p_9$、$p_{10}$和$p_{11}$相对于潜在分割点$k$均位于数据流分割点查找反方向。$k_a$为数据流分割点，图5中所示数据流分割点查找方向为从左向右，从数据流分割点k_a跳过最小数据块4KB后，最小数据块4KB结束位置作为下一个潜在分割点k_i，为潜在分割点k_i确定点p_{ix}，在本实施例中，根据在去重服务器103上预设的规则，x分别为1到11连续的自然数。在图5所示的实施方式中，为潜在分割点k_i确定的点为11个，分别为p_{i1}、p_{i2}、p_{i3}、p_{i4}、p_{i5}、p_{i6}、p_{i7}、p_{i8}、p_{i9}、p_{i10}和p_{i11}，点p_{i1}、p_{i2}、p_{i3}、p_{i4}、p_{i5}、p_{i6}、p_{i7}、p_{i8}、p_{i9}、p_{i10}和p_{i11}对应的窗口分别为$W_{i1}[p_{i1}\text{-}169,p_{i1}]$、$W_{i2}[p_{i2}\text{-}169,p_{i2}]$、$W_{i3}[p_{i3}\text{-}169,p_{i3}]$、$W_{i4}[p_{i4}\text{-}169,p_{i4}]$、$W_{i5}[p_{i5}\text{-}169,p_{i5}]$、$W_{i6}[p_{i6}\text{-}169,p_{i6}]$、$W_{i7}[p_{i7}\text{-}169,p_{i7}]$、$W_{i8}[p_{i8}\text{-}169,p_{i8}]$、$W_{i9}[p_{i9}\text{-}169,p_{i9}]$、$W_{i10}[p_{i10}\text{-}169,p_{i10}]$和$W_{i11}[p_{i11}\text{-}169,p_{i11}]$。上述窗口分别简称为$W_{i1}$、$W_{i2}$、$W_{i3}$、$W_{i4}$、$W_{i5}$、$W_{i6}$、$W_{i7}$、$W_{i8}$、$W_{i9}$、$W_{i10}$和$W_{i11}$。其中，点$p_{ix}$与潜在分割点$k_i$之间距离$d_x$个字节，具体的，$p_{i1}$与$k_i$间距0个字节、$p_{i2}$与$k_i$间距1个字节、$p_{i3}$与$k_i$间距2个字节、$p_{i4}$与$k_i$间距3个字节、$p_{i5}$与$k_i$间距4个字节、$p_{i6}$与$k_i$间距5个字节、$p_{i7}$与$k_i$

间距6个字节、p_{i8}与k_i间距7个字节、p_{i9}与k_i间距8个字节、p_{i10}与k_i间距9个字节，p_{i11}与k_i间距10个字节，并且p_{i2}、p_{i3}、p_{i4}、p_{i5}、p_{i6}、p_{i7}、p_{i8}、p_{i9}、p_{i10}和p_{i11}相对于潜在分割点k_i均位于数据流分割点查找反方向。判断$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据是否满足预定条件$C_1$、判断$W_{i2}[p_{i2}\text{-}169, p_{i2}]$中至少部分数据是否满足预定条件$C_2$、判断$W_{i3}[p_{i3}\text{-}169, p_{i3}]$中至少部分数据是否满足预定条件$C_3$、判断$W_{i4}[p_{i4}\text{-}169, p_{i4}]$中至少部分数据是否满足预定条件$C_4$、判断$W_{i5}[p_{i5}\text{-}169, p_{i5}]$中至少部分数据是否满足预定条件$C_5$、判断$W_{i6}[p_{i6}\text{-}169, p_{i6}]$中至少部分数据是否满足预定条件$C_6$、判断$W_{i7}[p_{i7}\text{-}169, p_{i7}]$中至少部分数据是否满足预定条件$C_7$、判断$W_{i8}[p_{i8}\text{-}169, p_{i8}]$中至少部分数据是否满足预定条件$C_8$、判断$W_{i9}[p_{i9}\text{-}169, p_{i9}]$中至少部分数据是否满足预定条件$C_9$、判断$W_{i10}[p_{i10}\text{-}169, p_{i10}]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{i11}[p_{i11}\text{-}169, p_{i11}]$中至少部分数据是否满足预定条件$C_{11}$。当判断窗口$W_{i1}$中至少部分数据满足预定条件$C_1$、窗口$W_{i2}$中至少部分数据满足预定条件$C_2$、窗口$W_{i3}$中至少部分数据满足预定条件$C_3$、窗口$W_{i4}$中至少部分数据满足预定条件$C_4$、窗口$W_{i5}$中至少部分数据满足预定条件$C_5$、窗口$W_{i6}$中至少部分数据满足预定条件$C_6$、窗口$W_{i7}$中至少部分数据满足预定条件$C_7$、窗口$W_{i8}$中至少部分数据满足预定条件$C_8$、窗口$W_{i9}$中至少部分数据满足预定条件$C_9$、窗口$W_{i10}$中至少部分数据满足预定条件$C_{10}$和窗口$W_{i11}$中至少部分数据满足预定条件$C_{11}$时，则当前潜在分割点$k_i$为数据流分割点。当11个窗口中任一个窗口中至少部分数据不满足对应的预定条件时，如图6所示，$W_{i5}[p_{i5}\text{-}169, p_{i5}]$中至少部分数据不满足对应的预定条件$C_5$，则从点$p_{i5}$沿着数据流分割点查找方向跳跃N个字节，其中N个字节不大于$\| B_5 \| + \max_x(\| A_x \| + \| (k_i - p_{ix}) \|)$，在图6所示的实施

方式中，跳跃N个字节不大于179字节，在本实施例中，N=11，得到下一个潜在分割点，为与潜在分割点k_i区别，这里将新的潜在分割点表示为k_j。根据图5所示的实施方式中在去重服务器103上预设的规则，为潜在分割点k_j确定的点为11个，分别为p_{j1}、p_{j2}、p_{j3}、p_{j4}、p_{j5}、p_{j6}、p_{j7}、p_{j8}、p_{j9}、p_{j10}和p_{j11}，确定点p_{j1}、p_{j2}、p_{j3}、p_{j4}、p_{j5}、p_{j6}、p_{j7}、p_{j8}、p_{j9}、p_{j10}和p_{j11}对应的窗口分别为$W_{j1}[p_{j1}\text{-}169, p_{j1}]$、$W_{j2}[p_{j2}\text{-}169, p_{j2}]$、$W_{j3}[p_{j3}\text{-}169, p_{j3}]$、$W_{j4}[p_{j4}\text{-}169, p_{j4}]$、$W_{j5}[p_{j5}\text{-}169, p_{j5}]$、$W_{j6}[p_{j6}\text{-}169, p_{j6}]$、$W_{j7}[p_{j7}\text{-}169, p_{j7}]$、$W_{j8}[p_{j8}\text{-}169, p_{j8}]$、$W_{j9}[p_{j9}\text{-}169, p_{j9}]$、$W_{j10}[p_{j10}\text{-}169, p_{j10}]$和$W_{j11}[p_{j11}\text{-}169, p_{j11}]$。其中，$p_{jx}$与潜在分割点$k_j$之间距离$d_x$个字节，具体的，$p_{j1}$与$k_j$间距0个字节、$p_{j2}$与$k_j$间距1个字节、$p_{j3}$与$k_j$间距2个字节、$p_{j4}$与$k_j$间距3个字节、$p_{j5}$与$k_j$间距4个字节、$p_{j6}$与$k_j$间距5个字节、$p_{j7}$与$k_j$间距6个字节、$p_{j8}$与$k_j$间距7个字节、$p_{j9}$与$k_j$间距8个字节、$p_{j10}$与$k_j$间距9个字节，$p_{j11}$与$k_j$间距10个字节，并且$p_{j1}$、$p_{j2}$、$p_{j3}$、$p_{j4}$、$p_{j5}$、$p_{j6}$、$p_{j7}$、$p_{j8}$、$p_{j9}$、$p_{j10}$和$p_{j11}$相对于潜在分割点$k_j$均位于数据流分割点查找反方向。如图6所示实施方式中，当为潜在分割点k_j确定的第11个窗口$W_{j11}[p_{j11}\text{-}169, p_{j11}]$，在保证潜在分割点$k_i$与潜在分割点$k_j$之间的范围都在判断范围之内，则在本实施方式中，必须保证窗口$W_{j11}[p_{j11}\text{-}169, p_{j11}]$的左边界与$W_{i5}[p_{i5}\text{-}169, p_{i5}]$的右边界$p_{i5}$重合或者位于$W_{i5}[p_{i5}\text{-}169, p_{i5}]$范围之内，其中，所述潜在分割点$k_j$确定的点$p_{j11}$是根据所述规则，为所述潜在分割点$k_j$确定的M个点按照数据流查找方向获得的序列中排序第一的点。因此，在这一限定内，当$W_{i5}[p_{i5}\text{-}169, p_{i5}]$中至少部分数据不满足预定条件$C_5$，从$p_{i5}$沿着数据流分割点查找方向跳跃的距离为不大于$\| B_5 \| + \max_x(\| A_x \| + \| (k_i - p_{ix}) \|)$，其中，M=11，11*U不大于$\max_x(\| A_x \| + \| (k_i - p_{ix}) \|)$，因此，

从p_{i5}沿着数据流分割点查找方向跳跃的距离为不大于179。判断W_{j1}[p_{j1}-169,p_{j1}]中至少部分数据是否满足预定条件C_1、判断W_{j2}[p_{j2}-169,p_{j2}]中至少部分数据是否满足预定条件C_2、判断W_{j3}[p_{j3}-169,p_{j3}]中至少部分数据是否满足预定条件C_3、判断W_{j4}[p_{j4}-169,p_{j4}]中至少部分数据是否满足预定条件C_4、判断W_{j5}[p_{j5}-169,p_{j5}]中至少部分数据是否满足预定条件C_5、判断W_{j6}[p_{j6}-169, p_{j6}]中至少部分数据是否满足预定条件C_6、判断W_{j7}[p_{j7}-169, p_{j7}]中至少部分数据是否满足预定条件C_7、判断W_{j8}[p_{j8}-169, p_{j8}]中至少部分数据是否满足预定条件C_8、判断W_{j9}[p_{j9}-169, p_{j9}]中至少部分数据是否满足预定条件C_9、判断W_{j10}[p_{j10}-169, p_{j10}]中至少部分数据是否满足预定条件C_{10}和判断W_{j11}[p_{j11}-169,p_{j11}]中至少部分数据是否满足预定条件C_{11}。当然在本发明实施例中，判断潜在分割点k_a是否为数据流分割点时也遵循该规则，具体实现不再描述，可以参照判断潜在分割点k_i的描述。当判断窗口W_{j1}中至少部分数据满足预定条件C_1、窗口W_{j2}中至少部分数据满足预定条件C_2、窗口W_{j3}中至少部分数据满足预定条件C_3、窗口W_{j4}中至少部分数据满足预定条件C_4、窗口W_{j5}中至少部分数据满足预定条件C_5、窗口W_{j6}中至少部分数据满足预定条件C_6、窗口W_{j7}中至少部分数据满足预定条件C_7、窗口W_{j8}中至少部分数据满足预定条件C_8、窗口W_{j9}中至少部分数据满足预定条件C_9、窗口W_{j10}中至少部分数据满足预定条件C_{10}和窗口W_{j11}中至少部分数据满足预定条件C_{11}时，则当前潜在分割点k_j为数据流分割点，k_j与k_a之间的数据构成1个数据块，同时按照与k_a相同的方式跳过最小分块大小4KB，获得下一个潜在分割点，并按照在去重服务器103上预设的规则，判断下一个潜在分割点是否为数据流分

割点。当判断潜在分割点k_j不是数据流分割点时，按照与k_i相同的方式跳跃11个字节获得下一个潜在分割点，并按照在去重服务器103上预设的规则及上述方法判断下一个潜在分割点是否为数据流分割点。当超过设定的最大数据块仍然没有找到数据流分割点时，则从最大数据块的结束位置作为强制分割点。

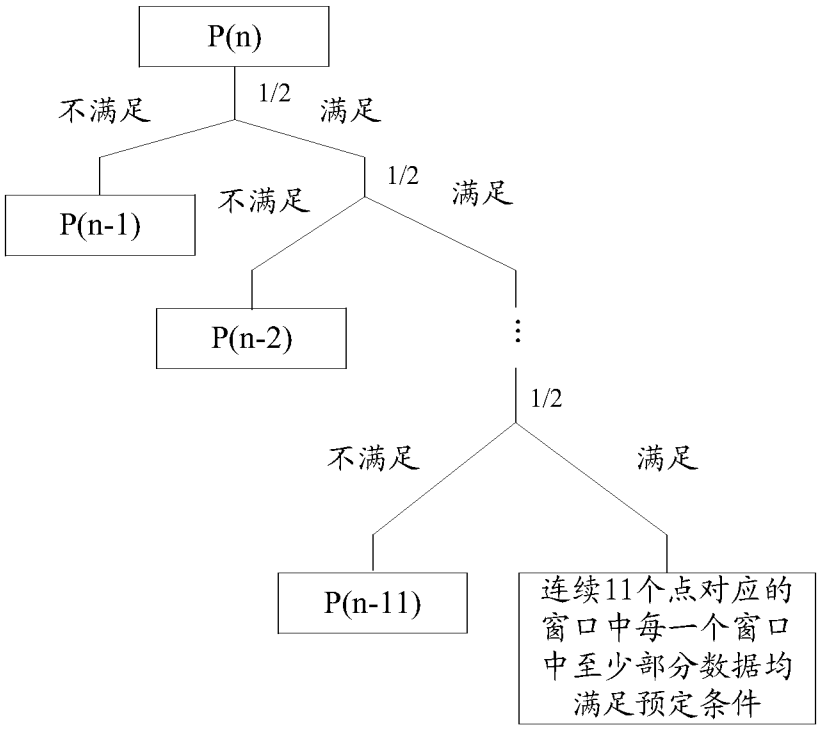
在图5所示的实施方式中，根据在去重服务器103上预设的规则，从判断$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据是否满足预定条件$C_1$开始，当判断$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据满足预定条件$C_1$、判断$W_{i2}[p_{i2}\text{-}169, p_{i2}]$中至少部分数据满足预定条件$C_2$、判断$W_{i3}[p_{i3}\text{-}169, p_{i3}]$中至少部分数据满足预定条件$C_3$和判断$W_{i4}[p_{i4}\text{-}169, p_{i4}]$中至少部分数据满足预定条件$C_4$，判断$W_{i5}[p_{i5}\text{-}169, p_{i5}]$中至少部分数据不满足预定条件$C_5$时，从点$p_{i5}$沿着数据流分割点查找方向跳跃10个字节，在第10个字节的结束位置获得新的潜在分割点，为与其他潜在分割点区别，这里表示为k_g，按照在去重服务器103上预设的规则，为潜在分割点k_g确定11个点p_{gx}，x分别为1到11连续的自然数，分别为p_{g1}、p_{g2}、p_{g3}、p_{g4}、p_{g5}、p_{g6}、p_{g7}、p_{g8}、p_{g9}、p_{g10}和p_{g11}，确定点p_{g1}、p_{g2}、p_{g3}、p_{g4}、p_{g5}、p_{g6}、p_{g7}、p_{g8}、p_{g9}、p_{g10}和p_{g11}对应的窗口分别为$W_{g1}[p_{g1}\text{-}169, p_{g1}]$、$W_{g2}[p_{g2}\text{-}169, p_{g2}]$、$W_{g3}[p_{g3}\text{-}169, p_{g3}]$、$W_{g4}[p_{g4}\text{-}169, p_{g4}]$、$W_{g5}[p_{g5}\text{-}169, p_{g5}]$、$W_{g6}[p_{g6}\text{-}169, p_{g6}]$、$W_{g7}[p_{g7}\text{-}169, p_{g7}]$、$W_{g8}[p_{g8}\text{-}169, p_{g8}]$、$W_{g9}[p_{g9}\text{-}169, p_{g9}]$、$W_{g10}[p_{g10}\text{-}169, p_{g10}]$和$W_{g11}[p_{g11}\text{-}169, p_{g11}]$。其中，$p_{gx}$与潜在分割点$k_g$之间距离$d_x$个字节，具体的，$p_{g1}$与$k_g$间距0个字节、$p_{g2}$与$k_g$间距1个字节、$p_{g3}$与$k_g$间距2个字节、$p_{g4}$与$k_g$间距3个字节、$p_{g5}$与$k_g$间距4个字节、$p_{g6}$与$k_g$间距5个字节、$p_{g7}$与$k_g$间距6个字节、$p_{g8}$与$k_g$间距7个字节、$p_{g9}$与$k_g$间距8个字节、$p_{g10}$与$k_g$间距9个字节，$p_{g11}$与$k_g$间

距10个字节，并且p_{g2}、p_{g3}、p_{g4}、p_{g5}、p_{g6}、p_{g7}、p_{g8}、p_{g9}、p_{g10}和p_{g11}相对于潜在分割点k_g均位于数据流分割点查找反方向。判断$W_{g1}[p_{g1}$-169，$p_{g1}]$中至少部分数据是否满足预定条件C_1、判断$W_{g2}[p_{g2}$-169，$p_{g2}]$中至少部分数据是否满足预定条件C_2、判断$W_{g3}[p_{g3}$-169，$p_{g3}]$中至少部分数据是否满足预定条件C_3、判断$W_{g4}[p_{g4}$-169，$p_{g4}]$中至少部分数据是否满足预定条件C_4、判断$W_{g5}[p_{g5}$-169，$p_{g5}]$中至少部分数据是否满足预定条件C_5、判断$W_{g6}[p_{g6}$-169，$p_{g6}]$中至少部分数据是否满足预定条件C_6、判断$W_{g7}[p_{g7}$-169，$p_{g7}]$中至少部分数据是否满足预定条件C_7、判断$W_{g8}[p_{g8}$-169，$p_{g8}]$中至少部分数据是否满足预定条件C_8、判断$W_{g9}[p_{g9}$-169，$p_{g9}]$中至少部分数据是否满足预定条件C_9、判断$W_{g10}[p_{g10}$-169，$p_{g10}]$中至少部分数据是否满足预定条件C_{10}和判断$W_{g11}[p_{g11}$-169，$p_{g11}]$中至少部分数据是否满足预定条件C_{11}。因此，潜在分割点k_g对应的点p_{g11}与潜在分割点k_i对应的点p_{i5}重合，并且点p_{g11}对应的窗口$W_{g11}[p_{g11}$-169，$p_{g11}]$与点p_{i5}对应的窗口$W_{i5}[p_{i5}$-169,$p_{i5}]$重合，并且C_5=C_{11}，因此，对当潜在分割点k_i，当判断$W_{i5}[p_{i5}$-169,$p_{i5}]$中至少部分数据不满足预定条件C_5时，从点p_{i5}沿着数据流分割点查找方向跳跃10个字节，获得的潜在分割点k_g仍然不符合作为数据流分割点的条件。因此，如果从点p_{i5}沿着数据流分割点查找方向跳跃10个字节会存在重复计算，从点p_{i5}沿着数据流分割点查找方向跳跃11个字节可以减少重复计算，效率更高。因此提高了查找数据流分割点的速度。当预设规定中点p_x对应的窗口$W_x[p_x$-A_x,p_x+$B_x]$中至少部分数据满足预定条件C_x的概率为1/2时，即是说以1/2的概率执行跳跃，每次最多可以跳跃179个字节。

在本实施方式中，预定规则为:为潜在分割点k确定11个点p_x、点

p_x对应的窗口$W_x[p_x-A_x,p_x+B_x]$和窗口$W_x[p_x-A_x,p_x+B_x]$对应的预定条件C_x，x分别为1到11连续的自然数，其中，点p_x对应的窗口$W_x[p_x-A_x,p_x+B_x]$中至少部分数据满足预定条件的概率为1/2，通过这两个因素可以计算P(n)。并且$A_1=A_2=A_3=A_4=A_5=A_6=A_7=A_8=A_9=A_{10}=A_{11}$=169，$B_1=B_2=B_3=B_4=B_5=B_6=B_7=B_8=B_9=B_{10}=B_{11}$=0，并且$C_1=C_2=C_3=C_4=C_5=C_6=C_7=C_8=C_9=C_{10}=C_{11}$，其中，$p_x$与潜在分割点$k$之间距离$d_x$个字节，具体的，$p_1$与潜在分割点$k$之间距离0个字节，$p_2$与$k$之间距离1个字节，$p_3$与$k$之间距离2个字节，$p_4$与$k$之间距离3个字节，$p_5$与$k$之间距离4个字节，$p_6$与$k$之间距离5个字节，$p_7$与$k$之间距离6个字节，$p_8$与$k$之间距离7个字节，$p_9$与$k$之间距离8个字节，$p_{10}$与$k$之间距离9个字节，$p_{11}$与$k$之间距离10个字节，并且$p_2$、$p_3$、$p_4$、$p_5$、$p_6$、$p_7$、$p_8$、$p_9$、$p_{10}$和$p_{11}$相对于潜在分割点$k$均位于数据流分割点查找反方向。因此是否存在连续11个点对应窗口中的每一个窗口中至少部分数据均满足预定条件C_x就决定潜在分割点k是否为数据流分割点。从数据流起始位置/上一数据流分割点跳跃最小分块长度4096个字节后，向数据流分割点查找反方向回退10个字节，找到第4086个点，在该点处不存在数据流分割点，所以P（4086）=1，依次类推，P（4087）=1，……P（4095）=1。在第4096个点处，即在最小分块大小处，以（1/2）^11的概率这11个点对应的窗口中每一个窗口中至少部分数据满足预定条件C_x，因此以（1/2）^11的概率存在数据流分割点，以1-（1/2）^11的概率不存在数据流分割点，所以P（11）=1-（1/2）^11。

在第n个点处，可以分为12种情况来递推P（n）。



情况1：第n个点对应的窗口中至少部分数据以1/2的概率不满足预定条件，此时第n个点前面的n-1个点以P（n-1）的概率不存在连续的11个点对应的窗口中每一个窗口中至少部分数据分别满足预定条件，因此P（n）包含1/2* P（n-1）。第n个点对应的窗口中至少部分数据不满足预定条件，并且第n个点前面的n-1个点存在连续的11个点对应的窗口中每一个窗口中至少部分数据分别满足预定条件的情况与P（n）无关。

情况2：第n个点对应的窗口中至少部分数据以1/2的概率满足预定条件，第n-1个点对应的窗口中至少部分数据以1/2的概率不满足预定条件，此时第n-1个点前面的n-2个点以P（n-2）的概率不存在连续的11个点对应的窗口中每一个窗口中至少部分数据分别满足预定条件，因此P（n）包含1/2*1/2*P（n-2）。第n个点对应的窗口中至少部分数据满足预定条件，第n-1个点对应的窗口中至少部分数据不满足预定条件，并且第n-1个点前面的n-2个点存在连续的11个点对应的窗口中每一个窗口中至少部分数据分别满足预定条件的情况与P（n）无

关。

依照上述描述，情况11：第n至n-9个点对应的窗口中至少部分数据以（1/2）^10的概率满足预定条件，第n-10个点对应的窗口中至少部分数据以1/2的概率不满足预定条件，此时第n-10个点前面的n-11个点以P（n-11）的概率不存在连续的11个点对应的窗口中每一个窗口中至少部分数据分别满足预定条件，因此P(n)包含(1/2)^10*1/2*P（n-11）。第n至n-9个点对应的窗口中至少部分数据均满足预定条件，第n-10个点对应的窗口中至少部分数据不满足预定条件，并且第n-10个点前面的n-11个点存在连续的11个点对应的窗口中每一个窗口中至少部分数据分别满足预定条件的情况与P（n）无关。

情况12：第n至n-10个点对应的窗口中至少部分数据以（1/2）^11的概率满足预定条件，该情况与P（n）无关。

因此，P（n）=1/2* P（n-1）+（1/2）^2* P（n-2）+……+（1/2）^11*P（n-11）。另一种预设规则：为潜在分割点k确定24个点p_x、点p_x对应的窗口$W_x[p_x-A_x,p_x+B_x]$和窗口$W_x[p_x-A_x,p_x+B_x]$对应的预定条件C_x，x分别为1到24连续的自然数，其中，点p_x对应的窗口$W_x[p_x-A_x,p_x+B_x]$中至少部分数据满足预定条件C_x的概率为3/4，通过这两个因素可以计算P(n)。并且$A_1=A_2=A_3=A_4=A_5=A_6=A_7=A_8=A_9=A_{10}=A_{11}$=169，$B_1=B_2=B_3=B_4=B_5=B_6=B_7=B_8=B_9=B_{10}=B_{11}$=0，并且$C_1=C_2=C_3=C_4=C_5=C_6=C_7=C_8=C_9=\ldots=C_{22}=C_{23}=C_{24}$，其中，$p_x$与潜在分割点$k$之间距离$d_x$个字节，具体的，$p_1$与潜在分割点$k$之间距离0个字节，$p_2$与$k$之间距离1个字节，$p_3$与$k$之间距离2个字节，$p_4$与$k$之间距离3个字节，$p_5$与$k$之间距离4个字节，$p_6$与$k$之间距离5个字节，$p_7$与$k$之间距离6个字节，$p_8$与$k$之间距离7个字节，$p_9$与$k$之间距离8个字节,...$p_{22}$与$k$之间距离21个字节，$p_{23}$与$k$之间距离22个字节，$p_{24}$与$k$之间距离23个字节,并且

p_2、p_3、p_4、p_5、p_6、p_7、p_8、$p_9 \cdots p_{22}$、p_{23}和p_{24}相对于潜在分割点k均位于数据流分割点查找反方向。因此是否存在连续24个点对应窗口中的每一个窗口中至少部分数据均满足预定条件C_x就决定潜在分割点k是否为数据流分割点，可以通过下面的公式计算:

P（4073）=1，P（4074）=1，……P（,4095）=1，P（4096）=1-（3/4）^24，

P（n）=1/4* P（n-1）+1/4*（3/4）* P（n-2）+……+1/4*（3/4）^23*P（n-24）。

经过计算，P(5*1024)=0.78，P(11*1024)=0.17，P(12*1024)=0.13，即从数据流起始位置/上一数据流分割点查找到12KB后以13%的概率仍未找到数据流分割点，强制进行分割。通过这个概率，求得数据流分割点的密度函数，经过积分求得大约平均在从数据流起始位置/上一数据流分割点查找7.6KB时找到数据流分割点，即平均分块长度大约为7.6KB。与连续的11个点对应的窗口中至少部分数据以1/2的概率满足预定条件不同，传统CDC算法采用一个窗口以1/2^12的概率满足条件时，方可达到平均分块长度7.6KB的效果。

在图3所示的数据流分割点查找的基础上，在图7所示的实施方式中，在去重服务器103上预设有规则，所述规则为：为潜在分割点k确定11个点p_x、点p_x对应的窗口$W_x[p_x-A_x, p_x+B_x]$和窗口$W_x[p_x-A_x, p_x+B_x]$对应的预定条件C_x，x分别为1到11连续的自然数，其中，点p_x对应的窗口$W_x[p_x-A_x, p_x+B_x]$中至少部分数据满足预定条件C_x的概率为1/2，并且$A_1=A_2=A_3=A_4=A_5=A_6=A_7=A_8=A_9=A_{10}=A_{11}=169$，$B_1=B_2=B_3=B_4=B_5=B_6=B_7=B_8=B_9=B_{10}=B_{11}=0$，并且$C_1=C_2=C_3=C_4=C_5=C_6=C_7=C_8=C_9=C_{10}=C_{11}$，其中，$p_x$与潜在分割点$k$之间距离$d_x$个字节，具体的，$p_1$与潜在分割点$k$之间距离2个字节，$p_2$与$k$之间距离3个字节，$p_3$与$k$之间距离

4个字节，p_4与k之间距离5个字节，p_5与k之间距离6个字节，p_6与k之间距离7个字节，p_7与k之间距离8个字节，p_8与k之间距离9个字节，p_9与k之间距离10个字节，p_{10}与k之间距离1个字节，p_{11}与k之间距离0个字节，并且p_1、p_2、p_3、p_4、p_5、p_6、p_7、p_8、p_9和p_{10}相对于潜在分割点k均位于数据流分割点查找反方向。k_a为数据流分割点，图7中所示数据流分割点查找方向为从左向右，从数据流分割点k_a跳过最小数据块4KB后，在最小数据块4KB结束位置作为下一个潜在分割点k_i，为潜在分割点k_i确定点p_{ix}，在本实施例中，根据在去重服务器103上预设的规则，x分别为1到11连续的自然数。在图7所示的实施方式中，依据预定规则，为潜在分割点k_i确定的点为11个，分别为p_{i1}、p_{i2}、p_{i3}、p_{i4}、p_{i5}、p_{i6}、p_{i7}、p_{i8}、p_{i9}、p_{i10}和p_{i11}，点p_{i1}、p_{i2}、p_{i3}、p_{i4}、p_{i5}、p_{i6}、p_{i7}、p_{i8}、p_{i9}、p_{i10}和p_{i11}对应的窗口分别为$W_{i1}[p_{i1}\text{-}169, p_{i1}]$、$W_{i2}[p_{i2}\text{-}169, p_{i2}]$、$W_{i3}[p_{i3}\text{-}169, p_{i3}]$、$W_{i4}[p_{i4}\text{-}169, p_{i4}]$、$W_{i5}[p_{i5}\text{-}169, p_{i5}]$、$W_{i6}[p_{i6}\text{-}169, p_{i6}]$、$W_{i7}[p_{i7}\text{-}169, p_{i7}]$、$W_{i8}[p_{i8}\text{-}169, p_{i8}]$、$W_{i9}[p_{i9}\text{-}169, p_{i9}]$、$W_{i10}[p_{i10}\text{-}169, p_{i10}]$和$W_{i11}[p_{i11}\text{-}169, p_{i11}]$。其中，点$p_{ix}$与潜在分割点$k_i$之间距离$d_{ix}$个字节，具体的，$p_{i1}$与$k_i$间距2个字节、$p_{i2}$与$k_i$间距3个字节、$p_{i3}$与$k_i$间距4个字节、$p_{i4}$与$k_i$间距5个字节、$p_{i5}$与$k_i$间距6个字节、$p_{i6}$与$k_i$间距7个字节、$p_{i7}$与$k_i$间距8个字节、$p_{i8}$与$k_i$间距9个字节、$p_{i9}$与$k_i$间距10个字节、$p_{i10}$与$k_i$间距1个字节，$p_{i11}$与$k_i$间距0个字节，并且$p_{i1}$、$p_{i2}$、$p_{i3}$、$p_{i4}$、$p_{i5}$、$p_{i6}$、$p_{i7}$、$p_{i8}$、$p_{i9}$和$p_{i10}$相对于潜在分割点$k_i$均位于数据流分割点查找反方向。判断$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据是否满足预定条件$C_1$、判断$W_{i2}[p_{i2}\text{-}169, p_{i2}]$中至少部分数据是否满足预定条件$C_2$、判断$W_{i3}[p_{i3}\text{-}169, p_{i3}]$中至少部分数据是否满足预定条件$C_3$、判断$W_{i4}[p_{i4}\text{-}169, p_{i4}]$中至少部分数据是否满足预定条件$C_4$、判断$W_{i5}$

$[p_{i5}\text{-}169, p_{i5}]$中至少部分数据是否满足预定条件$C_5$、判断$W_{i6}[p_{i6}\text{-}169, p_{i6}]$中至少部分数据是否满足预定条件$C_6$、判断$W_{i7}[p_{i7}\text{-}169, p_{i7}]$中至少部分数据是否满足预定条件$C_7$、判断$W_{i8}[p_{i8}\text{-}169, p_{i8}]$中至少部分数据是否满足预定条件$C_8$、判断$W_{i9}[p_{i9}\text{-}169, p_{i9}]$中至少部分数据是否满足预定条件$C_9$、判断$W_{i10}[p_{i10}\text{-}169, p_{i10}]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{i11}[p_{i11}\text{-}169, p_{i11}]$中至少部分数据是否满足预定条件$C_{11}$。当判断窗口$W_{i1}$中至少部分数据满足预定条件$C_1$、窗口$W_{i2}$中至少部分数据满足预定条件$C_2$、窗口$W_{i3}$中至少部分数据满足预定条件$C_3$、窗口$W_{i4}$中至少部分数据满足预定条件$C_4$、窗口$W_{i5}$中至少部分数据满足预定条件$C_5$、窗口$W_{i6}$中至少部分数据满足预定条件$C_6$、窗口$W_{i7}$中至少部分数据满足预定条件$C_7$、窗口$W_{i8}$中至少部分数据满足预定条件$C_8$、窗口$W_{i9}$中至少部分数据满足预定条件$C_9$、窗口$W_{i10}$中至少部分数据满足预定条件$C_{10}$和窗口$W_{i11}$中至少部分数据满足预定条件$C_{11}$时，则当前潜在分割点$k_i$为数据流分割点。当11个窗口中任一个窗口中至少部分数据不满足对应的预定条件时，如图8所示，$W_{i3}[p_{i3}\text{-}169, p_{i3}]$中至少部分数据不满足预定条件$C_3$，点$p_{i3}$沿着数据流分割点查找方向跳跃11个字节为例进行描述。如图8所示，当判断W_3不满足预定条件时，以p_3为起始点，沿着数据流分割点查找方向跳跃N个字节，其中N个字节不大于$\| B_3 \| + \max_x(\| A_x \| + \| (k_i - p_{ix}) \|)$，在图6所示的实施方式中，跳跃N个字节，具体为不大于179字节，在本实施例中，N=11，在第11个字节的结束位置，获得下一个潜在分割点，为与潜在分割点k_i区别，这里将新的潜在分割点表示为k_j，根据在去重服务器103上预设的规则，为潜在分割点k_j确定的点为11个，分别为p_{j1}、p_{j2}、p_{j3}、p_{j4}、p_{j5}、p_{j6}、p_{j7}、p_{j8}、p_{j9}、p_{j10}和p_{j11}，确定点p_{j1}、p_{j2}、p_{j3}、p_{j4}、p_{j5}、

p_{j6}、p_{j7}、p_{j8}、p_{j9}、p_{j10}和p_{j11}对应的窗口分别为$W_{j1}[p_{j1}\text{-}169, p_{j1}]$、$W_{j2}[p_{j2}\text{-}169, p_{j2}]$、$W_{j3}[p_{j3}\text{-}169, p_{j3}]$、$W_{j4}[p_{j4}\text{-}169, p_{j4}]$、$W_{j5}[p_{j5}\text{-}169, p_{j5}]$、$W_{j6}[p_{j6}\text{-}169, p_{j6}]$、$W_{j7}[p_{j7}\text{-}169, p_{j7}]$、$W_{j8}[p_{j8}\text{-}169, p_{j8}]$、$W_{j9}[p_{j9}\text{-}169, p_{j9}]$、$W_{j10}[p_{j10}\text{-}169, p_{j10}]$和$W_{j11}[p_{j11}\text{-}169, p_{j11}]$。其中，$p_{jx}$与潜在分割点$k_j$之间距离$d_x$个字节，具体的，$p_{j1}$与$k_j$间距2个字节、$p_{j2}$与$k_j$间距3个字节、$p_{j3}$与$k_j$间距4个字节、$p_{j4}$与$k_j$间距5个字节、$p_{j5}$与$k_j$间距6个字节、$p_{j6}$与$k_j$间距7个字节、$p_{j7}$与$k_j$间距8个字节、$p_{j8}$与$k_j$间距9个字节、$p_{j9}$与$k_j$间距10个字节、$p_{j10}$与$k_j$间距1个字节，$p_{j11}$与$k_j$间距0个字节，并且$p_{j1}$、$p_{j2}$、$p_{j3}$、$p_{j4}$、$p_{j5}$、$p_{j6}$、$p_{j7}$、$p_{j8}$、$p_{j9}$和$p_{j10}$相对于潜在分割点$k_j$均位于数据流分割点查找反方向。判断$W_{j1}[p_{j1}\text{-}169, p_{j1}]$中至少部分数据是否满足预定条件$C_1$、判断$W_{j2}[p_{j2}\text{-}169, p_{j2}]$中至少部分数据是否满足预定条件$C_2$、判断$W_{j3}[p_{j3}\text{-}169, p_{j3}]$中至少部分数据是否满足预定条件$C_3$、判断$W_{j4}[p_{j4}\text{-}169, p_{j4}]$中至少部分数据是否满足预定条件$C_4$、判断$W_{j5}[p_{j5}\text{-}169, p_{j5}]$中至少部分数据是否满足预定条件$C_5$、判断$W_{j6}[p_{j6}\text{-}169, p_{j6}]$中至少部分数据是否满足预定条件$C_6$、判断$W_{j7}[p_{j7}\text{-}169, p_{j7}]$中至少部分数据是否满足预定条件$C_7$、判断$W_{j8}[p_{j8}\text{-}169, p_{j8}]$中至少部分数据是否满足预定条件$C_8$、判断$W_{j9}[p_{j9}\text{-}169, p_{j9}]$中至少部分数据是否满足预定条件$C_9$、判断$W_{j10}[p_{j10}\text{-}169, p_{j10}]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{j11}[p_{j11}\text{-}169, p_{j11}]$中至少部分数据是否满足预定条件$C_{11}$。当然在本发明实施例中，判断潜在分割点$k_a$是否为数据流分割点时也遵循该原则，具体实现不再描述，可以参照判断潜在分割点k_i的描述。当判断窗口W_{j1}中至少部分数据满足预定条件C_1、窗口W_{j2}中至少部分数据满足预定条件C_2、窗口W_{j3}中至少部

分数据满足预定条件C_3、窗口W_{j4}中至少部分数据满足预定条件C_4、窗口W_{j5}中至少部分数据满足预定条件C_5、窗口W_{j6}中至少部分数据满足预定条件C_6、窗口W_{j7}中至少部分数据满足预定条件C_7、窗口W_{j8}中至少部分数据满足预定条件C_8、窗口W_{j9}中至少部分数据满足预定条件C_9、窗口W_{j10}中至少部分数据满足预定条件C_{10}和窗口W_{j11}中至少部分数据满足预定条件C_{11}时，则当前潜在分割点k_j为数据流分割点，k_j与k_a之间的数据构成1个数据块，同时按照与k_a相同的方式跳过最小分块大小4KB，获得下一个潜在分割点，并按照在去重服务器103上预设的规则，判断下一个潜在分割点是否为数据流分割点。当判断潜在分割点k_j不是数据流分割点时，按照与k_i相同的方式跳跃11个字节获得下一个潜在分割点，并按照在去重服务器103上预设的规则及上述方法判断下一个潜在分割点是否为数据流分割点。当超过设定的最大数据块仍然没有找到数据流分割点时，则从最大数据块的结束位置作为强制分割点。当然该方法的实施受最大数据块长度和构成该数据流的文件的大小约束，在此不再赘述。

在图3所示的数据流分割点查找的基础上，在图9所示的实施方式中，在去重服务器103上预设有规则，所述规则为：为潜在分割点k确定11个点p_x、点p_x对应的窗口$W_x[p_x-A_x,p_x+B_x]$和窗口$W_x[p_x-A_x,p_x+B_x]$对应的预定条件C_x，其中$A_1=A_2=A_3=A_4=A_5=A_6=A_7=A_8=A_9=A_{10}=A_{11}=169$，$B_1=B_2=B_3=B_4=B_5=B_6=B_7=B_8=B_9=B_{10}=B_{11}=0$，并且$C_1=C_2=C_3=C_4=C_5=C_6=C_7=C_8=C_9=C_{10}=C_{11}$。其中，$p_x$与潜在分割点$k$之间距离$d_x$个字节，具体的，$p_1$与潜在分割点$k$之间距离3个字节，$p_2$与$k$之间距离2个字节，$p_3$与$k$之间距离1个字节，$p_4$与$k$之间距离0个字节，$p_5$与$k$之间距离1个字节，$p_6$与$k$之间距离2个字节，$p_7$与$k$之间距离3个字节，$p_8$与$k$之

间距离4个字节，p_9与k之间距离5个字节，p_{10}与k之间距离6个字节，p_{11}与k之间距离7个字节，并且p_5、p_6、p_7、p_8、p_9、p_{10}和p_{11}相对于潜在分割点k均位于数据流分割点查找反方向，p_1、p_2和p_3相对于潜在分割点k均位于数据流分割点查找方向。k_a为数据流分割点，图9中所示数据流分割点查找方向为从左向右，从数据流分割点k_a跳过最小数据块4KB后，最小数据块4KB结束位置作为下一个潜在分割点k_i，为潜在分割点k_i确定点p_{ix}，在本实施例中，根据在去重服务器103上预设的规则，x分别为1到11连续的自然数。在图9所示的实施方式中，为潜在分割点k_i确定的点为11个，分别为p_{i1}、p_{i2}、p_{i3}、p_{i4}、p_{i5}、p_{i6}、p_{i7}、p_{i8}、p_{i9}、p_{i10}和p_{i11}，点p_{i1}、p_{i2}、p_{i3}、p_{i4}、p_{i5}、p_{i6}、p_{i7}、p_{i8}、p_{i9}、p_{i10}和p_{i11}对应的窗口分别为$W_{i1}[p_{i1}\text{-}169, p_{i1}]$、$W_{i2}[p_{i2}\text{-}169, p_{i2}]$、$W_{i3}[p_{i3}\text{-}169, p_{i3}]$、$W_{i4}[p_{i4}\text{-}169, p_{i4}]$、$W_{i5}[p_{i5}\text{-}169, p_{i5}]$、$W_{i6}[p_{i6}\text{-}169, p_{i6}]$、$W_{i7}[p_{i7}\text{-}169, p_{i7}]$、$W_{i8}[p_{i8}\text{-}169, p_{i8}]$、$W_{i9}[p_{i9}\text{-}169, p_{i9}]$、$W_{i10}[p_{i10}\text{-}169, p_{i10}]$和$W_{i11}[p_{i11}\text{-}169, p_{i11}]$。其中，$p_{ix}$与潜在分割点$k_i$之间距离$d_x$个字节，具体的，$p_{i1}$与$k_i$间距3个字节、$p_{i2}$与$k_i$间距2个字节、$p_{i3}$与$k_i$间距1个字节、$p_{i4}$与$k_i$间距0个字节、$p_{i5}$与$k_i$间距1个字节、$p_{i6}$与$k_i$间距2个字节、$p_{i7}$与$k_i$间距3个字节、$p_{i8}$与$k_i$间距4个字节、$p_{i9}$与$k_i$间距5个字节、$p_{i10}$与$k_i$间距6个字节，$p_{i11}$与$k_i$间距7个字节，并且$p_{i5}$、$p_{i6}$、$p_{i7}$、$p_{i8}$、$p_{i9}$、$p_{i10}$和$p_{i11}$相对于潜在分割点$k_i$均位于数据流分割点查找反方向，$p_{i1}$、$p_{i2}$和$p_{i3}$相对于潜在分割点$k_i$均位于数据流分割点查找方向。判断$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据是否满足预定条件$C_1$、判断$W_{i2}[p_{i2}\text{-}169, p_{i2}]$中至少部分数据是否满足预定条件$C_2$、判断$W_{i3}[p_{i3}\text{-}169, p_{i3}]$中至少部分数据是否满足预定条件$C_3$、判断$W_{i4}[p_{i4}\text{-}169, p_{i4}]$中至少部分数据是否满足预定条件$C_4$、判断$W_{i5}[p_{i5}\text{-}169, p_{i5}]$中至少部分数据是否满足预

定条件C_5、判断$W_{i6}[p_{i6}\text{-}169,p_{i6}]$中至少部分数据是否满足预定条件$C_6$、判断$W_{i7}[p_{i7}\text{-}169,p_{i7}]$中至少部分数据是否满足预定条件$C_7$、判断$W_{i8}[p_{i8}\text{-}169,p_{i8}]$中至少部分数据是否满足预定条件$C_8$、判断$W_{i9}[p_{i9}\text{-}169,p_{i9}]$中至少部分数据是否满足预定条件$C_9$、判断$W_{i10}[p_{i10}\text{-}169,p_{i10}]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{i11}[p_{i11}\text{-}169,p_{i11}]$中至少部分数据是否满足预定条件$C_{11}$。当判断窗口$W_{i1}$中至少部分数据满足预定条件$C_1$、窗口$W_{i2}$中至少部分数据满足预定条件$C_2$、窗口$W_{i3}$中至少部分数据满足预定条件$C_3$、窗口$W_{i4}$中至少部分数据满足预定条件$C_4$、窗口$W_{i5}$中至少部分数据满足预定条件$C_5$、窗口$W_{i6}$中至少部分数据满足预定条件$C_6$、窗口$W_{i7}$中至少部分数据满足预定条件$C_7$、窗口$W_{i8}$中至少部分数据满足预定条件$C_8$、窗口$W_{i9}$中至少部分数据满足预定条件$C_9$、窗口$W_{i10}$中至少部分数据满足预定条件$C_{10}$和窗口$W_{i11}$中至少部分数据满足预定条件$C_{11}$时，则当前潜在分割点$k_i$为数据流分割点。当11个窗口中任一个窗口中至少部分数据不满足对应的预定条件时，如图10所示，$W_{i7}[p_{i7}\text{-}169,p_{i7}]$中至少部分数据不满足对应的预定条件，则从点$p_{i7}$沿着数据流分割点查找方向跳跃N个字节，其中N个字节不大于$\|B_4\|+\max_x(\|A_x\|+\|(k_i-p_{ix})\|)$，在图10所示的实施方式中，跳跃N个字节，具体为不大于179个字节，在本实施例中，具体取N=8，得到新的潜在分割点，为与潜在分割点k_i区别，这里将新的潜在分割点表示为k_j，根据图9所示的实施方式中在去重服务器103上预设的规则，为潜在分割点k_j确定的点为11个，分别为p_{j1}、p_{j2}、p_{j3}、p_{j4}、p_{j5}、p_{j6}、p_{j7}、p_{j8}、p_{j9}、p_{j10}和p_{j11}，确定点p_{j1}、p_{j2}、p_{j3}、p_{j4}、p_{j5}、p_{j6}、p_{j7}、p_{j8}、p_{j9}、p_{j10}和p_{j11}对应的窗口分别为$W_{j1}[p_{j1}\text{-}169,p_{j1}]$、$W_{j2}[p_{j2}\text{-}169,p_{j2}]$、$W_{j3}[p_{j3}\text{-}169,p_{j3}]$、$W_{j4}[p_{j4}\text{-}169,p_{j4}]$、$W_{j5}[p_{j5}\text{-}169,p_{j5}]$、$W_{j6}[p_{j6}\text{-}169,\ p_{j6}]$、

$W_{j7}[p_{j7}\text{-}169,\ p_{j7}]$、$W_{j8}[p_{j8}\text{-}169,\ p_{j8}]$、$W_{j9}[p_{j9}\text{-}169,\ p_{j9}]$、$W_{j10}[p_{j10}\text{-}169,\ p_{j10}]$和$W_{j11}[p_{j11}\text{-}169,\ p_{j11}]$。其中，$p_{jx}$与潜在分割点$k_j$之间距离$d_x$个字节，具体的，$p_{j1}$与$k_j$间距3个字节、$p_{j2}$与$k_j$间距2个字节、$p_{j3}$与$k_j$间距1个字节、$p_{j4}$与$k_j$间距0个字节、$p_{j5}$与$k_j$间距1个字节、$p_{j6}$与$k_j$间距2个字节、$p_{j7}$与$k_j$间距3个字节、$p_{j8}$与$k_j$间距4个字节、$p_{j9}$与$k_j$间距5个字节、$p_{j10}$与$k_j$间距6个字节，$p_{j11}$与$k_j$间距7个字节，并且$p_{j5}$、$p_{j6}$、$p_{j7}$、$p_{j8}$、$p_{j9}$、$p_{j10}$和$p_{j11}$相对于潜在分割点$k_j$均位于数据流分割点查找反方向，$p_{j1}$、$p_{j2}$和$p_{j3}$相对于潜在分割点$k_j$均位于数据流分割点查找方向。判断$W_{j1}[p_{j1}\text{-}169,p_{j1}]$中至少部分数据是否满足预定条件$C_1$、判断$W_{j2}[p_{j2}\text{-}169,p_{j2}]$中至少部分数据是否满足预定条件$C_2$、判断$W_{j3}[p_{j3}\text{-}169,p_{j3}]$中至少部分数据是否满足预定条件$C_3$、判断$W_{j4}[p_{j4}\text{-}169,p_{j4}]$中至少部分数据是否满足预定条件$C_4$、判断$W_{j5}[p_{j5}\text{-}169,p_{j5}]$中至少部分数据是否满足预定条件$C_5$、判断$W_{j6}[p_{j6}\text{-}169,\ p_{j6}]$中至少部分数据是否满足预定条件$C_6$、判断$W_{j7}[p_{j7}\text{-}169,\ p_{j7}]$中至少部分数据是否满足预定条件$C_7$、判断$W_{j8}[p_{j8}\text{-}169,\ p_{j8}]$中至少部分数据是否满足预定条件$C_8$、判断$W_{j9}[p_{j9}\text{-}169,\ p_{j9}]$中至少部分数据是否满足预定条件$C_9$、判断$W_{j10}[p_{j10}\text{-}169,\ p_{j10}]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{j11}[p_{j11}\text{-}169,\ p_{j11}]$中至少部分数据是否满足预定条件$C_{11}$。当然在本发明实施例中，判断潜在分割点$k_a$是否为数据流分割点时也遵循该原则，具体实现不再描述，可以参照判断潜在分割点k_i的描述。当判断窗口W_{j1}中至少部分数据满足预定条件C_1、窗口W_{j2}中至少部分数据满足预定条件C_2、窗口W_{j3}中至少部分数据满足预定条件C_3、窗口W_{j4}中至少部分数据满足预定条件C_4、窗口W_{j5}中至少部分数据满足

预定条件C_5、窗口W_{j6}中至少部分数据满足预定条件C_6、窗口W_{j7}中至少部分数据满足预定条件C_7、窗口W_{j8}中至少部分数据满足预定条件C_8、窗口W_{j9}中至少部分数据满足预定条件C_9、窗口W_{j10}中至少部分数据满足预定条件C_{10}和窗口W_{j11}中至少部分数据满足预定条件C_{11}时，则当前潜在分割点k_j为数据流分割点，k_j与k_a之间的数据构成1个数据块，同时按照与k_a相同的方式跳过最小分块大小4KB，获得下一个潜在分割点，并按照在去重服务器103上预设的规则，判断下一个潜在分割点是否为数据流分割点。当判断潜在分割点k_j不是数据流分割点时，按照与k_i相同的方式跳跃8个字节获得下一个潜在分割点，并按照在去重服务器103上预设的规则及上述方法判断下一个潜在分割点是否为数据流分割点。当超过设定的最大数据块仍然没有找到数据流分割点时，则从最大数据块的结束位置作为强制分割点。

在图3所示的数据流分割点查找的基础上，在图11所示的实施方式中，在去重服务器103上预设有规则，所述规则为：为潜在分割点k确定11个点p_x、点p_x对应的窗口$W_x[p_x-A_x,p_x+B_x]$和窗口$W_x[p_x-A_x,p_x+B_x]$对应的预定条件C_x，其中$A_1=A_2=A_3=A_4=A_5=A_6=A_7=A_8=A_9=A_{10}=169$，$A_{11}=182$，$B_1=B_2=B_3=B_4=B_5=B_6=B_7=B_8=B_9=B_{10}=B_{11}=0$，并且$C_1=C_2=C_3=C_4=C_5=C_6=C_7=C_8=C_9=C_{10}\neq C_{11}$。其中，$p_x$与潜在分割点$k$之间距离$d_x$个字节，具体的，$p_1$与潜在分割点$k$之间距离0个字节，$p_2$与$k$之间距离1个字节，$p_3$与$k$之间距离2个字节，$p_4$与$k$之间距离3个字节，$p_5$与$k$之间距离4个字节，$p_6$与$k$之间距离5个字节，$p_7$与$k$之间距离6个字节，$p_8$与$k$之间距离7个字节，$p_9$与$k$之间距离8个字节，$p_{10}$与$k$之间距离1个字节，$p_{11}$与$k$之间距离3个字节，并且、$p_2$、$p_3$、$p_4$、$p_5$、$p_6$、$p_7$、$p_8$和$p_9$相对于潜在分割点$k$均位于数据流分割点查找反方向，$p_{10}$和

p_{11}相对于潜在分割点k均位于数据流分割点查找方向。k_a为数据流分割点，图11中所示数据流分割点查找方向为从左向右，从数据流分割点k_a跳过最小数据块4KB后，最小数据块4KB结束位置作为下一个潜在分割点k_i，为潜在分割点k_i确定点p_{ix}，在本实施例中，根据在去重服务器103上预设的规则，x分别为1到11连续的自然数。在图11所示的实施方式中，为潜在分割点k_i确定的点为11个，分别为p_{i1}、p_{i2}、p_{i3}、p_{i4}、p_{i5}、p_{i6}、p_{i7}、p_{i8}、p_{i9}、p_{i10}和p_{i11}，点p_{i1}、p_{i2}、p_{i3}、p_{i4}、p_{i5}、p_{i6}、p_{i7}、p_{i8}、p_{i9}、p_{i10}和p_{i11}对应的窗口分别为$W_{i1}[p_{i1}\text{-}169, p_{i1}]$、$W_{i2}[p_{i2}\text{-}169, p_{i2}]$、$W_{i3}[p_{i3}\text{-}169, p_{i3}]$、$W_{i4}[p_{i4}\text{-}169, p_{i4}]$、$W_{i5}[p_{i5}\text{-}169, p_{i5}]$、$W_{i6}[p_{i6}\text{-}169, p_{i6}]$、$W_{i7}[p_{i7}\text{-}169, p_{i7}]$、$W_{i8}[p_{i8}\text{-}169, p_{i8}]$、$W_{i9}[p_{i9}\text{-}169, p_{i9}]$、$W_{i10}[p_{i10}\text{-}169, p_{i10}]$和$W_{i11}[p_{i11}\text{-}182, p_{i11}]$。其中，$p_{ix}$与潜在分割点$k_i$之间距离$d_x$个字节，具体的，$p_{i1}$与$k_i$间距0个字节、$p_{i2}$与$k_i$间距1个字节、$p_{i3}$与$k_i$间距2个字节、$p_{i4}$与$k_i$间距3个字节、$p_{i5}$与$k_i$间距4个字节、$p_{i6}$与$k_i$间距5个字节、$p_{i7}$与$k_i$间距6个字节、$p_{i8}$与$k_i$间距7个字节、$p_{i9}$与$k_i$间距8个字节、$p_{i10}$与$k_i$间距1个字节，$p_{i11}$与$k_i$间距3个字节，并且$p_{i2}$、$p_{i3}$、$p_{i4}$、$p_{i5}$、$p_{i6}$、$p_{i7}$、$p_{i8}$和$p_{i9}$相对于潜在分割点$k_i$均位于数据流分割点查找反方向，$p_{i10}$和$p_{i11}$相对于潜在分割点$k_i$均位于数据流分割点查找方向。判断$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据是否满足预定条件$C_1$、判断$W_{i2}[p_{i2}\text{-}169, p_{i2}]$中至少部分数据是否满足预定条件$C_2$、判断$W_{i3}[p_{i3}\text{-}169, p_{i3}]$中至少部分数据是否满足预定条件$C_3$、判断$W_{i4}[p_{i4}\text{-}169, p_{i4}]$中至少部分数据是否满足预定条件$C_4$、判断$W_{i5}[p_{i5}\text{-}169, p_{i5}]$中至少部分数据是否满足预定条件$C_5$、判断$W_{i6}[p_{i6}\text{-}169, p_{i6}]$中至少部分数据是否满足预定条件$C_6$、判断$W_{i7}[p_{i7}\text{-}169, p_{i7}]$中至少部分数据是否满足预定条件$C_7$、判断$W_{i8}[p_{i8}\text{-}169, p_{i8}]$中至少部分数据是否满足预定条件

C_8、判断$W_{i9}[p_{i9}\text{-}169, p_{i9}]$中至少部分数据是否满足预定条件$C_9$、判断$W_{i10}[p_{i10}\text{-}169, p_{i10}]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{i11}[p_{i11}\text{-}169, p_{i11}]$中至少部分数据是否满足预定条件$C_{11}$。当判断窗口$W_{i1}$中至少部分数据满足预定条件$C_1$、窗口$W_{i2}$中至少部分数据满足预定条件$C_2$、窗口$W_{i3}$中至少部分数据满足预定条件$C_3$、窗口$W_{i4}$中至少部分数据满足预定条件$C_4$、窗口$W_{i5}$中至少部分数据满足预定条件$C_5$、窗口$W_{i6}$中至少部分数据满足预定条件$C_6$、窗口$W_{i7}$中至少部分数据满足预定条件$C_7$、窗口$W_{i8}$中至少部分数据满足预定条件$C_8$、窗口$W_{i9}$中至少部分数据满足预定条件$C_9$、窗口$W_{i10}$中至少部分数据满足预定条件$C_{10}$和窗口$W_{i11}$中至少部分数据满足预定条件$C_{11}$时，则当前潜在分割点$k_i$为数据流分割点。当判断窗口$W_{i11}$中至少部分数据不满足预定条件$C_{11}$时，则从潜在分割点$k_i$沿着数据流分割点查找方向跳跃1个字节，得到新的潜在分割点，为与潜在分割点k_i区别，这里将新的潜在分割点表示为k_j。当W_{i1}、W_{i2}、W_{i3}、W_{i4}、W_{i5}、W_{i6}、W_{i7}、W_{i8}、W_{i9}和W_{i10}10个窗口中任一个窗口中至少部分数据不满足对应的预定条件时，如图12所示，$W_{i4}[p_{i4}\text{-}169, p_{i4}]$，则从点$p_{i4}$沿着数据流分割点查找方向跳跃N个字节，其中N个字节不大于$\| B_4 \| + \max_x(\| A_x \| + \| (k_i - p_{ix}) \|)$，在图12所示的实施方式中，跳跃N个字节，具体为不大于179，在本实施例中，具体取N=9，得到新的潜在分割点，为与潜在分割点k_i区别，这里将新的潜在分割点表示为k_j，根据图11所示的实施方式中在去重服务器103上预设的规则，为潜在分割点k_j确定的点为11个，分别为p_{j1}、p_{j2}、p_{j3}、p_{j4}、p_{j5}、p_{j6}、p_{j7}、p_{j8}、p_{j9}、p_{j10}和p_{j11}，确定点p_{j1}、p_{j2}、p_{j3}、p_{j4}、p_{j5}、p_{j6}、p_{j7}、p_{j8}、p_{j9}、p_{j10}和p_{j11}对应的窗口分别为$W_{j1}[p_{j1}\text{-}169, p_{j1}]$、$W_{j2}[p_{j2}\text{-}169, p_{j2}]$、$W_{j3}[p_{j3}\text{-}169, p_{j3}]$、$W_{j4}[p_{j4}\text{-}169, p_{j4}]$、$W_{j5}[p_{j5}\text{-}169,$

p_{j5}]、W_{j6}[p_{j6}-169, p_{j6}]、W_{j7}[p_{j7}-169, p_{j7}]、W_{j8}[p_{j8}-169, p_{j8}]、W_{j9}[p_{j9}-169, p_{j9}]、W_{j10}[p_{j10}-169, p_{j10}]和W_{j11}[p_{j8}-182, p_{j8}]。其中，p_{jx}与潜在分割点k_j之间距离d_x个字节，具体的，p_{j1}与k_j间距0个字节、p_{j2}与k_j间距1个字节、p_{j3}与k_j间距2个字节、p_{j4}与k_j间距3个字节、p_{j5}与k_j间距4个字节、p_{j6}与k_j间距5个字节、p_{j7}与k_j间距6个字节、p_{j8}与k_j间距7个字节、p_{j9}与k_j间距8个字节、p_{j10}与k_j间距1个字节，p_{j11}与k_j间距3个字节，并且p_{j2}、p_{j3}、p_{j4}、p_{j5}、p_{j6}、p_{j7}、p_{j8}和p_{j9}相对于潜在分割点k_j均位于数据流分割点查找反方向，p_{j10}和p_{j11}相对于潜在分割点k_j均位于数据流分割点查找方向。判断W_{j1}[p_{j1}-169,p_{j1}]中至少部分数据是否满足预定条件C_1、判断W_{j2}[p_{j2}-169,p_{j2}]中至少部分数据是否满足预定条件C_2、判断W_{j3}[p_{j3}-169,p_{j3}]中至少部分数据是否满足预定条件C_3、判断W_{j4}[p_{j4}-169,p_{j4}]中至少部分数据是否满足预定条件C_4、判断W_{j5}[p_{j5}-169,p_{j5}]中至少部分数据是否满足预定条件C_5、判断W_{j6}[p_{j6}-169, p_{j6}]中至少部分数据是否满足预定条件C_6、判断W_{j7}[p_{j7}-169, p_{j7}]中至少部分数据是否满足预定条件C_7、判断W_{j8}[p_{j8}-169, p_{j8}]中至少部分数据是否满足预定条件C_8、判断W_{j9}[p_{j9}-169, p_{j9}]中至少部分数据是否满足预定条件C_9、判断W_{j10}[p_{j10}-169, p_{j10}]中至少部分数据是否满足预定条件C_{10}和判断W_{j11}[p_{j11}-182, p_{j11}]中至少部分数据是否满足预定条件C_{11}。当然在本发明实施例中，判断潜在分割点k_a是否为数据流分割点时也遵循该原则，具体实现不再描述，可以参照判断潜在分割点k_i的描述。当判断窗口W_{j1}中至少部分数据满足预定条件C_1、窗口W_{j2}中至少部分数据满足预定条件C_2、窗口W_{j3}中至少部分数据满足预定条件C_3、窗口W_{j4}中至少部分数据满足预定条件C_4、窗口W_{j5}中

至少部分数据满足预定条件C_5、窗口W_{j6}中至少部分数据满足预定条件C_6、窗口W_{j7}中至少部分数据满足预定条件C_7、窗口W_{j8}中至少部分数据满足预定条件C_8、窗口W_{j9}中至少部分数据满足预定条件C_9、窗口W_{j10}中至少部分数据满足预定条件C_{10}和窗口W_{j11}中至少部分数据满足预定条件C_{11}时，则当前潜在分割点k_j为数据流分割点，k_j与k_a之间的数据构成1个数据块，同时按照与k_a相同的方式跳过最小分块大小4KB，获得下一个潜在分割点，并按照在去重服务器103上预设的规则，判断下一个潜在分割点是否为数据流分割点。当判断潜在分割点k_j不是数据流分割点时，按照与k_i相同的方式获得下一个潜在分割点，并按照在去重服务器103上预设的规则及上述方法判断下一个潜在分割点是否为数据流分割点。当超过设定的最大数据块仍然没有找到数据流分割点时，则从最大数据块的结束位置作为强制分割点。

在图3所示的数据流分割点查找的基础上，在图13所示的实施方式中，在去重服务器103上预设有规则为：为潜在分割点k确定11个点p_x、点p_x对应的窗口$W_x[p_x-A_x,p_x+B_x]$和窗口$W_x[p_x-A_x,p_x+B_x]$对应的预定条件C_x，x分别为1到11连续的自然数，其中，点p_x对应的窗口$W_x[p_x-A_x,p_x+B_x]$中至少部分数据满足预定条件的概率为1/2，并且$A_1=A_2=A_3=A_4=A_5=A_6=A_7=A_8=A_9=A_{10}=A_{11}=169$，$B_1=B_2=B_3=B_4=B_5=B_6=B_7=B_8=B_9=B_{10}=B_{11}=0$，并且$C_1=C_2=C_3=C_4=C_5=C_6=C_7=C_8=C_9=C_{10}=C_{11}$，其中，$p_x$与潜在分割点$k$之间距离$d_x$个字节，具体的，$p_1$与潜在分割点$k$之间距离0个字节，$p_2$与$k$之间距离2个字节，$p_3$与$k$之间距离4个字节，$p_4$与$k$之间距离6个字节，$p_5$与$k$之间距离8个字节，$p_6$与$k$之间距离10个字节，$p_7$与$k$之间距离12个字节，$p_8$与$k$之间距离14个字节，$p_9$与$k$之间距离16个字节，$p_{10}$与$k$之间距离18个字节，$p_{11}$与$k$之间距离20个字节，并

且p_2、p_3、p_4、p_5、p_6、p_7、p_8、p_9、p_{10}和p_{11}相对于潜在分割点k均位于数据流分割点查找反方向。k_a为数据流分割点，图13中所示数据流分割点查找方向为从左向右，从数据流分割点k_a跳过最小数据块4KB后，在最小数据块4KB结束位置作为下一个潜在分割点k_i，为潜在分割点k_i确定点p_{ix}，在本实施例中，根据在去重服务器103上预设的规则，x分别为1到11连续的自然数。在图13所示的实施方式中，依据预定规则，为潜在分割点k_i确定的点为11个，分别为p_{i1}、p_{i2}、p_{i3}、p_{i4}、p_{i5}、p_{i6}、p_{i7}、p_{i8}、p_{i9}、p_{i10}和p_{i11}，点p_{i1}、p_{i2}、p_{i3}、p_{i4}、p_{i5}、p_{i6}、p_{i7}、p_{i8}、p_{i9}、p_{i10}和p_{i11}对应的窗口分别为$W_{i1}[p_{i1}\text{-}169, p_{i1}]$、$W_{i2}[p_{i2}\text{-}169, p_{i2}]$、$W_{i3}[p_{i3}\text{-}169, p_{i3}]$、$W_{i4}[p_{i4}\text{-}169, p_{i4}]$、$W_{i5}[p_{i5}\text{-}169, p_{i5}]$、$W_{i6}[p_{i6}\text{-}169, p_{i6}]$、$W_{i7}[p_{i7}\text{-}169, p_{i7}]$、$W_{i8}[p_{i8}\text{-}169, p_{i8}]$、$W_{i9}[p_{i9}\text{-}169, p_{i9}]$、$W_{i10}[p_{i10}\text{-}169, p_{i10}]$和$W_{i11}[p_{i11}\text{-}169, p_{i11}]$。其中，$p_{ix}$与潜在分割点$k_i$之间距离$d_x$个字节，具体的，$p_{i1}$与$k_i$间距0个字节、$p_{i2}$与$k_i$间距2个字节、$p_{i3}$与$k_i$间距4个字节、$p_{i4}$与$k_i$间距6个字节、$p_{i5}$与$k_i$间距8个字节、$p_{i6}$与$k_i$间距10个字节、$p_{i7}$与$k_i$间距12个字节、$p_{i8}$与$k_i$间距14个字节、$p_{i9}$与$k_i$间距16个字节、$p_{i10}$与$k_i$间距18个字节，$p_{i11}$与$k_i$间距20个字节，并且$p_{i2}$、$p_{i3}$、$p_{i4}$、$p_{i5}$、$p_{i6}$、$p_{i7}$、$p_{i8}$、$p_{i9}$、$p_{i10}$和$p_{i11}$相对于潜在分割点$k_i$均位于数据流分割点查找反方向。判断$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据是否满足预定条件$C_1$、判断$W_{i2}[p_{i2}\text{-}169, p_{i2}]$中至少部分数据是否满足预定条件$C_2$、判断$W_{i3}[p_{i3}\text{-}169, p_{i3}]$中至少部分数据是否满足预定条件$C_3$、判断$W_{i4}[p_{i4}\text{-}169, p_{i4}]$中至少部分数据是否满足预定条件$C_4$、判断$W_{i5}[p_{i5}\text{-}169, p_{i5}]$中至少部分数据是否满足预定条件$C_5$、判断$W_{i6}[p_{i6}\text{-}169, p_{i6}]$中至少部分数据是否满足预定条件$C_6$、判断$W_{i7}[p_{i7}\text{-}169, p_{i7}]$中至少部分数据是否满足预定条件$C_7$、判断$W_{i8}[p_{i8}\text{-}169, p_{i8}]$中至少部分数据是

否满足预定条件C_8、判断$W_{i9}[p_{i9}\text{-}169,p_{i9}]$中至少部分数据是否满足预定条件$C_9$、判断$W_{i10}[p_{i10}\text{-}169,p_{i10}]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{i11}[p_{i11}\text{-}169,p_{i11}]$中至少部分数据是否满足预定条件$C_{11}$。当判断窗口$W_{i1}$中至少部分数据满足预定条件$C_1$、窗口$W_{i2}$中至少部分数据满足预定条件$C_2$、窗口$W_{i3}$中至少部分数据满足预定条件$C_3$、窗口$W_{i4}$中至少部分数据满足预定条件$C_4$、窗口$W_{i5}$中至少部分数据满足预定条件$C_5$、窗口$W_{i6}$中至少部分数据满足预定条件$C_6$、窗口$W_{i7}$中至少部分数据满足预定条件$C_7$、窗口$W_{i8}$中至少部分数据满足预定条件$C_8$、窗口$W_{i9}$中至少部分数据满足预定条件$C_9$、窗口$W_{i10}$中至少部分数据满足预定条件$C_{10}$和窗口$W_{i11}$中至少部分数据满足预定条件$C_{11}$时，则当前潜在分割点$k_i$为数据流分割点。当11个窗口中任一个窗口中至少部分数据不满足对应的预定条件时，如图14所示，$W_{i4}[p_{i4}\text{-}169,p_{i4}]$中至少部分数据不满足预定条件$C_4$，则选择下一个潜在分割点，为与潜在分割点$k_i$区别，这里表示为$k_j$，$k_j$位于$k_i$右边，并且$k_j$与$k_i$间距1个字节。如图14所示，依据在去重服务器103上预设的规则，为潜在分割点k_j确定11个点，分别为p_{j1}、p_{j2}、p_{j3}、p_{j4}、p_{j5}、p_{j6}、p_{j7}、p_{j8}、p_{j9}、p_{j10}和p_{j11}，确定点p_{j1}、p_{j2}、p_{j3}、p_{j4}、p_{j5}、p_{j6}、p_{j7}、p_{j8}、p_{j9}、p_{j10}和p_{j11}对应的窗口分别为$W_{j1}[p_{j1}\text{-}169,p_{j1}]$、$W_{j2}[p_{j2}\text{-}169,p_{j2}]$、$W_{j3}[p_{j3}\text{-}169,p_{j3}]$、$W_{j4}[p_{j4}\text{-}169,p_{j4}]$、$W_{j5}[p_{j5}\text{-}169,p_{j5}]$、$W_{j6}[p_{j6}\text{-}169,\ p_{j6}]$、$W_{j7}[p_{j7}\text{-}169,\ p_{j7}]$、$W_{j8}[p_{j8}\text{-}169,\ p_{j8}]$、$W_{j9}[p_{j9}\text{-}169,\ p_{j9}]$、$W_{j10}[p_{j10}\text{-}169,\ p_{j10}]$和$W_{j11}[p_{j11}\text{-}169,\ p_{j11}]$，其中，$A_1=A_2=A_3=A_4=A_5=A_6=A_7=A_8=A_9=A_{10}=A_{11}=169$，$B_1=B_2=B_3=B_4=B_5=B_6=B_7=B_8=B_9=B_{10}=B_{11}=0$，并且$C_1=C_2=C_3=C_4=C_5=C_6=C_7=C_8=C_9=C_{10}=C_{11}$。其中，$p_{jx}$与潜在分割点$k_j$之间距离$d_x$个字节，具体的，$p_{j1}$与$k_j$间距0个字节、$p_{j2}$与$k_j$间距2个字节、$p_{j3}$与$k_j$间距4个

字节、p_{j4}与k_j间距6个字节、p_{j5}与k_j间距8个字节、p_{j6}与k_j间距10个字节、p_{j7}与k_j间距12个字节、p_{j8}与k_j间距14个字节、p_{j9}与k_j间距16个字节、p_{j10}与k_j间距18个字节，p_{j11}与k_j间距20个字节，并且p_{j2}、p_{j3}、p_{j4}、p_{j5}、p_{j6}、p_{j7}、p_{j8}、p_{j9}、p_{j10}和p_{j11}相对于潜在分割点k_j均位于数据流分割点查找反方向。判断$W_{j1}[p_{j1}\text{-}169, p_{j1}]$中至少部分数据是否满足预定条件$C_1$、判断$W_{j2}[p_{j2}\text{-}169, p_{j2}]$中至少部分数据是否满足预定条件$C_2$、判断$W_{j3}[p_{j3}\text{-}169, p_{j3}]$中至少部分数据是否满足预定条件$C_3$、判断$W_{j4}[p_{j4}\text{-}169, p_{j4}]$中至少部分数据是否满足预定条件$C_4$、判断$W_{j5}[p_{j5}\text{-}169, p_{j5}]$中至少部分数据是否满足预定条件$C_5$、判断$W_{j6}[p_{j6}\text{-}169, p_{j6}]$中至少部分数据是否满足预定条件$C_6$、判断$W_{j7}[p_{j7}\text{-}169, p_{j7}]$中至少部分数据是否满足预定条件$C_7$、判断$W_{j8}[p_{j8}\text{-}169, p_{j8}]$中至少部分数据是否满足预定条件$C_8$、判断$W_{j9}[p_{j9}\text{-}169, p_{j9}]$中至少部分数据是否满足预定条件$C_9$、判断$W_{j10}[p_{j10}\text{-}169, p_{j10}]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{j11}[p_{j11}\text{-}169, p_{j11}]$中至少部分数据是否满足预定条件$C_{11}$。当判断窗口$W_{j1}$中至少部分数据满足预定条件$C_1$、窗口$W_{j2}$中至少部分数据满足预定条件$C_2$、窗口$W_{j3}$中至少部分数据满足预定条件$C_3$、窗口$W_{j4}$中至少部分数据满足预定条件$C_4$、窗口$W_{j5}$中至少部分数据满足预定条件$C_5$、窗口$W_{j6}$中至少部分数据满足预定条件$C_6$、窗口$W_{j7}$中至少部分数据满足预定条件$C_7$、窗口$W_{j8}$中至少部分数据满足预定条件$C_8$、窗口$W_{i9}$中至少部分数据满足预定条件$C_9$、窗口$W_{j10}$中至少部分数据满足预定条件$C_{10}$和窗口$W_{j11}$中至少部分数据满足预定条件$C_{11}$时，则当前潜在分割点$k_j$为数据流分割点。当判断窗口$W_{j1}$、$W_{j2}$、$W_{j3}$、$W_{j4}$、$W_{j5}$、$W_{j6}$、$W_{j7}$、$W_{j8}$、$W_{j9}$、$W_{j10}$和$W_{j11}$中任一个窗口中至少

部分数据不满足预定条件时，如图15所示，$W_{j3}[p_{j3}\text{-}169,\ p_{j3}]$中至少部分数据不满足预定条件$C_3$时，点$p_{i4}$相对于数据流分割点查找方向位于点$p_{j3}$左边，从点$p_{i4}$沿着数据流分割点查找方向跳跃21个字节，获得下一个潜在分割点，为与潜在分割点k_i、k_j相区别，表示为k_l。根据图13所实施方式中在去重服务器103上预设的规则，为潜在分割点k_l确定的点为11个，分别为p_{l1}、p_{l2}、p_{l3}、p_{l4}、p_{l5}、p_{l6}、p_{l7}、p_{l8}、p_{l9}、p_{l10}和p_{l11}，点p_{l1}、p_{l2}、p_{l3}、p_{l4}、p_{l5}、p_{l6}、p_{l7}、p_{l8}、p_{l9}、p_{l10}和p_{l11}对应的窗口分别为$W_{l1}[p_{l1}\text{-}169,\ p_{l1}]$、$W_{l2}[p_{l2}\text{-}169,\ p_{l2}]$、$W_{l3}[p_{l3}\text{-}169,\ p_{l3}]$、$W_{l4}[p_{l4}\text{-}169,\ p_{l4}]$、$W_{l5}[p_{l5}\text{-}169,\ p_{l5}]$、$W_{l6}[p_{l6}\text{-}169,\ p_{l6}]$、$W_{l7}[p_{l7}\text{-}169,\ p_{l7}]$、$W_{l8}[p_{l8}\text{-}169,\ p_{l8}]$、$W_{l9}[p_{l9}\text{-}169,\ p_{l9}]$、$W_{l10}[p_{l10}\text{-}169,\ p_{l10}]$和$W_{l11}[p_{l11}\text{-}169,\ p_{l11}]$，其中，$p_{lx}$与潜在分割点$k_l$之间距离$d_x$个字节，具体的，$p_{l1}$与潜在分割点$k_l$之间距离0个字节，$p_{l2}$与$k_l$之间距离2个字节，$p_{l3}$与$k_l$之间距离4个字节，$p_{l4}$与$k_l$之间距离6个字节，$p_{l5}$与$k_l$之间距离8个字节，$p_{l6}$与$k_l$之间距离10个字节，$p_{l7}$与$k_l$之间距离12个字节，$p_{l8}$与$k_l$之间距离14个字节，$p_{l9}$与$k_l$之间距离16个字节，$p_{l10}$与$k_l$之间距离18个字节，$p_{l11}$与$k_l$之间距离20个字节，并且$p_{l2}$、$p_{l3}$、$p_{l4}$、$p_{l5}$、$p_{l6}$、$p_{l7}$、$p_{l8}$、$p_{l9}$、$p_{l10}$和$p_{l11}$相对于潜在分割点$k_l$均位于数据流分割点查找反方向。判断$W_{l1}[p_{l1}\text{-}169,\ p_{l1}]$中至少部分数据是否满足预定条件$C_1$、判断$W_{l2}[p_{l2}\text{-}169,\ p_{l2}]$中至少部分数据是否满足预定条件$C_2$、判断$W_{l3}[p_{l3}\text{-}169,\ p_{l3}]$中至少部分数据是否满足预定条件$C_3$、判断$W_{l4}[p_{l4}\text{-}169,\ p_{l4}]$中至少部分数据是否满足预定条件$C_4$、判断$W_{l5}[p_{l5}\text{-}169,\ p_{l5}]$中至少部分数据是否满足预定条件$C_5$、判断$W_{l6}[p_{l6}\text{-}169,\ p_{l6}]$中至少部分数据是否满足预定条件$C_6$、判断$W_{l7}[p_{l7}\text{-}169,\ p_{l7}]$中至少部分数据是否满足预定条件$C_7$、判断$W_{l8}[p_{l8}\text{-}169,\ p_{l8}]$中至少部分数据是否满足预

定条件C_8、判断$W_{l9}[p_{l9}\text{-}169,\ p_{l9}]$中至少部分数据是否满足预定条件$C_9$、判断$W_{l10}[p_{l10}\text{-}169,\ p_{l10}]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{l11}[p_{l11}\text{-}169,\ p_{l11}]$中至少部分数据是否满足预定条件$C_{11}$。当判断窗口$W_{l1}$中至少部分数据满足预定条件$C_1$、窗口$W_{l2}$中至少部分数据满足预定条件$C_2$、窗口$W_{l3}$中至少部分数据满足预定条件$C_3$、窗口$W_{l4}$中至少部分数据满足预定条件$C_4$、窗口$W_{l5}$中至少部分数据满足预定条件$C_5$、窗口$W_{l6}$中至少部分数据满足预定条件$C_6$、窗口$W_{l7}$中至少部分数据满足预定条件$C_7$、窗口$W_{l8}$中至少部分数据满足预定条件$C_8$、窗口$W_{l9}$中至少部分数据满足预定条件$C_9$、窗口$W_{l10}$中至少部分数据满足预定条件$C_{10}$和窗口$W_{l11}$中至少部分数据满足预定条件$C_{11}$时，则当前潜在分割点$k_l$为数据流分割点。当窗口$W_{l1}$、$W_{l2}$、$W_{l3}$、$W_{l4}$、$W_{l5}$、$W_{l6}$、$W_{l7}$、$W_{l8}$、$W_{l9}$、$W_{l10}$和$W_{l11}$中任一窗口中至少部分数据不满足预定条件时，选择下一个潜在分割点，为与潜在分割点k_i、k_j和k_l区别，表示为k_m，k_m位于k_l右边，并且k_m与k_l间距1个字节。根据图13所示实施例在去重服务器103上预设的规则，为潜在分割点k_m确定的点为11个，分别为p_{m1}、p_{m2}、p_{m3}、p_{m4}、p_{m5}、p_{m6}、p_{m7}、p_{m8}、p_{m9}、p_{m10}和p_{m11}，点p_{m1}、p_{m2}、p_{m3}、p_{m4}、p_{m5}、p_{m6}、p_{m7}、p_{m8}、p_{m9}、p_{m10}和p_{m11}对应的窗口分别为$W_{m1}[p_{m1}\text{-}169,\ p_{m1}]$、$W_{m2}[p_{m2}\text{-}169,\ p_{m2}]$、$W_{m3}[p_{m3}\text{-}169,\ p_{m3}]$、$W_{m4}[p_{m4}\text{-}169,\ p_{m4}]$、$W_{m5}[p_{m5}\text{-}169,\ p_{m5}]$、$W_{m6}[p_{m6}\text{-}169,\ p_{m6}]$、$W_{m7}[p_{m7}\text{-}169,\ p_{m7}]$、$W_{m8}[p_{m8}\text{-}169,\ p_{m8}]$、$W_{m9}[p_{m9}\text{-}169,\ p_{m9}]$、$W_{m10}[p_{m10}\text{-}169,\ p_{m10}]$和$W_{m11}[p_{m11}\text{-}169,\ p_{m11}]$，其中，$p_{mx}$与潜在分割点$k_m$之间距离$d_x$个字节，具体的，$p_{m1}$与潜在分割点$k_m$之间距离0个字节，$p_{m2}$与$k_m$之间距离2个字节，$p_{m3}$与$k_m$之间距离4个字节，$p_{m4}$与$k_m$之间距离6个字节，$p_{m5}$与$k_m$之间距离8个字节，$p_{m6}$与$k_m$之间距离10个字节，$p_{m7}$与$k_m$之间距离12个字

节，p_{m8}与k_m之间距离14个字节，p_{m9}与k_m之间距离16个字节，p_{m10}与k_m之间距离18个字节，p_{m11}与k_m之间距离20个字节，并且p_{m2}、p_{m3}、p_{m4}、p_{m5}、p_{m6}、p_{m7}、p_{m8}、p_{m9}、p_{m10}和p_{m11}相对于潜在分割点k_m均位于数据流分割点查找反方向。判断$W_{m1}[p_{m1}\text{-}169, p_{m1}]$中至少部分数据是否满足预定条件$C_1$、判断$W_{m2}[p_{m2}\text{-}169, p_{m2}]$中至少部分数据是否满足预定条件$C_2$、判断$W_{m3}[p_{m3}\text{-}169, p_{m3}]$中至少部分数据是否满足预定条件$C_3$、判断$W_{m4}[p_{m4}\text{-}169, p_{m4}]$中至少部分数据是否满足预定条件$C_4$、判断$W_{m5}[p_{m5}\text{-}169, p_{m5}]$中至少部分数据是否满足预定条件$C_5$、判断$W_{m6}[p_{m6}\text{-}169, p_{m6}]$中至少部分数据是否满足预定条件$C_6$、判断$W_{m7}[p_{m7}\text{-}169, p_{m7}]$中至少部分数据是否满足预定条件$C_7$、判断$W_{m8}[p_{m8}\text{-}169, p_{m8}]$中至少部分数据是否满足预定条件$C_8$、判断$W_{m9}[p_{m9}\text{-}169, p_{m9}]$中至少部分数据是否满足预定条件$C_9$、判断$W_{m10}[p_{m10}\text{-}169, p_{m10}]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{m11}[p_{m11}\text{-}169, p_{m11}]$中至少部分数据是否满足预定条件$C_{11}$。当判断窗口$W_{m1}$中至少部分数据满足预定条件$C_1$、窗口$W_{m2}$中至少部分数据满足预定条件$C_2$、窗口$W_{m3}$中至少部分数据满足预定条件$C_3$、窗口$W_{m4}$中至少部分数据满足预定条件$C_4$、窗口$W_{m5}$中至少部分数据满足预定条件$C_5$、窗口$W_{m6}$中至少部分数据满足预定条件$C_6$、窗口$W_{m7}$中至少部分数据满足预定条件$C_7$、窗口$W_{m8}$中至少部分数据满足预定条件$C_8$、窗口$W_{m9}$中至少部分数据满足预定条件$C_9$、窗口$W_{m10}$中至少部分数据满足预定条件$C_{10}$和窗口$W_{m11}$中至少部分数据满足预定条件$C_{11}$时，则当前潜在分割点$k_m$为数据流分割点。当任一个窗口中至少部分数据不满足预定条件时，则按照前面描述的方案执行跳跃，以获得下一个潜在分割点并判断是否为数据流分割点。

本发明实施例提供了一种判断窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部

分数据是否满足预定条件C_z的方法，本实施例中使用随机函数判断窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足预定条件$C_z$，以图5所示的实施方式为例，根据在去重服务器103上预设的规则，为潜在分割点k_i确定点p_{i1}及点p_{i1}对应的窗口$W_{i1}[p_{i1}\text{-}169, p_{i1}]$，判断$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据是否满足预定的条件$C_1$，如图16所示，$W_{i1}$表示窗口$W_{i1}[p_{i1}\text{-}169, p_{i1}]$，为判断$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据是否满足预定条件$C_1$，选择5个字节，图16中“■”表示选择的1个字节，相邻两个选择的字节之间相差42个字节。将选择的5字节数据反复利用51次，共获得255字节，以增加随机性。其中每个字节由8位组成，记为$a_{m,1}\ldots a_{m,8}$，表示255个字节中第m个字节的第1到第8位，因此，255个字节对应的位可以表示为：$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,8} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ a_{255,1} & a_{255,2} & \cdots & a_{255,8} \end{pmatrix}$，当$a_{m,n}$=1时，$V_{am,n}$ =1，当$a_{m,n}$=0时，$V_{am,n}$=-1，其中$a_{m,n}$表示$a_{m,1}\ldots a_{m,8}$中的任一个，255个字节对应的位按照$a_{m,n}$与$V_{am,n}$的转换关系得到矩阵V_a，可以表示为：

$\begin{pmatrix} V_{a1,1} & V_{a1,2} & \cdots & V_{a1,8} \\ V_{a2,1} & V_{a2,2} & \cdots & V_{a2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{a255,1} & V_{a255,2} & \cdots & V_{a255,8} \end{pmatrix}$。选取大量随机数，组成矩阵，由随机数据组成的矩阵一旦组成，保持不变，如从服从特定分布（这里以正态分布为例）的随机数中选择255*8个随机数组成矩阵R:

$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,8} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{255,1} & h_{255,2} & \cdots & h_{255,8} \end{pmatrix}$，将矩阵$V_a$的第m行与矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为$S_{am}=V_{am,1}*h_{m,1}+V_{am,2}*h_{m,2}+\ldots+V_{am,8}$

$*h_{m,8}$。根据该方法，获得S_{a1}、S_{a2}...到S_{a255}，统计S_{a1}、S_{a2}...到S_{a255}中满足特定条件（这里以大于0为例）的值的个数K。由于矩阵R服从正态分布，则S_{am}与矩阵R一样，仍然服从正态分布，根据概率论，正态分布随机数大于0的概率为1/2，在S_{a1}、S_{a2}...到S_{a255}中，每个值大于0的概率为1/2，所以K满足二项分布：$P(k=n)=C_{255}^{n}(\frac{1}{2})^{n}(\frac{1}{2})^{255-n}=C_{255}^{n}(\frac{1}{2})^{255}$。根据统计结果，判断$S_{a1}$、$S_{a2}$...到$S_{a255}$的值大于0的个数K是否为偶数，二项分布的随机数为偶数的概率为为1/2，所以K以1/2的概率满足条件。当K为偶数时，表明$W_{i1}[p_{i1}\text{-}169,p_{i1}]$中至少部分数据满足预定条件$C_1$；当K为奇数时，表明$W_{i1}[p_{i1}\text{-}169,p_{i1}]$中至少部分数据不满足预定条件$C_1$,这里$C_1$即指根据上述方式获得的$S_{a1}$、$S_{a2}$...到$S_{a255}$的值大于0的个数K为偶数。在图5所示的实施方式中，在$W_{i1}[p_{i1}\text{-}169,p_{i1}]$、$W_{i2}[p_{i2}\text{-}169,p_{i2}]$、$W_{i3}[p_{i3}\text{-}169,p_{i3}]$、$W_{i4}[p_{i4}\text{-}169,p_{i4}]$、$W_{i5}[p_{i5}\text{-}169,p_{i5}]$、$W_{i6}[p_{i6}\text{-}169,p_{i6}]$、$W_{i7}[p_{i7}\text{-}169,p_{i7}]$、$W_{i8}[p_{i8}\text{-}169,p_{i8}]$、$W_{i9}[p_{i9}\text{-}169,p_{i9}]$、$W_{i10}[p_{i10}\text{-}169,p_{i10}]$和$W_{i11}[p_{i11}\text{-}169,p_{i11}]$中，各窗口大小相同,即窗口大小均为169字节，同时判断窗口中至少部分数据是否满足预定条件的方式也相同,具体见上述判断$W_{i1}[p_{i1}\text{-}169,p_{i1}]$中至少部分数据是否满足预定条件$C_1$的描述。因此，如图16所示，“⊠”表示判断窗口$W_{i2}[p_{i2}\text{-}169,p_{i2}]$中至少部分数据是否满足预定条件$C_2$时选择的1个字节，相邻两个选择的字节之间相差42个字节。将选择的5字节数据反复利用51次，共获得255字节，以增加随机性。其中每个字节由8位组成，记为$b_{m,1}...b_{m,8}$,表示255个字节中第m个字节的第1到第8位，因此，255个字节对应的位可以表示为：$\begin{pmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,8} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ b_{255,1} & b_{255,2} & \cdots & b_{255,8} \end{pmatrix}$，当$b_{m,n}$=1时，$V_{bm,n}$=1，当$b_{m,n}$=0时，$V_{bm,n}$

=-1，其中$b_{m,n}$表示$b_{m,1}\ldots b_{m,8}$中的任一个，255个字节对应的位按照$b_{m,n}$与$V_{bm,n}$的转换关系得到矩阵V_b，可以表示为：$\begin{pmatrix} V_{b1,1} & V_{b1,2} & \cdots & V_{b1,8} \\ V_{b2,1} & V_{b2,2} & \cdots & V_{b2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{b255,1} & V_{b255,2} & \cdots & V_{b255,8} \end{pmatrix}$。

判断$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据是否满足预定条件的方式与判断窗口$W_{i2}[p_{i2}\text{-}169, p_{i2}]$中至少部分数据是否满足预定条件的方式相同，因此使用矩阵R: $\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,8} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{255,1} & h_{255,2} & \cdots & h_{255,8} \end{pmatrix}$,将矩阵$V_b$的第m行与矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为$S_{bm}=V_{bm,1}*h_{m,1}+V_{bm,2}*h_{m,2}+\ldots+V_{bm,8}*h_{m,8}$。根据该方法，获得$S_{b1}$、$S_{b2}\ldots$到$S_{b255}$，统计$S_{b1}$、$S_{b2}\ldots$到$S_{b255}$中满足特定条件（这里以大于0为例）的值的个数K。由于矩阵R服从正态分布，则S_{bm}与矩阵R一样，仍然服从正态分布，根据概率论，正态分布随机数大于0的概率为1/2，在S_{b1}、$S_{b2}\ldots$到S_{b255}中，每个值大于0的概率为1/2，所以K满足二项分布：

$P(k=n)=C_{255}^{n}(\frac{1}{2})^{n}(\frac{1}{2})^{255-n}=C_{255}^{n}(\frac{1}{2})^{255}$。根据统计结果，判断$S_{b1}$、$S_{b2}\ldots$到$S_{b255}$的值大于0的个数K是否为偶数，二项分布的随机数为偶数的概率为为1/2，所以K以1/2的概率满足条件。当K为偶数时，表明$W_{i2}[p_{i2}\text{-}169, p_{i2}]$中至少部分数据满足预定条件$C_2$；当K为奇数时，表明$W_{i2}[p_{i2}\text{-}169, p_{i2}]$中至少部分数据不满足预定条件$C_2$,这里$C_2$即指根据上述方式获得的$S_{b1}$、$S_{b2}\ldots$到$S_{b255}$的值大于0的个数K为偶数。图3所示的实施方式中，$W_{i2}[p_{i2}\text{-}169, p_{i2}]$中至少部分数据满足预定条件$C_2$。

因此，如图16所示，“▨”表示判断窗口$W_{i3}[p_{i3}\text{-}169, p_{i3}]$中至少部

分数据是否满足预定条件C_3时选择的1个字节，相邻两个选择的字节之间相差42个字节。将选择的5字节数据反复利用51次，共获得255字节，以增加随机性。然后使用判断窗口$W_{i1}[p_{i1}\text{-}169, p_{i1}]$和$W_{i2}[p_{i2}\text{-}169, p_{i2}]$中至少部分数据是否满足预定条件的方法，判断$W_{i3}[p_{i3}\text{-}169, p_{i3}]$中至少数据是否满足预定条件$C_3$。图5所示的实施方式中，$W_{i3}[p_{i3}\text{-}169, p_{i3}]$中至少部分数据满足预定条件。如图16所示，“ ”表示判断窗口$W_{i4}[p_{i4}\text{-}169, p_{i4}]$中至少部分数据是否满足预定条件$C_4$时选择的1个字节，相邻两个选择的字节之间相差42个字节。将选择的5字节数据反复利用51次，共获得255字节，以增加随机性。然后使用判断窗口$W_{i1}[p_{i1}\text{-}169, p_{i1}]$、$W_{i2}[p_{i2}\text{-}169, p_{i2}]$和$W_{i3}[p_{i3}\text{-}169, p_{i3}]$中至少部分数据是否满足预定条件的方法，判断$W_{i4}[p_{i4}\text{-}169, p_{i4}]$中至少部分数据是否满足预定条件$C_4$。图5所示的实施方式中，$W_{i4}[p_{i4}\text{-}169, p_{i4}]$中至少部分数据满足预定条件$C_4$。如图16所示，“ ”表示判断窗口$W_{i5}[p_{i5}\text{-}169, p_{i5}]$中至少部分数据是否满足预定条件$C_5$时选择的1个字节，相邻两个选择的字节之间相差42个字节。将选择的5字节数据反复利用51次，共获得255字节，以增加随机性。然后使用判断窗口$W_{i1}[p_{i1}\text{-}169, p_{i1}]$、$W_{i2}[p_{i2}\text{-}169, p_{i2}]$、$W_{i3}[p_{i3}\text{-}169, p_{i3}]$和$W_{i4}[p_{i4}\text{-}169, p_{i4}]$中至少部分数据是否满足预定条件的方法，判断$W_{i5}[p_{i5}\text{-}169, p_{i5}]$中至少数据是否满足预定条件$C_5$。图5所示的实施方式中，$W_{i5}[p_{i5}\text{-}169, p_{i5}]$中至少部分数据不满足预定条件$C_5$。

当$W_{i5}[p_{i5}\text{-}169, p_{i5}]$中至少部分数据不满足预定条件时$C_5$，从点$p_{i5}$沿着数据流分割点查找方向跳跃11个字节，在第11个字节的结束位置获得下一个潜在分割点k_j，如图6所示，根据在去重服务器103上预设的规则，为潜在分割点k_j确定点p_{j1}、点p_{j1}对应的窗口$W_{j1}[p_{j1}\text{-}169, p_{j1}]$，

判断窗口 $W_{j1}[p_{j1}\text{-}169,p_{j1}]$ 中至少部分数据是否满足预定条件 C_1 的方式与判断窗口 $W_{i1}[p_{i1}\text{-}169,p_{i1}]$ 中至少部分数据是否满足预定条件 C_1 的方式相同，因此如图17所示，W_{j1} 表示窗口 $W_{j1}[p_{j1}\text{-}169,p_{j1}]$，为判断 $W_{j1}[p_{j1}\text{-}169,p_{j1}]$ 中至少部分数据是否满足预定条件 C_1，选择5个字节，图17中“■”表示选择的1个字节，相邻两个选择的字节之间相差42个字节。将选择的5字节数据反复利用51次，共获得255字节，以增加随机性。其中每个字节由8位组成，记为 $a_{m,1}'\ldots a_{m,8}'$，表示255个字节中第m个字节的第1到第8位，因此，255个字节对应的位可以表示为：

$\begin{pmatrix} a_{1,1}' & a_{1,2}' & \ldots & a_{1,8}' \\ a_{2,1}' & a_{2,2}' & \ldots & a_{2,8}' \\ \vdots & \vdots & \vdots & \vdots \\ a_{255,1}' & a_{255,2}' & \ldots & a_{255,8}' \end{pmatrix}$，当 $a_{m,n}'$=1时，$V_{am,n}'$ =1，当 $a_{m,n}'$=0时，$V_{am,n}'$=-1，其中 $a_{m,n}'$ 表示 $a_{m,1}'\ldots a_{m,8}'$ 中的任一个，255个字节对应的位按照 $a_{m,n}'$ 与 $V_{am,n}'$ 的转换关系得到矩阵 V_a'，可以表示为：$\begin{pmatrix} V_{a1,1}' & V_{a1,2}' & \cdots & V_{a1,8}' \\ V_{a2,1}' & V_{a2,2}' & \cdots & V_{a2,8}' \\ \vdots & \vdots & \vdots & \vdots \\ V_{a255,1}' & V_{a255,2}' & \cdots & V_{a255,8}' \end{pmatrix}$。

判断窗口 $W_{j1}[p_{j1}\text{-}169,p_{j1}]$ 中至少部分数据是否满足预定的条件与判断窗口 $W_{i1}[p_{i1}\text{-}169,p_{i1}]$ 中至少部分数据是否满足预定的条件的方式相同，因此使用矩阵R：$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,8} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{255,1} & h_{255,2} & \cdots & h_{255,8} \end{pmatrix}$，将矩阵 V_a' 的第m行与矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为 $S_{am}'=V_{am,1}'*h_{m,1}+V_{am,2}'*h_{m,2}+\ldots+V_{am,8}'*h_{m,8}$。根据该方法，获得 S_{a1}'、$S_{a2}'\ldots$到 S_{a255}'，统计 S_{a1}'、$S_{a2}'\ldots$到 S_{a255}' 中满足特定条件（这里以大于0为例）的值的个数K。由于矩阵R服从正态分布，则 S_{am}' 与矩阵R一样，仍然服从正

态分布，根据概率论，正态分布随机数大于0的概率为1/2，在S_{a1}'、S_{a2}'...到S_{a255}'中，每个值大于0的概率为1/2，所以K满足二项分布：$P(k=n)=C_{255}^{n}(\frac{1}{2})^{n}(\frac{1}{2})^{255-n}=C_{255}^{n}(\frac{1}{2})^{255}$。根据统计结果，判断$S_{a1}$'、$S_{a2}$'...到$S_{a255}$'的值大于0的个数K是否为偶数，二项分布的随机数为偶数的概率为1/2，所以K以1/2的概率满足条件。当K为偶数时，表明$W_{j1}[p_{j1}\text{-}169, p_{j1}]$中至少部分数据满足预定条件$C_1$；当K为奇数时，表明$W_{j1}[p_{j1}\text{-}169, p_{j1}]$中至少部分数据不满足预定条件$C_1$。

判断$W_{i2}[p_{i2}\text{-}169, p_{i2}]$中至少部分数据是否满足预定条件$C_2$的方式和判断$W_{j2}[p_{j2}\text{-}169, p_{j2}]$中至少部分数据是否满足预定条件$C_2$的方式相同，因此，如图17所示，“▨”表示判断窗口$W_{j2}[p_{j2}\text{-}169, p_{j2}]$中至少部分数据是否满足预定条件$C_2$时选择的1个字节，相邻两个选择的字节之间相差42个字节。将选择的5字节数据反复利用51次，共获得255字节，以增加随机性。其中每个字节由8位组成，记为$b_{m,1}'...b_{m,8}'$,表示255个字节中第m个字节的第1到第8位，因此，255个字节对应的位可以表示为：$\begin{pmatrix} b_{1,1}' & b_{1,2}' & \cdots & b_{1,8}' \\ b_{2,1}' & b_{2,2}' & \cdots & b_{2,8}' \\ \vdots & \vdots & \vdots & \vdots \\ b_{255,1}' & b_{255,2}' & \cdots & b_{255,8}' \end{pmatrix}$，当$b_{m,n}'=1$时，$V_{bm,n}'=1$，当$b_{m,n}'=0$时，$V_{bm,n}'=-1$，其中$b_{m,n}'$表示$b_{m,1}'...b_{m,8}'$中的任一个，255个字节对应的位按照$b_{m,n}'$与$V_{bm,n}'$的转换关系得到矩阵$V_b'$，可以表示为：

$\begin{pmatrix} V_{b1,1} & V_{b1,2} & \cdots & V_{b1,8} \\ V_{b2,1} & V_{b2,2} & \cdots & V_{b2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{b255,1} & V_{b255,2} & \cdots & V_{b255,8} \end{pmatrix}$。窗口$W_2[p_2\text{-}169, p_2]$和$W_2[q_2\text{-}169, q_2]$中至少部分数据是否满足预定条件的方式相同，因此仍使用矩阵R：

$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,8} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{255,1} & h_{255,2} & \cdots & h_{255,8} \end{pmatrix}$,将矩阵$V_b$'的第m行与矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为S_{bm}'=$V_{bm,1}$'*$h_{m,1}$+$V_{bm,2}$'*$h_{m,2}$+...+$V_{bm,8}$'*$h_{m,8}$。根据该方法，获得S_{b1}'、S_{b2}'...到S_{b255}'，统计S_{b1}'、S_{b2}'...到S_{b255}'中满足特定条件（这里以大于0为例）的值的个数K。由于矩阵R服从正态分布，则S_{bm}'与矩阵R一样，仍然服从正态分布，根据概率论，正态分布随机数大于0的概率为1/2，在S_{b1}'、S_{b2}'...到S_{b255}'中，每个值大于0的概率为1/2，所以K满足二项分布：

$P(k=n)=C_{255}^{n}(\frac{1}{2})^{n}(\frac{1}{2})^{255-n}=C_{255}^{n}(\frac{1}{2})^{255}$。根据统计结果，判断$S_{b1}$'、$S_{b2}$'...到$S_{b255}$'的值大于0的个数K是否为偶数，二项分布的随机数为偶数的概率为为1/2，所以K以1/2的概率满足条件。当K为偶数时，表明W_{j2}[p_{j2}-169,p_{j2}]中至少部分数据满足预定条件C_2；当K为奇数时，表明W_{j2}[p_{j2}-169,p_{j2}]中至少部分数据不满足预定条件C_2。同理，判断W_{i3}[p_{i3}-169,p_{i3}]中至少部分数据是否满足预定条件C_3的方式与判断W_{j3}[p_{j3}-169,p_{j3}]中至少部分数据是否满足预定条件C_3的方式相同，同理，判断W_{j4}[p_{j4}-169,p_{j4}]中至少部分数据是否满足预定条件C_4、判断W_{j5}[p_{j5}-169,p_{j5}]中至少部分数据是否满足预定条件C_5、判断W_{j6}[p_{j6}-169, p_{j6}]中至少部分数据是否满足预定条件C_6、判断W_{j7}[p_{j7}-169, p_{j7}]中至少部分数据是否满足预定条件C_7、判断W_{j8}[p_{j8}-169, p_{j8}]中至少部分数据是否满足预定条件C_8、判断W_{j9}[p_{j9}-169, p_{j9}]中至少部分数据是否满足预定条件C_9、判断W_{j10}[p_{j10}-169, p_{j10}]中至少部分数据是否满足预定条件C_{10}和判断W_{j11}[p_{j11}-169, p_{j11}]中至少部分数据是否满足预定条件C_{11}，

在此不再赘述。

仍然以图5所示实施方式为例，提供了一种判断窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足预定条件$C_z$的方法，本实施例中使用随机函数判断窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足预定条件$C_z$，根据在去重服务器103上预设的规则，为潜在分割点k_i确定点p_{i1}及p_{i1}对应的窗口$W_{i1}[p_{i1}\text{-}169, p_{i1}]$，判断$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据是否满足预定的条件$C_1$，如图16所示，$W_{i1}$表示窗口$W_{i1}[p_{i1}\text{-}169, p_{i1}]$，为判断$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据是否满足预定条件$C_1$，选择5个字节，图16中“■”表示选择的1个字节，相邻两个选择“■”的字节之间相差42个字节。其中一种实现方式为使用HASH函数计算选择的5个字节，使用HASH函数计算得到的数值是一个固定均匀分布，如果使用HASH函数计算得到的数值为偶数，则判断$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据满足预定条件$C_1$，即$C_1$表示根据上述方式使用HASH函数计算得到的数值为偶数。因此，$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据是否满足预定条件的概率为1/2。在图5所示的实施方式中，使用Hash函数判断$W_{i2}[p_{i2}\text{-}169, p_{i2}]$中至少部分数据是否满足预定条件$C_2$、$W_{i3}[p_{i3}\text{-}169, p_{i3}]$中至少部分数据是否满足预定条件$C_3$、$W_{i4}[p_{i4}\text{-}169, p_{i4}]$中至少部分数据是否满足预定条件$C_4$和$W_{i5}[p_{i5}\text{-}169, p_{i5}]$中至少部分数据是否满足预定条件$C_5$，具体实现可参考描述图5所示实施方式使用Hash函数判断$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据是否满足预定条件的方式$C_1$，在此不再赘述。

当$W_{i5}[p_{i5}\text{-}169, p_{i5}]$中至少部分数据不满足预定条件$C_5$时，从点$p_{i5}$沿着数据流分割点查找方向跳跃11个字节，在第11个字节的结束位置获得当前潜在分割点k_j，如图6所示，根据在去重服务器103上预设的

规则，为潜在分割点k_j确定点p_{j1}、点p_{j1}对应的窗口$W_{j1}[p_{j1}\text{-}169, p_{j1}]$，判断窗口$W_{j1}[p_{j1}\text{-}169, p_{j1}]$中至少部分数据是否满足预定条件$C_1$的方式与判断窗口$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据是否满足预定条件$C_1$的方式相同，因此如图17所示，$W_{j1}$表示窗口$W_{j1}[p_{j1}\text{-}169, p_{j1}]$，为判断$W_{j1}[p_{j1}\text{-}169, p_{j1}]$中至少部分数据是否满足预定条件$C_1$，选择5个字节，图17中“■”表示选择的1个字节，相邻两个选择的字节“■”之间相差42个字节。使用Hash函数计算从窗口$W_{j1}[p_{j1}\text{-}169, p_{j1}]$中选取的5个字节，如果得到的数值为偶数，则$W_{j1}[p_{j1}\text{-}169, p_{j1}]$中至少部分数据满足预定条件$C_1$。图17中，判断$W_{i2}[p_{i2}\text{-}169, p_{i2}]$中至少部分数据是否满足预定条件$C_2$的方式和判断$W_{j2}[p_{j2}\text{-}169, p_{j2}]$中至少部分数据是否满足预定条件$C_2$的方式相同，因此，如图17所示，“”表示判断窗口$W_{j2}[p_{j2}\text{-}169, p_{j2}]$中至少部分数据是否满足预定条件$C_2$时选择的1个字节，相邻两个选择的字节“ ”之间相差42个字节。使用Hash函数计算选择的5个字节，如果得到的数值为偶数，则$W_{j2}[p_{j2}\text{-}169, p_{j2}]$中至少部分数据满足预定条件$C_2$。图17中，判断$W_{i3}[p_{i3}\text{-}169, p_{i3}]$中至少部分数据是否满足预定条件$C_3$的方式与判断$W_{j3}[p_{j3}\text{-}169, p_{j3}]$中至少部分数据是否满足预定条件$C_3$的方式相同，因此，如图17所示，“ ”表示判断窗口$W_{j3}[p_{j3}\text{-}169, p_{j3}]$中至少部分数据是否满足预定条件$C_3$时选择的1个字节，相邻两个选择的字节“ ”之间相差42个字节。使用Hash函数计算选择的5个字节，得到的数值为偶数，则$W_{j3}[p_{j3}\text{-}169, p_{j3}]$中至少部分数据满足预定条件$C_3$。图17中，判断$W_{j4}[p_{j4}\text{-}169, p_{j4}]$中至少部分数据是否满足预定条件$C_4$的方式和判断窗口$W_{i4}[p_{i4}\text{-}169, p_{i4}]$中至少部分数据是否满足预定条件$C_4$的方式，因此，如图17所示，“ ”表示判断窗口$W_{j4}$

$[p_{j4}\text{-}169, p_{j4}]$中至少部分数据是否满足预定条件$C_4$时选择的1个字节，相邻两个选择的字节“▨”之间相差42个字节。使用Hash函数计算选择的5个字节，得到的数值为偶数，则$W_{j4}[p_{j4}\text{-}169, p_{j4}]$中至少部分数据满足预定条件$C_4$。根据上述方法，判断$W_{j5}[p_{j5}\text{-}169, p_{j5}]$中至少部分数据是否满足预定条件$C_5$、判断$W_{j6}[p_{j6}\text{-}169, p_{j6}]$中至少部分数据是否满足预定条件$C_6$、判断$W_{j7}[p_{j7}\text{-}169, p_{j7}]$中至少部分数据是否满足预定条件$C_7$、判断$W_{j8}[p_{j8}\text{-}169, p_{j8}]$中至少部分数据是否满足预定条件$C_8$、判断$W_{j9}[p_{j9}\text{-}169, p_{j9}]$中至少部分数据是否满足预定条件$C_9$、判断$W_{j10}[p_{j10}\text{-}169, p_{j10}]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{j11}[p_{j11}\text{-}169, p_{j11}]$中至少部分数据是否满足预定条件$C_{11}$，在此不再赘述。

以图5所示的实施方式为例，提供了一种判断窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足预定条件$C_z$的方法，本实施例中使用随机函数判断窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足预定条件$C_z$，根据在去重服务器103上预设的规则，为潜在分割点k_i确定点p_{i1}及p_{i1}对应的窗口$W_{i1}[p_{i1}\text{-}169, p_{i1}]$，判断$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据是否满足预定条件$C_1$，如图16所示，$W_{i1}$表示窗口$W_{i1}[p_{i1}\text{-}169, p_{i1}]$，为判断$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据是否满足预定条件$C_1$，选择5个字节，图16中序号为169、127、85、43和1的字节“■”分别表示选择的1个字节，相邻两个选择的字节之间相差42个字节。将序号为169、127、85、43和1的字节“■”分别转换成一个十进制数值，分别表示为a_1、a_2、a_3、a_4和a_5。因为1个字节由8位组成，所以每个字节“■”作为一个数值，则a_1、a_2、a_3、a_4和a_5中的任一个a_r均满足$0\leq a_r\leq 255$。a_1、a_2、a_3、a_4和a_5组成1*5的矩阵。从服从二项分布的随机数中选

择256*5个随机数，组成矩阵R，表示为：$\begin{pmatrix} h_{0,1} & h_{0,2} & \cdots & h_{0,5} \\ h_{1,1} & h_{1,2} & \cdots & h_{1,5} \\ \vdots & \vdots & \vdots & \vdots \\ h_{255,5} & h_{255,5} & \cdots & h_{255,5} \end{pmatrix}$，

根据a_1的值和所在的列，从矩阵R中查找对应的值，如a_1=36，a_1位于第1列，则查找$h_{36,1}$对应的值；根据a_2的值和所在的列，从矩阵R中查找对应的值，如a_2=48，a_2位于第2列，则查找$h_{48,2}$对应的值；根据a_3的值和所在的列，从矩阵R中查找对应的值，如a_3=26，a_3位于第3列，则查找$h_{26,3}$对应的值；根据a_4的值和所在的列，从矩阵R中查找对应的值，如a_4=26，a_4位于第4列，则查找$h_{26,4}$对应的值；根据a_5的值和所在的列，从矩阵R中查找对应的值，如a_5=88，a_5位于第5列，则查找$h_{88,5}$对应的值。$S_1=h_{36,1}+h_{48,2}+h_{26,3}+h_{26,4}+h_{88,5}$，因为矩阵R服从二项分布，因此，$S_1$也服从二项分布。当$S_1$为偶数，则$W_{i1}[p_{i1}\text{-}169,p_{i1}]$中至少部分数据满足预定条件$C_1$，当$S_1$为奇数，则$W_{i1}[p_{i1}\text{-}169,p_{i1}]$中至少部分数据不满足预定条件$C_1$，$S_1$为偶数的概率为1/2，$C_1$表示按上述方式计算$S_1$为偶数。在图5所示实施例中，$W_{i1}[p_{i1}\text{-}169,p_{i1}]$中至少部分数据满足预定条件$C_1$。如图16所示，“▨”表示判断窗口$W_{i2}[p_{i2}\text{-}169,p_{i2}]$中至少部分数据是否满足预定条件$C_2$时分别选择的1个字节，在图16中，分别用序号170、128、86、44和2表示，相邻两个选择的字节之间相差42个字节。将序号170、128、86、44和2的字节“▨”分别转换成一个十进制数值，分别表示为b_1、b_2、b_3、b_4和b_5。因为1个字节由8位组成，所以每个字节“▨”作为一个数值，则b_1、b_2、b_3、b_4和b_5中的任一个b_r均满足$0\leq b_r\leq 255$。b_1、b_2、b_3、b_4和b_5组成1*5的矩阵。本实施方式中，判断W_{i1}和W_{i2}中至少部分数据是否满足预定条件的方式相同，因此仍然使用矩阵R，根据b_1的值和所在的列，从矩阵R中

查找对应的值，如b_1=66，b_1位于第1列，则查找$h_{66,1}$对应的值；根据b_2的值和所在的列，从矩阵R中查找对应的值，如b_2=48，b_2位于第2列，则查找$h_{48,2}$对应的值；根据b_3的值和所在的列，从矩阵R中查找对应的值，如b_3=99，b_3位于第3列，则查找$h_{99,3}$对应的值；根据b_4的值和所在的列，从矩阵R中查找对应的值，如b_4=26，b_4位于第4列，则查找$h_{26,4}$对应的值；根据b_5的值和所在的列，从矩阵R中查找对应的值，如b_5=90，b_5位于第5列，则查找$h_{90,5}$对应的值。$S_2=h_{66,1}+h_{48,2}+h_{99,3}+h_{26,4}+h_{90,5}$，因为矩阵R服从二项分布，因此，$S_2$也服从二项分布。当$S_2$为偶数，则$W_{i2}$[$p_{i2}$-169,$p_{i2}$]中至少部分数据满足预定条件$C_2$，当$S_2$为奇数，则$W_{i2}$[$p_{i2}$-169,$p_{i2}$]中至少部分数据不满足预定条件$C_2$，$S_2$为偶数的概率为1/2。在图5所示实施例中，$W_{i2}$[$p_{i2}$-169,$p_{i2}$]中至少部分数据满足预定条件$C_2$。使用同样的规则，分别判断$W_{i3}$[$p_{i3}$-169,$p_{i3}$]中至少部分数据是否满足预定条件$C_3$、判断$W_{i4}$[$p_{i4}$-169,$p_{i4}$]中至少部分数据是否满足预定条件$C_4$、判断$W_{i5}$[$p_{i5}$-169,$p_{i5}$]中至少部分数据是否满足预定条件$C_5$、判断$W_{i6}$[$p_{i6}$-169,$p_{i6}$]中至少部分数据是否满足预定条件$C_6$、判断$W_{i7}$[$p_{i7}$-169,$p_{i7}$]中至少部分数据是否满足预定条件$C_7$、判断$W_{i8}$[$p_{i8}$-169,$p_{i8}$]中至少部分数据是否满足预定条件$C_8$、判断$W_{i9}$[$p_{i9}$-169,$p_{i9}$]中至少部分数据是否满足预定条件$C_9$、判断$W_{i10}$[$p_{i10}$-169,$p_{i10}$]中至少部分数据是否满足预定条件$C_{10}$和判断$W_{i11}$[$p_{i11}$-169,$p_{i11}$]中至少部分数据是否满足预定条件$C_{11}$。图5所示的实施方式中，$W_{i5}$[$p_{i5}$-169,$p_{i5}$]中至少部分数据不满足预定条件$C_5$，从点$p_{i5}$沿着数据流分割点查找方向跳跃11个字节，在第11个字节的结束位置获得当前潜在分割点k_j，如图6所示，根据在去重服务器103上预设的规则，为潜在分割点k_j确定点p_{j1}、点p_{j1}对应的窗口W_{j1}[p_{j1}-169,p_{j1}]，判断窗口W_{j1}[p_{j1}-169,p_{j1}]中至少部分数据是

否满足预定条件C_1的方式与判断窗口$W_{i1}[p_{i1}\text{-}169,p_{i1}]$中至少部分数据是否满足预定条件$C_1$的方式相同，因此如图17所示，$W_{j1}$表示窗口$W_{j1}[p_{j1}\text{-}169,p_{j1}]$，为判断$W_{j1}[p_{j1}\text{-}169,p_{j1}]$中至少部分数据是否满足预定条件$C_1$，图17中序号为169、127、85、43和1的字节“■”分别表示选择的1个字节，相邻两个选择的字节之间相差42个字节。将序号为169、127、85、43和1的字节“■”分别转换成一个十进制数值，分别表示为a_1'、a_2'、a_3'、a_4'和a_5'。因为1个字节由8位组成，所以每个字节“■”作为一个数值，则a_1'、a_2'、a_3'、a_4'和a_5'中的任一个a_r'均满足$0\le a_r'\le 255$。a_1'、a_2'、a_3'、a_4'和a_5'组成1*5的矩阵。判断窗口$W_{j1}[p_{j1}\text{-}169,p_{j1}]$中至少部分数据是否满足预定条件$C_1$的方式与判断窗口$W_{i1}[p_{i1}\text{-}169,p_{i1}]$中至少部分数据是否满足预定条件$C_1$的方式相同，因此，仍然使用矩阵R，表示为：$\begin{pmatrix} h_{0,1} & h_{0,2} & \cdots & h_{0,5} \\ h_{1,1} & h_{1,2} & \cdots & h_{1,5} \\ \vdots & \vdots & \vdots & \vdots \\ h_{255,5} & h_{255,5} & \cdots & h_{255,5} \end{pmatrix}$，

根据a_1'的值和所在的列，从矩阵R中查找对应的值，如a_1'=16，a_1'位于第1列，则查找$h_{16,1}$对应的值；根据a_2'的值和所在的列，从矩阵R中查找对应的值，如a_2'=98，a_2'位于第2列，则查找$h_{98,2}$对应的值；根据a_3'的值和所在的列，从矩阵R中查找对应的值，如a_3'=56，a_3'位于第3列，则查找$h_{56,3}$对应的值；根据a_4'的值和所在的列，从矩阵R中查找对应的值，如a_4'=36，a_4'位于第4列，则查找$h_{36,4}$对应的值；根据a_5'的值和所在的列，从矩阵R中查找对应的值，如a_5'=99，a_5'位于第5列，则查找$h_{99,5}$对应的值。$S_1'=h_{16,1}+h_{98,2}+h_{56,3}+h_{36,4}+h_{99,5}$，因为矩阵R服从二项分布，因此，$S_1'$也服从二项分布。当$S_1'$为偶数，则$W_{j1}[p_{j1}\text{-}169,p_{j1}]$中至少部分数据满足预定条件$C_1$，当$S_1'$为奇数，则$W_{j1}[p_{j1}\text{-}169,p_{j1}]$中至

少部分数据不满足预定条件C_1，S_1'为偶数的概率为1/2。

判断$W_{i2}[p_{i2}\text{-}169,p_{i2}]$中至少部分数据是否满足预定条件$C_2$的方式和判断$W_{j2}[p_{j2}\text{-}169,p_{j2}]$中至少部分数据是否满足预定条件$C_2$的方式相同，因此，如图17所示，“ ”表示判断窗口$W_{j2}[p_{j2}\text{-}169,p_{j2}]$中至少部分数据是否满足预定条件$C_2$时选择的1个字节，相邻两个选择的字节之间相差42个字节，分别用序号170、128、86、44和2表示，相邻两个选择的字节之间相差42个字节。将序号170、128、86、44和2的字节“ ”分别转换成一个十进制数值，分别表示为b_1'、b_2'、b_3'、b_4'和b_5'。因为1个字节由8位组成，所以每个字节“ ”作为一个数值，则b_1'、b_2'、b_3'、b_4'和b_5'中的任一个b_r'均满足$0\leq b_r'\leq 255$。b_1'、b_2'、b_3'、b_4'和b_5'组成1*5的矩阵。与判断窗口$W_{i2}[p_{i2}\text{-}169,p_{i2}]$中至少部分数据是否满足预定条件$C_2$使用相同的矩阵R，根据$b_1'$的值和所在的列，从矩阵R中查找对应的值，如$b_1'$=210，$b_1'$位于第1列，则查找$h_{210,1}$对应的值；根据$b_2'$的值和所在的列，从矩阵R中查找对应的值，如$b_2'$=156，$b_2'$位于第2列，则查找$h_{156,2}$对应的值；根据$b_3'$的值和所在的列，从矩阵R中查找对应的值，如$b_3'$=144，$b_3'$位于第3列，则查找$h_{144,3}$对应的值；根据$b_4'$的值和所在的列，从矩阵R中查找对应的值，如$b_4'$=60，$b_4'$位于第4列，则查找$h_{60,4}$对应的值；根据$b_5'$的值和所在的列，从矩阵R中查找对应的值，如$b_5'$=90，$b_5'$位于第5列，则查找$h_{90,5}$对应的值。$S_2'=h_{210,1}+h_{156,2}+h_{144,3}+h_{60,4}+h_{90,5}$，与$S_2$的判断条件相同，当$S_2'$为偶数，则$W_{j2}[p_{j2}\text{-}169,p_{j2}]$中至少部分数据满足预定条件$C_2$，当$S_2'$为奇数，则$W_{j2}[p_{j2}\text{-}169,p_{j2}]$中至少部分数据不满足预定条件$C_2$，$S_2'$为偶数的概率为1/2。

同理，判断$W_{i3}[p_{i3}\text{-}169,p_{i3}]$中至少部分数据是否满足预定条件$C_3$

的方式与判断$W_{j3}[p_{j3}\text{-}169,p_{j3}]$中至少部分数据是否满足预定条件$C_3$的方式相同，同理，判断$W_{j4}[p_{j4}\text{-}169,p_{j4}]$中至少部分数据是否满足预定条件$C_4$、判断$W_{j5}[p_{j5}\text{-}169,p_{j5}]$中至少部分数据是否满足预定条件$C_5$、判断$W_{j6}[p_{j6}\text{-}169,\ p_{j6}]$中至少部分数据是否满足预定条件$C_6$、判断$W_{j7}[p_{j7}\text{-}169,\ p_{j7}]$中至少部分数据是否满足预定条件$C_7$、判断$W_{j8}[p_{j8}\text{-}169,p_{j8}]$中至少部分数据是否满足预定条件$C_8$、判断$W_{j9}[p_{j9}\text{-}169,\ p_{j9}]$中至少部分数据是否满足预定条件$C_9$、判断$W_{j10}[p_{j10}\text{-}169,\ p_{j10}]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{j11}[p_{j11}\text{-}169,\ p_{j11}]$中至少部分数据是否满足预定条件$C_{11}$，在此不再赘述。

以图5所示的实施方式为例，提供了一种判断窗口$W_{iz}[p_{iz}-A_z,\ p_{iz}+B_z]$中至少部分数据是否满足预定条件$C_z$的方法，本实施例中使用随机函数判断窗口$W_{iz}[p_{iz}-A_z,\ p_{iz}+B_z]$中至少部分数据是否满足预定条件$C_z$，根据在去重服务器103上预设的规则，为潜在分割点k_i确定点p_{i1}及p_{i1}对应的窗口$W_{i1}[p_{i1}\text{-}169,p_{i1}]$，判断$W_{i1}[p_{i1}\text{-}169,p_{i1}]$中至少部分数据是否满足预定的条件$C_1$，如图16所示，$W_{i1}$表示窗口$W_{i1}[p_{i1}\text{-}169,p_{i1}]$，为判断$W_{i1}[p_{i1}\text{-}169,p_{i1}]$中至少部分数据是否满足预定条件$C_1$，选择5个字节，图16中序号为169、127、85、43和1的字节“■”分别表示选择的1个字节，相邻两个选择的字节之间相差42个字节。将序号为169、127、85、43和1的字节“■”分别转换成一个十进制数值，分别表示为a_1、a_2、a_3、a_4和a_5。因为1个字节由8位组成，所以每个字节“■”作为一个数值，则a_1、a_2、a_3、a_4和a_5中的任一个a_s均满足$0\leq a_s\leq 255$。a_1、a_2、a_3、a_4和a_5组成1*5的矩阵。从服从二项分布的随机数中选

择256*5个随机数，组成矩阵R，表示为：$\begin{pmatrix} h_{0,1} & h_{0,2} & \cdots & h_{0,5} \\ h_{1,1} & h_{1,2} & \cdots & h_{1,5} \\ \vdots & \vdots & \vdots & \vdots \\ h_{255,5} & h_{255,5} & \cdots & h_{255,5} \end{pmatrix}$，从服从二项分布的随机数中选择256*5个随机数，组成矩阵G，表示为：$\begin{pmatrix} g_{0,1} & g_{0,2} & \cdots & g_{0,5} \\ g_{1,1} & g_{1,2} & \cdots & g_{1,5} \\ \vdots & \vdots & \vdots & \vdots \\ g_{255,5} & g_{255,5} & \cdots & g_{255,5} \end{pmatrix}$。

根据 a_1 的值和所在的列，如 a_1=36， a_1 位于第1列，则从矩阵R中查找查找 $h_{36,1}$ 对应的值，从矩阵G中查找 $g_{36,1}$ 对应的值；根据 a_2 的值和所在的列，如 a_2=48， a_2 位于第2列，则从矩阵R中查 $h_{48,2}$ 对应的值，从矩阵G中查找 $g_{48,2}$ 对应的值；根据 a_3 的值和所在的列，如 a_3=26， a_3 位于第3列，则从矩阵R中查找 $h_{26,3}$ 对应的值，从矩阵G中查找 $g_{26,3}$ 对应的值；根据 a_4 的值和所在的列，如 a_4=26， a_4 位于第4列，则从矩阵R中查找 $h_{26,4}$ 对应的值，从矩阵G中查找 $g_{26,4}$ 对应的值；根据 a_5 的值和所在的列，如 a_5=88， a_5 位于第5列，则从矩阵R中查找 $h_{88,5}$ 对应的值，从矩阵G中查找 $g_{88,5}$ 对应的值。$S_{1h}=h_{36,1}+h_{48,2}+h_{26,3}+h_{26,4}+h_{88,5}$，因为矩阵R服从二项分布，因此， S_{1h} 也服从二项分布；$S_{1g}=g_{36,1}+g_{48,2}+g_{26,3}+g_{26,4}+g_{88,5}$，因为矩阵G服从二项分布，因此 S_{1g} 也服从二项分布。当 S_{1h} 和 S_{1g} 中有1个为偶数，则 $W_{i1}[p_{i1}\text{-}169, p_{i1}]$ 中至少部分数据满足预定条件 C_1，当 S_{1h} 和 S_{1g} 均为奇数，则 $W_{i1}[p_{i1}\text{-}169, p_{i1}]$ 中至少部分数据不满足预定条件 C_1，C_1 表述按照上述方法获得的 S_{1h} 和 S_{1g} 中有1个为偶数。因为 S_{1h} 和 S_{1g} 均服从二项分布，因此 S_{1h} 为偶数的概率为1/2， S_{1g} 为偶数的概率为1/2， S_{1h} 和 S_{1g} 中有1个为偶数的概率为1-1/4=3/4，因此， $W_{i1}[p_{i1}\text{-}169, p_{i1}]$ 中至少部分数据满足预定条件 C_1 的概率为3/4。在图5所示实施例中，

$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据满足预定条件$C_1$。在图5所示的实施方式中，在$W_{i1}[p_{i1}\text{-}169, p_{i1}]$ 、$W_{i2}[p_{i2}\text{-}169, p_{i2}]$、$W_{i3}[p_{i3}\text{-}169, p_{i3}]$、$W_{i4}[p_{i4}\text{-}169, p_{i4}]$、$W_{i5}[p_{i5}\text{-}169, p_{i5}]$、$W_{i6}[p_{i6}\text{-}169, p_{i6}]$、$W_{i7}[p_{i7}\text{-}169, p_{i7}]$、$W_{i8}[p_{i8}\text{-}169, p_{i8}]$、$W_{i9}[p_{i9}\text{-}169, p_{i9}]$、$W_{i10}[p_{i10}\text{-}169, p_{i10}]$和$W_{i11}[p_{i11}\text{-}169, p_{i11}]$中，各窗口大小相同,即窗口大小均为169字节，同时判断窗口中至少部分数据是否满足预定条件的方式也相同,具体见上述判断$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据是否满足预定条件$C_1$的描述。因此，如图16所示，“ ”表示判断窗口$W_{i2}[p_{i2}\text{-}169, p_{i2}]$中至少部分数据是否满足预定条件$C_2$时分别选择的1个字节，在图16中，分别用序号170、128、86、44和2表示，相邻两个选择的字节之间相差42个字节。将序号170、128、86、44和2的字节“ ” 分别转换成一个十进制数值，分别表示为b_1、b_2、b_3、b_4和b_5。因为1个字节由8位组成，所以每个字节“ ”作为一个数值，则b_1、b_2、b_3、b_4和b_5中的任一个b_s均满足$0 \le b_s \le 255$。b_1、b_2、b_3、b_4和b_5组成1*5的矩阵。本实施方式中，判断各窗口中至少部分数据是否满足预定条件的方式相同，因此仍然使用相同矩阵R和G。根据b_1的值和所在的列，如b_1=66，b_1位于第1列，则从矩阵R中查找$h_{66,1}$对应的值，从矩阵G中查找$g_{66,1}$对应的值；根据b_2的值和所在的列，如b_2=48，b_2位于第2列，则从矩阵R中查找$h_{48,2}$对应的值，从矩阵G中查找$g_{48,2}$对应的值；根据b_3的值和所在的列，如b_3=99，b_3位于第3列，则从矩阵R中查找$h_{99,3}$对应的值，从矩阵G中查找$g_{99,3}$对应的值；根据b_4的值和所在的列，如b_4=26，b_4位于第4列，则从矩阵R中查找$h_{26,4}$对应的值，从矩阵G中查找$g_{26,4}$对应的值；根据b_5的值和所在的列，如b_5=90，b_5位于第5列，则从矩阵R中查找$h_{90,5}$对应的值，从矩阵G中查找$g_{90,5}$对应

的值。$S_{2h}=h_{66,1}+h_{48,2}+h_{99,3}+h_{26,4}+h_{90,5}$,因为矩阵R服从二项分布,因此,$S_{2h}$也服从二项分布。$S_{2g}=g_{66,1}+g_{48,2}+g_{99,3}+g_{26,4}+g_{90,5}$,因为矩阵G服从二项分布,因此,$S_{2g}$也服从二项分布。当$S_{2h}$和$S_{2g}$中有1个为偶数,则$W_{i2}[p_{i2}\text{-}169,p_{i2}]$中至少部分数据满足预定条件$C_2$,当$S_{2h}$和$S_{2g}$均为奇数,则$W_{i2}[p_{i2}\text{-}169,p_{i2}]$中至少部分数据不满足预定条件$C_2$,$S_{2h}$和$S_{2g}$中有1个为偶数的概率为3/4。在图5所示实施例中,$W_{i2}[p_{i2}\text{-}169,p_{i2}]$中至少部分数据满足预定条件$C_2$。使用同样的规则,分别判断$W_{i3}[p_{i3}\text{-}169,p_{i3}]$中至少部分数据是否满足预定条件$C_3$、判断$W_{i4}[p_{i4}\text{-}169,p_{i4}]$中至少部分数据是否满足预定条件$C_4$、判断$W_{i5}[p_{i5}\text{-}169,p_{i5}]$中至少部分数据是否满足预定条件$C_5$、判断$W_{i6}[p_{i6}\text{-}169,p_{i6}]$中至少部分数据是否满足预定条件$C_6$、判断$W_{i7}[p_{i7}\text{-}169,p_{i7}]$中至少部分数据是否满足预定条件$C_7$、判断$W_{i8}[p_{i8}\text{-}169,p_{i8}]$中至少部分数据是否满足预定条件$C_8$、判断$W_{i9}[p_{i9}\text{-}169,p_{i9}]$中至少部分数据是否满足预定条件$C_9$、判断$W_{i10}[p_{i10}\text{-}169,p_{i10}]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{i11}[p_{i11}\text{-}169,p_{i11}]$中至少部分数据是否满足预定条件$C_{11}$。图5所示的实施方式中,$W_{i5}[p_{i5}\text{-}169,p_{i5}]$中至少部分数据不满足预定条件$C_5$,从点$p_{i5}$沿着数据流分割点查找方向跳跃11个字节,在第11个字节的结束位置获得当前潜在分割点k_j,如图6所示,根据在去重服务器103上预设的规则,为潜在分割点k_j确定点p_{j1}、点p_{j1}对应的窗口$W_{j1}[p_{j1}\text{-}169,p_{j1}]$,判断窗口$W_{j1}[p_{j1}\text{-}169,p_{j1}]$中至少部分数据是否满足预定条件$C_1$的方式与判断窗口$W_{i1}[p_{i1}\text{-}169,p_{i1}]$中至少部分数据是否满足预定条件$C_1$的方式相同,因此如图17所示,$W_{j1}$表示窗口$W_{j1}[p_{j1}\text{-}169,p_{j1}]$,为判断$W_{j1}[p_{j1}\text{-}169,p_{j1}]$中至少部分数据是否满足预定条件$C_1$,图17中序号为169、127、85、43和1的字节“■”分别表示选择的1个字节,相邻两个选择

的字节之间相差42个字节。将序号为169、127、85、43和1的字节“■”分别转换成一个十进制数值，分别表示为 a_1'、a_2'、a_3'、a_4'和a_5'。因为1个字节由8位组成，所以每个字节“■”作为一个数值，则 a_1'、a_2'、a_3'、a_4'和a_5'中的任一个a_s'均满足$0\leq a_s'\leq 255$。a_1'、a_2'、a_3'、a_4'和a_5'组成1*5的矩阵。使用与判断窗口$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据是否满足预定条件$C_1$相同的矩阵R和G，分别表示为：

$$\begin{pmatrix} h_{0,1} & h_{0,2} & \cdots & h_{0,5} \\ h_{1,1} & h_{1,2} & \cdots & h_{1,5} \\ \vdots & \vdots & \vdots & \vdots \\ h_{255,5} & h_{255,5} & \cdots & h_{255,5} \end{pmatrix} \text{和} \begin{pmatrix} g_{0,1} & g_{0,2} & \cdots & g_{0,5} \\ g_{1,1} & g_{1,2} & \cdots & g_{1,5} \\ \vdots & \vdots & \vdots & \vdots \\ g_{255,5} & g_{255,5} & \cdots & g_{255,5} \end{pmatrix}。$$

根据a_1'的值和所在的列，如a_1'=16，a_1'位于第1列，则从矩阵R中查找$h_{16,1}$对应的值，从矩阵G中查找$g_{16,1}$对应的值；根据a_2'的值和所在的列，如a_2'=98，a_2'位于第2列，则从矩阵R中查找$h_{98,2}$对应的值，从矩阵G中查找$g_{98,2}$对应的值；根据a_3'的值和所在的列，如a_3'=56，a_3'位于第3列，则从矩阵R中查找$h_{56,3}$对应的值，从矩阵G中查找$g_{56,3}$对应的值；根据a_4'的值和所在的列，如a_4'=36，a_4'位于第4列，则从矩阵R中查找$h_{36,4}$对应的值，从矩阵G中查找$g_{36,4}$对应的值；根据a_5'的值和所在的列，如a_5'=99，a_5'位于第5列，则从矩阵R中查找$h_{99,5}$对应的值，从矩阵G中查找$g_{99,5}$对应的值。$S_{1h}'=h_{16,1}+h_{98,2}+h_{56,3}+h_{36,4}+h_{99,5}$，因为矩阵R服从二项分布，因此，$S_{1h}'$也服从二项分布；$S_{1g}'=g_{16,1}+g_{98,2}+g_{56,3}+g_{36,4}+g_{99,5}$，因为矩阵G服从二项分布，因此$S_{1g}'$也服从二项分布。当$S_{1h}'$和$S_{1g}'$中有1个为偶数，则$W_{j1}[p_{j1}\text{-}169, p_{j1}]$中至少部分数据满足预定条件$C_1$，当$S_{1h}'$和$S_{1g}'$均为奇数，则$W_{j1}[p_{j1}\text{-}169, p_{j1}]$中至少部分数据不满足预定条件$C_1$，$S_{1h}'$和$S_{1g}'$有1个为偶数的概率为3/4。

判断$W_{i2}[p_{i2}\text{-}169, p_{i2}]$中至少部分数据是否满足预定条件$C_2$的方式

和判断$W_{j2}[p_{j2}\text{-}169,p_{j2}]$中至少部分数据是否满足预定条件$C_2$的方式相同，因此，如图17所示，“ ”表示判断窗口$W_{j2}[p_{j2}\text{-}169,p_{j2}]$中至少部分数据是否满足预定条件$C_2$时选择的1个字节，相邻两个选择的字节之间相差42个字节。在图17中，分别用序号170、128、86、44和2表示，相邻两个选择的字节之间相差42个字节。将序号170、128、86、44和2的字节“ ”分别转换成一个十进制数值，分别表示为b_1'、b_2'、b_3'、b_4'和b_5'。因为1个字节由8位组成，所以每个字节“ ”作为一个数值，则b_1'、b_2'、b_3'、b_4'和b_5'中的任一个b_s'均满足$0\leq b_s'\leq 255$。b_1'、b_2'、b_3'、b_4'和b_5'组成1*5的矩阵。使用与判断窗口$W_{i2}[p_{i2}\text{-}169,p_{i2}]$中至少部分数据是否满足预定条件$C_2$相同的矩阵R和G，根据$b_1'$的值和所在的列，如$b_1'$=210，$b_1'$位于第1列，则从矩阵R中查找$h_{210,1}$对应的值，从矩阵G中查找$g_{210,1}$对应的值；根据$b_2'$的值和所在的列，如$b_2'$=156，$b_2'$位于第2列，则从矩阵R中查找$h_{156,2}$对应的值，从矩阵G中查找$g_{156,2}$对应的值；根据$b_3'$的值和所在的列，如$b_3'$=144，$b_3'$位于第3列，则从矩阵R中查找$h_{144,3}$对应的值，从矩阵G中查找$g_{144,3}$对应的值；根据$b_4'$的值和所在的列，如$b_4'$=60，$b_4'$位于第4列，则从矩阵R中查找$h_{60,4}$对应的值，从矩阵G中查找$g_{60,4}$对应的值；根据$b_5'$的值和所在的列，如$b_5'$=90，$b_5'$位于第5列，则从矩阵R中查找$h_{90,5}$对应的值，从矩阵G中查找$g_{90,5}$对应的值。$S_{2h}'=h_{210,1}+h_{156,2}+h_{144,3}+h_{60,4}+h_{90,5}$，$S_{2g}'=g_{210,1}+g_{156,2}+g_{144,3}+g_{60,4}+g_{90,5}$。当$S_{2h}'$和$S_{2g}'$中有1个为偶数，则$W_{j2}[p_{j2}\text{-}169,p_{j2}]$中至少部分数据满足预定条件$C_2$，当$S_{2h}'$和$S_{2g}'$均为奇数，则$W_{j2}[p_{j2}\text{-}169,p_{j2}]$中至少部分数据不满足预定条件$C_2$，$S_{2h}'$和$S_{2g}'$中有1个为偶数的概率为3/4。

同理，判断$W_{i3}[p_{i3}\text{-}169,p_{i3}]$中至少部分数据是否满足预定条件$C_3$的方式与判断$W_{j3}[p_{j3}\text{-}169,p_{j3}]$中至少部分数据是否满足预定条件$C_3$的方式相同，同理，判断$W_{j4}[p_{j4}\text{-}169,p_{j4}]$中至少部分数据是否满足预定条件$C_4$、判断$W_{j5}[p_{j5}\text{-}169,p_{j5}]$中至少部分数据是否满足预定条件$C_5$、判断$W_{j6}[p_{j6}\text{-}169,\ p_{j6}]$中至少部分数据是否满足预定条件$C_6$、判断$W_{j7}[p_{j7}\text{-}169,\ p_{j7}]$中至少部分数据是否满足预定条件$C_7$、判断$W_{j8}[p_{j8}\text{-}169,p_{j8}]$中至少部分数据是否满足预定条件$C_8$、判断$W_{j9}[p_{j9}\text{-}169,\ p_{j9}]$中至少部分数据是否满足预定条件$C_9$、判断$W_{j10}[p_{j10}\text{-}169,\ p_{j10}]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{j11}[p_{j11}\text{-}169,\ p_{j11}]$中至少部分数据是否满足预定条件$C_{11}$，在此不再赘述。

以图5所示的实施方式为例，提供了一种判断窗口$W_{iz}[p_{iz}-A_z,\ p_{iz}+B_z]$中至少部分数据是否满足预定条件$C_z$的方法，本实施例中使用随机函数判断窗口$W_{iz}[p_{iz}-A_z,\ p_{iz}+B_z]$中至少部分数据是否满足预定条件$C_z$，根据在去重服务器103上预设的规则，为潜在分割点k_i确定点p_{i1}及p_{i1}对应的窗口$W_{i1}[p_{i1}\text{-}169,p_{i1}]$，判断$W_{i1}[p_{i1}\text{-}169,p_{i1}]$中至少部分数据是否满足预定的条件$C_1$，如图16所示，$W_{i1}$表示窗口$W_{i1}[p_{i1}\text{-}169,p_{i1}]$，为判断$W_{i1}[p_{i1}\text{-}169,p_{i1}]$中至少部分数据是否满足预定条件$C_1$，选择5个字节，图16中序号为169、127、85、43和1的字节“■”分别表示选择的1个字节，相邻两个选择的字节之间相差42个字节。将序号为169、127、85、43和1的字节“■”依次看成40个位，分别表示为a_1、a_2、a_3、$a_4\ldots a_{40}$。a_1、a_2、a_3、$a_4\ldots a_{40}$中的任一a_t，当a_t=0时，V_{at}=-1，当a_t=1时，V_{at}=1，根据a_t与V_{at}对应关系，生成V_{a1}、V_{a2}、V_{a3}、$V_{a4}\ldots V_{a40}$。从服从正态分布的随机数中选择40个随机数，分别表示为：h_1、h_2、h_3、$h_4\ldots h_{40}$。$S_a=V_{a1}*h_1+V_{a2}*h_2+V_{a3}*h_3+V_{a4}*h_4+\ldots+V_{a40}*h_{40}$。因为$h_1$、

h_2、h_3、$h_4 \ldots h_{40}$服从正态分布，因此，S_a也服从正态分布。当S_a为正数，则$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据满足预定条件$C_1$，当$S_a$为负数或0，则$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据不满足预定条件$C_1$，$S_a$为正数的概率为1/2。在图5所示实施例中，$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据满足预定条件$C_1$。如图16所示，“ ”表示判断窗口$W_{i2}[p_{i2}\text{-}169, p_{i2}]$中至少部分数据是否满足预定条件$C_2$时分别选择的1个字节，在图16中，分别用序号170、128、86、44和2表示，相邻两个选择的字节之间相差42个字节。将序号170、128、86、44和2的字节“ ”依次看成40个位，分别表示为b_1、b_2、b_3、$b_4 \ldots b_{40}$。b_1、b_2、b_3、$b_4 \ldots b_{40}$中的任一b_t，当b_t=0时，V_{bt}=-1，当b_t=1时，V_{bt}=1，根据b_t与V_{bt}对应关系，生成V_{b1}、V_{b2}、V_{b3}、$V_{b4} \ldots V_{b40}$。判断窗口$W_{i1}[p_{i1}\text{-}169, p_{i1}]$中至少部分数据是否满足预定条件$C_1$的方式与判断窗口$W_{i2}[p_{i2}\text{-}169, p_{i2}]$中至少部分数据是否满足预定条件$C_2$的方式相同，因此，使用相同的随机数：$h_1$、$h_2$、$h_3$、$h_4 \ldots h_{40}$，$S_b = V_{b1} * h_1 + V_{b2} * h_2 + V_{b3} * h_3 + V_{b4} * h_4 + \ldots + V_{b40} * h_{40}$。因为$h_1$、$h_2$、$h_3$、$h_4 \ldots h_{40}$服从正态分布，因此，$S_b$也服从正态分布。当$S_b$为正数，则$W_{i2}[p_{i2}\text{-}169, p_{i2}]$中至少部分数据满足预定条件$C_2$，当$S_b$为负数或0，则$W_{i2}[p_{i2}\text{-}169, p_{i2}]$中至少部分数据不满足预定条件$C_2$，$S_b$为正数的概率为1/2。在图5所示实施例中，$W_{i2}[p_{i2}\text{-}169, p_{i2}]$中至少部分数据满足预定条件$C_2$。使用同样的规则，分别判断$W_{i3}[p_{i3}\text{-}169, p_{i3}]$中至少部分数据是否满足预定条件$C_3$、判断$W_{i4}[p_{i4}\text{-}169, p_{i4}]$中至少部分数据是否满足预定条件$C_4$、判断$W_{i5}[p_{i5}\text{-}169, p_{i5}]$中至少部分数据是否满足预定条件$C_5$、判断$W_{i6}[p_{i6}\text{-}169, p_{i6}]$中至少部分数据是否满足预定条件$C_6$、判断$W_{i7}[p_{i7}\text{-}169, p_{i7}]$中至少部分数据是否满足预定条件$C_7$、判断$W_{i8}[p_{i8}\text{-}169, p_{i8}]$中至少部分数据是否满足预定条件$C_8$、判断$W_{i9}[p_{i9}$

-169, p_{i9}]中至少部分数据是否满足预定条件C_9、判断W_{i10}[p_{i10}-169, p_{i10}]中至少部分数据是否满足预定条件C_{10}和判断W_{i11}[p_{i11}-169, p_{i11}]中至少部分数据是否满足预定条件C_{11}。图5所示的实施方式中，W_{i5}[p_{i5}-169, p_{i5}]中至少部分数据不满足预定条件C_5，从点p_{i5}沿着数据流分割点查找方向跳跃11个字节，在第11个字节的结束位置获得当前潜在分割点k_j，如图6所示，根据在去重服务器103上预设的规则，为潜在分割点k_j确定点p_{j1}、点p_{j1}对应的窗口W_{j1}[p_{j1}-169, p_{j1}]，判断窗口W_{j1}[p_{j1}-169, p_{j1}]中至少部分数据是否满足预定条件C_1的方式与判断窗口W_{i1}[p_{i1}-169, p_{i1}]中至少部分数据是否满足预定条件C_1的方式相同，因此如图17所示，W_{j1}表示窗口W_{j1}[p_{j1}-169, p_{j1}]，为判断W_{j1}[p_{j1}-169, p_{j1}]中至少部分数据是否满足预定条件C_1，为判断W_{j1}[p_{j1}-169, p_{j1}]中至少部分数据是否满足预定条件C_1，选择5个字节，图17中序号为169、127、85、43和1的字节“■”分别表示选择的1个字节，相邻两个选择的字节之间相差42个字节。将序号为169、127、85、43和1的字节“■”依次看成40个位，分别表示为a_1'、a_2'、a_3'、a_4'...a_{40}'。a_1'、a_2'、a_3'、a_4'...a_{40}'中的任一a_t'，当a_t'=0时，V_{at}' =-1，当a_t'=1时，V_{at}'=1，根据a_t'与V_{at}'对应关系，生成V_{a1}'、V_{a2}'、V_{a3}'、V_{a4}'...V_{a40}'。判断窗口W_{j1}[p_{j1}-169, p_{j1}]中至少部分数据是否满足预定条件C_1的方式与判断窗口W_{i1}[p_{i1}-169, p_{i1}]中至少部分数据是否满足预定条件C_1的方式相同，因此使用相同的随机数：h_1、h_2、h_3、h_4...h_{40}。S_a'=V_{a1}'*h_1+V_{a2}'*h_2+V_{a3}'*h_3+V_{a4}'*h_4+...+V_{a40}'*h_{40}。因为h_1、h_2、h_3、h_4...h_{40}服从正态分布，因此，S_a'也服从正态分布。当S_a'为正数，则W_{j1}[p_{j1}-169, p_{j1}]中至少部分数据满足预定条件C_1，当S_a'为负数或0，则W_{j1}[p_{j1}-169, p_{j1}]中至少部分数据不满足预定条件C_1，S_a'为正数的概率为1/2。

判断$W_{i2}[p_{i2}\text{-}169,p_{i2}]$中至少部分数据是否满足预定条件$C_2$的方式和判断$W_{j2}[p_{j2}\text{-}169,p_{j2}]$中至少部分数据是否满足预定条件$C_2$的方式相同，因此，如图17所示，“ ”表示判断窗口$W_{j2}[p_{j2}\text{-}169,p_{j2}]$中至少部分数据是否满足预定条件$C_2$时选择的1个字节，相邻两个选择的字节之间相差42个字节。在图17中，分别用序号170、128、86、44和2表示，相邻两个选择的字节之间相差42个字节。将序号170、128、86、44和2的字节“ ”依次看成40个位，分别表示为b_1'、b_2'、b_3'、$b_4'\ldots b_{40}'$。b_1'、b_2'、b_3'、$b_4'\ldots b_{40}'$中的任一b_t'，当b_t'=0时，V_{bt}'=-1，当b_t'=1时，V_{bt}'=1，根据b_t'与V_{bt}'对应关系，生成V_{b1}'、V_{b2}'、V_{b3}'、$V_{b4}'\ldots V_{b40}'$。判断$W_{i2}[p_{i2}\text{-}169,p_{i2}]$中至少部分数据是否满足预定条件$C_2$的方式和判断$W_{j2}[p_{j2}\text{-}169,p_{j2}]$中至少部分数据是否满足预定条件$C_2$的方式相同，因此，使用相同的随机数：$h_1$、$h_2$、$h_3$、$h_4\ldots h_{40}$，$S_b'=V_{b1}'*h_1+V_{b2}'*h_2+V_{b3}'*h_3+V_{b4}'*h_4+\ldots+V_{b40}'*h_{40}$。因为$h_1$、$h_2$、$h_3$、$h_4\ldots h_{40}$服从正态分布，因此，$S_b'$也服从正态分布。当$S_b'$为正数，则$W_{j2}[p_{j2}\text{-}169,p_{j2}]$中至少部分数据满足预定条件$C_2$，当$S_b'$为负数或0，则$W_{j2}[p_{j2}\text{-}169,p_{j2}]$中至少部分数据不满足预定条件$C_2$，$S_b'$为正数的概率为1/2。

同理，判断$W_{i3}[p_{i3}\text{-}169,p_{i3}]$中至少部分数据是否满足预定条件$C_3$的方式与判断$W_{j3}[p_{j3}\text{-}169,p_{j3}]$中至少部分数据是否满足预定条件$C_3$的方式相同，同理，判断$W_{j4}[p_{j4}\text{-}169,p_{j4}]$中至少部分数据是否满足预定条件$C_4$、判断$W_{j5}[p_{j5}\text{-}169,p_{j5}]$中至少部分数据是否满足预定条件$C_5$、判断$W_{j6}[p_{j6}\text{-}169,\ p_{j6}]$中至少部分数据是否满足预定条件$C_6$、判断$W_{j7}[p_{j7}\text{-}169,\ p_{j7}]$中至少部分数据是否满足预定条件$C_7$、判断$W_{j8}[p_{j8}\text{-}169,p_{j8}]$中至少部分数据是否满足预定条件$C_8$、判断$W_{j9}[p_{j9}\text{-}169,\ p_{j9}]$中至

少部分数据是否满足预定条件C_9、判断$W_{j10}[p_{j10}\text{-}169,\ p_{j10}]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{j11}[p_{j11}\text{-}169,\ p_{j11}]$中至少部分数据是否满足预定条件$C_{11}$，在此不再赘述。

仍然以图5所示实施方式为例，提供了一种判断窗口$W_{iz}[p_{iz}-A_z,\ p_{iz}+B_z]$中至少部分数据是否满足预定条件$C_z$的方法，本实施例中使用随机函数判断窗口$W_{iz}[p_{iz}-A_z,\ p_{iz}+B_z]$中至少部分数据是否满足预定条件$C_z$，根据在去重服务器103上预设的规则，为潜在分割点k_i确定点p_{i1}及p_{i1}对应的窗口$W_{i1}[p_{i1}\text{-}169,p_{i1}]$，判断$W_{i1}[p_{i1}\text{-}169,p_{i1}]$中至少部分数据是否满足预定条件$C_1$，如图16所示，$W_{i1}$表示窗口$W_{i1}[p_{i1}\text{-}169,p_{i1}]$，为判断$W_{i1}[p_{i1}\text{-}169,p_{i1}]$中至少部分数据是否满足预定条件$C_1$，选择5个字节，图16中序号为169、127、85、43和1的字节“■”分别表示选择的1个字节，相邻两个选择的字节之间相差42个字节。将序号为169、127、85、43和1的字节“■”转换成1个十进制数，范围为0-（2^40-1），使用均匀分布随机数生成器为0-（2^40-1）中的每一个十进制数生成1个指定值，记录0-（2^40-1）中的每一个十进制数与指定值之间的对应关系R，一旦指定则该十进制数对应的指定值就不变，该指定值服从均匀分布，如果该指定值为偶数，则$W_{i1}[p_{i1}\text{-}169,p_{i1}]$中至少部分数据满足预定条件$C_1$，如果该指定值为奇数，则$W_{i1}[p_{i1}\text{-}169,p_{i1}]$中至少部分数据不满足预定条件$C_1$，$C_1$表示按照上述方法获得的指定值为偶数。因为均匀分布的随机数为偶数的概率为1/2，因此，$[p_{i1}\text{-}169,p_{i1}]$中至少部分数据满足预定条件$C_1$的概率为1/2。在图5所示的实施方式中，使用同样的规则，分别判断$W_{i2}[p_{i2}\text{-}169,p_{i2}]$中至少部分数据是否满足预定条件$C_2$，判断$W_{i3}[p_{i3}\text{-}169,p_{i3}]$中至少部分数据是否满足预定条件$C_3$、判断$W_{i4}[p_{i4}\text{-}169,p_{i4}]$中至少部分数据是否满足预定条件$C_4$、

判断$W_{i5}[p_{i5}\text{-}169,p_{i5}]$中至少部分数据是否满足预定条件$C_5$，在此不再赘述。

当$W_{i5}[p_{i5}\text{-}169,p_{i5}]$中至少部分数据不满足预定条件$C_5$，从点$p_{i5}$沿着数据流分割点查找方向跳跃11个字节，在第11个字节的结束位置获得当前潜在分割点k_j，如图6所示，根据在去重服务器103上预设的规则，为潜在分割点k_j确定点p_{j1}、点p_{j1}对应的窗口$W_{j1}[p_{j1}\text{-}169,p_{j1}]$，判断窗口$W_{j1}[p_{j1}\text{-}169,p_{j1}]$中至少部分数据是否满足预定条件$C_1$的方式与判断窗口$W_{i1}[p_{i1}\text{-}169,p_{i1}]$中至少部分数据是否满足预定条件$C_1$的方式相同，因此，使用相同的0-（2^40-1）中的每一个十进制数与指定值之间的对应关系R，如图17所示，W_{j1}表示窗口$W_{j1}[p_{j1}\text{-}169,p_{j1}]$，为判断$W_{j1}[p_{j1}\text{-}169,p_{j1}]$中至少部分数据是否满足预定条件$C_1$，选择5个字节，图17中“■”表示选择的1个字节，相邻两个选择的字节“■”之间相差42个字节。将序号为169、127、85、43和1的字节“■”转换成1个十进制数，在R查找该十进制数对应的指定值，如果该指定值为偶数，则$W_{j1}[p_{j1}\text{-}169,p_{j1}]$中至少部分数据满足预定条件$C_1$，如果该指定值为奇数，则$W_{j1}[p_{j1}\text{-}169,p_{j1}]$中至少部分数据不满足预定条件$C_1$，因为均匀分布的随机数为偶数的概率为1/2，因此，$W_{j1}[p_{j1}\text{-}169,p_{j1}]$中至少部分数据满足预定条件$C_1$的概率为1/2。同理，判断$W_{i2}[p_{i2}\text{-}169,p_{i2}]$中至少部分数据是否满足预定条件$C_2$的方式和判断$W_{j2}[p_{j2}\text{-}169,p_{j2}]$中至少部分数据是否满足预定条件$C_2$的方式相同，判断$W_{i3}[p_{i3}\text{-}169,p_{i3}]$中至少部分数据是否满足预定条件$C_3$的方式与判断$W_{j3}[p_{j3}\text{-}169,p_{j3}]$中至少部分数据是否满足预定条件$C_3$的方式相同，同理，判断$W_{j4}[p_{j4}\text{-}169,p_{j4}]$中至少部分数据是否满足预定条件$C_4$、判断$W_{j5}[p_{j5}\text{-}169,p_{j5}]$中至少部分数据是否满足预定条件$C_5$、判断$W_{j6}[p_{j6}\text{-}169,\ p_{j6}]$

中至少部分数据是否满足预定条件C_6、判断$W_{j7}[p_{j7}\text{-}169,\ p_{j7}]$中至少部分数据是否满足预定条件$C_7$、判断$W_{j8}[p_{j8}\text{-}169,\ p_{j8}]$中至少部分数据是否满足预定条件$C_8$、判断$W_{j9}[p_{j9}\text{-}169,\ p_{j9}]$中至少部分数据是否满足预定条件$C_9$、判断$W_{j10}[p_{j10}\text{-}169,\ p_{j10}]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{j11}[p_{j11}\text{-}169,\ p_{j11}]$中至少部分数据是否满足预定条件$C_{11}$，在此不再赘述。

图1所示的本发明实施例中的去重服务器103，是指能够实现本发明实施例所描述的技术方案的装置，如图18所示，通常包括中央处理单元、主存储器以及输入输出接口。中央处理单元、主存储器与输入输出接口之间相互通信，主存储器存储可执行指令，中央处理单元执行主存储器中存储的可执行指令，从而执行特定的功能，如本发明实施例图4至图17所描述的查找数据流分割点。因此，如图19所示，根据图4至图17所示的本发明实施例，去重服务器103，在去重服务器103上预设有规则，所述规则为：为潜在分割点k确定M个点p_x、点p_x对应的窗口$W_x[p_x\text{-}A_x, p_x+B_x]$和窗口$W_x[p_x\text{-}A_x, p_x+B_x]$对应的预定条件$C_x$，其中，x为1到M连续的自然数，M≥2，$A_x$和$B_x$为整数；去重服务器103包括确定单元1901和判断处理单元1902。其中，确定单元1901，用于用于执行步骤a)：a)依据所述规则为当前潜在分割点k_i确定点p_{iz}及所述点p_{iz}对应的窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$，i和z为整数，并且1≤z≤M；判断处理单元1902，用于所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足预定条件$C_z$；

当所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据不满足所述预定条件$C_z$，从所述点$p_{iz}$沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，N*U不大于$\|B_z\|+\max_x(\|A_x\|+\|(k_i\text{-}p_{ix})\|)$，

获得新的潜在分割点，则所述确定单元为所述新的潜在分割点执行步骤a)；当所述当前潜在分割点k_i的M个窗口中的每一个窗口$W_{ix}[p_{ix}-A_x, p_{ix}+B_x]$中至少部分数据满足预定条件$C_x$，则所述当前潜在分割点$k_i$为数据流分割点。

进一步地，所述规则还包括：至少两个点p_e和p_f，满足条件$A_e=A_f$，$B_e=B_f$，$C_e=C_f$。进一步地，所述规则还包括：所述至少两个点p_e和p_f，相对于所述潜在分割点k，在所述数据流分割点查找反方向上。

进一步地，所述规则还包括：所述至少两个点p_e和p_f之间的距离为1个U。

进一步地，所述判断处理单元1902具体用于使用随机函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$。具体地，所述判断处理单元1902具体用于使用hash函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$。具体地，所述判断处理单元1902具体用于使用随机函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$，具体包括：

在所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中选择F个字节，将所述F个字节反复利用H次，共获得F*H个字节，其中每个字节由8位组成，记为$a_{m,1}\ldots a_{m,8}$，表示所述F*H个字节中第m个字节的第1到第8位，所述F*H个字节对应的位可以表示为：$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,8} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ a_{F*H,1} & a_{F*H,2} & \cdots & a_{F*H,8} \end{pmatrix}$，当$a_{m,n}$=1时，$V_{am,n}$=1，当$a_{m,n}$=0时，$V_{am,n}$=-1，其中$a_{m,n}$表示$a_{m,1}\ldots a_{m,8}$中的任一个，所述F*H个字节对应的位按照$a_{m,n}$与$V_{am,n}$的转换关系得到矩阵$V_a$，所述矩阵

V_a表示为：$\begin{pmatrix} V_{a1,1} & V_{a1,2} & \cdots & V_{a1,8} \\ V_{a2,1} & V_{a2,2} & \cdots & V_{a2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{aF*H,1} & V_{aF*H,2} & \cdots & V_{aF*H,8} \end{pmatrix}$，从服务正态分布的随机数中选择F*H*8个随机数组成矩阵R，所述矩阵R表示为：

$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,8} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{F*H,1} & h_{F*H,2} & \cdots & h_{F*H,8} \end{pmatrix}$，将所述矩阵$V_a$的第m行与所述矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为$S_{am}=V_{am,1}*h_{m,1}+V_{am,2}*h_{m,2}+\ldots+V_{am,8}*h_{m,8}$，同理，获得$S_{a1}$、$S_{a2}$...到$S_{aF*H}$，统计$S_{a1}$、$S_{a2}$...到$S_{aF*H}$中满足大于0的值的个数K，当K为偶数，则所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据满足所述预定条件$C_z$。

进一步地，所述判断处理单元1902用于当所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据不满足所述预定条件$C_z$，从所述点$p_{iz}$沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，获得所述新的潜在分割点，所述确定单元1901为所述新的潜在分割点执行步骤a)，根据所述规则，为所述新的潜在分割点确定的点p_{ic}对应的窗口$W_{ic}[p_{ic}-A_c, p_{ic}+B_c]$的左边界与所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$的右边界重合或者为所述新的潜在分割点确定的所述窗口$W_{ic}[p_{ic}-A_c, p_{ic}+B_c]$的左边界位于所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$范围之内；其中，为所述新的潜在分割点确定的所述窗口$W_{ic}[p_{ic}-A_c, p_{ic}+B_c]$是根据所述规则，为所述新的潜在分割点确定的M个点按照数据流查找方向获得的序列中排序第一的点。

根据图4至图17所示的本发明实施例提供的基于服务器查找数据流分割点的方法中，为潜在分割点k_i确定点p_{ix}及点p_{ix}的窗口$W_{ix}[p_{ix}-A_x, p_{ix}+B_x]$，其中，x分别为1到M连续的自然数，M≥2，可以并行判

断M个窗口中每一个窗口中至少部分数据是否满足预定条件C_x，或者依次判断窗口中至少部分数据是否满足预定条件，也可以判断窗口W_{i1} [p_{i1}-A_1， p_{i1}+B_1]中至少部分数据满足预定条件C_1时，再判断W_{i2} [p_{i2}-A_2， p_{i2}+B_2]中至少部分数据满足预定条件C_2时，直到判断W_{im} [p_{im}-A_m， p_{im}+B_m]中至少部分数据满足预定条件C_m。实施例中其他窗口的判断与此相同，不再赘述。

另外，根据根据图4至图17所示的本发明实施例，实际应用中，在去重服务器103上预设有规则，所述规则为：为潜在分割点k确定M个点p_x、点p_x对应的窗口W_x[p_x-A_x, p_x+B_x]和窗口W_x[p_x-A_x, p_x+B_x]对应的预定条件C_x，x分别为1到M连续的自然数，M≥2，在该预设规则中，A_1、A_2、A_3…A_m可以不全部相等，B_1、B_2、B_3…B_m可以不全部相等，C_1、C_2、C_3…C_M也可以不全部相同。在图5所示的实施方式中，在窗口W_{i1}[p_{i1}-169, p_{i1}]、W_{i2}[p_{i2}-169, p_{i2}]、W_{i3}[p_{i3}-169, p_{i3}]、W_{i4}[p_{i4}-169, p_{i4}]、W_{i5}[p_{i5}-169, p_{i5}]、W_{i6}[p_{i6}-169, p_{i6}]、W_{i7}[p_{i7}-169, p_{i7}]、W_{i8}[p_{i8}-169, p_{i8}]、W_{i9}[p_{i9}-169, p_{i9}]、W_{i10}[p_{i10}-169, p_{i10}]和W_{i11}[p_{i11}-169, p_{i11}]中，各窗口大小相同,即窗口大小均为169字节，同时判断窗口中至少部分数据是否满足预定条件的方式也相同,具体见上述判断W_{i1}[p_{i1}-169, p_{i1}]中至少部分数据是否满足预定条件C_1的描述，但在图11所示的实施方式中，W_{i1}[p_{i1}-169, p_{i1}]、W_{i2}[p_{i2}-169, p_{i2}]、W_{i3}[p_{i3}-169, p_{i3}]、W_{i4}[p_{i4}-169, p_{i4}]、W_{i5}[p_{i5}-169, p_{i5}]、W_{i6}[p_{i6}-169, p_{i6}]、W_{i7}[p_{i7}-169, p_{i7}]、W_{i8}[p_{i8}-169, p_{i8}]、W_{i9}[p_{i9}-169, p_{i9}]、W_{i10}[p_{i10}-169, p_{i10}]与W_{i11}[p_{i11}-182, p_{i11}]窗口大小可以不相同,同时判断窗口中至少部分数据是否满足预定条件的方式也可以不相同。在所有实施例中，根据在去重服务器103上预设的规则，判断窗口W_{i1}中至少部分数据是否满足预定条件

C_1的方式与判断窗口W_{j1}中至少部分数据是否满足预定条件C_1的方式必然相同，判断W_{i2}中至少部分数据是否满足预定条件C_2的方式与判断W_{j2}中至少部分数据是否满足预定条件C_2的方式必然相同...判断窗口W_{iM}中至少部分数据是否满足预定条件C_M的方式与判断窗口W_{jM}中至少部分数据是否满足预定条件C_M的方式必然相同。在此不再赘述，同时根据图4至图17所示的本发明实施例，虽然均以M=11为例，但根据实际需要，M的取值并不限于11，本领域技术人员根据本发明实施例中的描述，确定M的值。

根据图4至图17所示的本发明实施例，在去重服务器103上预设有规则，k_a、k_i、k_j、k_l和k_m为沿着数据流分割点查找方向查找分割点时获得的潜在分割点，k_a、k_i、k_j、k_l和k_m都依据该规则。本发明实施例中的窗口$W_x[p_x-A_x,p_x+B_x]$表示一个特定范围，在该特定范围选择数据以判断这些数据是否满足预定条件C_x，具体地，可以在该特定范围内选择部分数据，也可以选择全部数据以判断这些数据是否满足预定条件C_x。本发明实施例中具体使用的窗口概念可参照窗口$W_x[p_x-A_x,p_x+B_x]$，在此不再赘述。

根据图4至图17所示的本发明实施例，窗口$W_x[p_x-A_x,p_x+B_x]$中，（p_x-A_x）和（p_x+B_x）表示该窗口$W_x[p_x-A_x,p_x+B_x]$的两个边界，其中（p_x-A_x）表示窗口$W_x[p_x-A_x,p_x+B_x]$相对于点p_x位于数据流分割点查找反方向的边界，（p_x+B_x）表示窗口$W_x[p_x-A_x,p_x+B_x]$相对于点p_x位于数据流分割点查找方向的边界。具体地，在本发明实施例中，在图3至图15所示的数据流分割点查找方向为从左向右，则其中（p_x-A_x）表示窗口$W_x[p_x-A_x,p_x+B_x]$相对于点p_x位于数据流分割点查找反方向的边界（即左边界），（p_x+B_x）表示窗口$W_x[p_x-A_x,p_x$

$+B_x$]相对于点p_x位于数据流分割点查找方向的边界（即右边界）。如果在图3至图15所示的数据流分割点查找方向为从右向左，则其中（p_x-A_x）表示窗口$W_x[p_x-A_x, p_x+B_x]$相对于点p_x位于数据流分割点查找反方向的边界（即右边界），（p_x+B_x）表示窗口$W_x[p_x-A_x, p_x+B_x]$相对于点p_x位于数据流分割点查找方向的边界（即左边界）。

本领域普通技术人员可以意识到，结合本发明实施例描述的各示例的单元及算法步骤，本发明实施例的关键特征可以与其他技术相结合，以更为复杂的形式呈现，但仍会包含本发明的关键特征。在真实环境中可能使用备用分割点，例如一种实施方式为，根据在去重服务器103上预设的规则，为潜在分割点k_i确定11个点p_x，x为1到11连续的自然数，确定p_x对应的窗口$W_x[p_x-A_x, p_x+B_x]$及窗口$W_x[p_x-A_x, p_x+B_x]$对应的预定条件C_x，当11个窗口中每一个窗口$W_x[p_x-A_x, p_x+B_x]$中至少部分数据均满足预定条件C_x，则潜在分割点k_i为数据流分割点，当超过设定的最大数据块时，仍未查找到分割点，这时可能使用备用的预设规则，备用的预设规则与在去重服务器103上预设的规则类似，备用的预设规则为：例如为潜在分割点k_i确定10个点p_x，x为1到10连续的自然数，确定p_x对应的窗口$W_x[p_x-A_x, p_x+B_x]$及窗口$W_x[p_x-A_x, p_x+B_x]$对应的预定条件C_x，当10个窗口中每一个窗口$W_x[p_x-A_x, p_x+B_x]$中至少部分数据均满足预定条件C_x，则潜在分割点k_i为数据流分割点，当超过设定的最大数据块时，仍未查找到数据流分割点时，从最大数据块的结束位置作为强制分割点。

在去重服务器 103 上预设有规则，所述规则中为潜在分割点 k 确定 M 个点，并不一定要求先有一个潜在分割点 k，可以通过确定的 M 个点来判断潜在分割点 k。

本发明实施例提供一种基于去重服务器查找数据流分割点的方

法，如图20所示，包括:

在去重服务器103上预设有规则，所述规则为：为潜在分割点k确定M个窗口$W_x[k-A_x,k+B_x]$和窗口$W_x[k-A_x,k+B_x]$对应的预定条件C_x，其中，x为1到M连续的自然数，M≥2，、A_x和B_x为整数；在图3所示的实施方式中，关于M的取值，其中一种实现方式，M*U取值不大于预设的两个相邻的数据流分割点之间的最大距离，即预设的数据块最大长度。判断窗口$W_z[k-A_z，k+B_z]$中至少部分数据是否满足预定条件C_z，其中，z为整数，1≤z≤M，（$k-A_z$）与（$k+B_z$）分别表示窗口W_z的两个边界。当判断任意一个窗口$W_z[k-A_z，k+B_z]$中至少部分数据不满足预定条件C_z，则从潜在分割点k沿数据流分割点查找方向跳跃N个字节，N≤‖B_z‖+max$_x$(‖A_x‖)。其中，‖B_z‖表示$W_z[k-A_z，k+B_z]$中B_z的绝对值，max$_x$(‖A_x‖)表示M个窗口中A_x绝对值中的最大值，将在下面实施例中具体介绍N取值的原理。当判断M个窗口中的每一个窗口$W_x[k-A_x,k+B_x]$中至少部分数据满足预定条件C_x，则潜在分割点k为数据流分割点。

具体地，对当前潜在分割点k_i，依据所述规则，执行以下步骤:

步骤2001:依据所述规则为当前潜在分割点k_i确定对应的窗口$W_{iz}[k_i-A_z,k_i+B_z]$，i和z为整数,并且1≤z≤M；

步骤2002:判断所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据是否满足预定条件C_z；

当所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据不满足所述预定条件C_z，从所述当前潜在分割点k_i沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，N*U不大于‖B_z‖+max$_x$(‖A_x‖)，获得新的潜在分割点，执行步骤2001 ；

当所述当前潜在分割点k_i的M个窗口中的每一个窗口$W_{ix}[k_i-A_x,k_i$

$+B_x]$中至少部分数据满足预定条件C_x，则所述当前潜在分割点k_i为数据流分割点。

进一步地,所述规则还包括：至少两个窗口$W_{ie}[k_i-A_e,k_i+B_e]$与$W_{if}[k_i-A_f,k_i+B_f]$，满足条件：$|A_e+B_e|=|A_f+B_f|$，$C_e=C_f$；进一步地，所述规则还包括：A_e和A_f为正整数;更进一步地,所述规则还包括：$A_e-1=A_f$，$B_e+1=B_f$。其中，$|A_e+B_e|$表示窗口W_{ie}的大小，$|A_f+B_f|$表示窗口W_{if}的大小。

进一步地，判断所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据是否满足所述预定条件C_z，具体包括：使用随机函数判断所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据是否满足所述预定条件C_z;更进一步地，所述使用随机函数判断所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据是否满足所述预定条件C_z，具体为使用hash函数判断所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据是否满足所述预定条件C_z。

当所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据不满足所述预定条件C_z，从所述当前潜在分割点k_i沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，获得所述新的潜在分割点，根据所述规则，为所述新的潜在分割点确定的窗口$W_{ic}[k_i-A_c,k_i+B_c]$的左边界与所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$的右边界重合或者为所述新的潜在分割点确定的所述窗口$W_{ic}[k_i-A_c,k_i+B_c]$的左边界位于所述窗口$W_{iz}[k_i-A_z,k_i+B_z]$范围之内；其中，为所述新的潜在分割点确定的所述窗口$W_{ic}[k_i-A_c,k_i+B_c]$是根据所述规则，为所述新的潜在分割点确定的M个窗口按照数据流查找方向获得的序列中排序第一的窗口。

本发明实施例中通过判断M个窗口中某一个窗口中至少部分数据是否满足预定条件，来查找数据流分割点，当某一个窗口中至少部

分数据不满足预定条件，则跳过N*U个长度，其中，N*U不大于 $\| B_z \| + \max_x(\| A_x \|)$，获得下一个潜在分割点，提高了数据流分割点查找效率。

在重复数据删除过程中，为保证数据块大小均匀，会考虑平均数据块（也可以称为平均分块）大小，即在满足最小数据块大小和最大数据块大小限定的同时，会确定平均数据块大小，以保证获得的数据块大小均匀。窗口 $W_x[k-A_x, k+B_x]$ 的个数M与窗口 $W_x[k-A_x, k+B_x]$ 中至少部分数据满足预设条件的概率这两个因素决定了找到数据流分割点的概率（以P(n)表示），前者影响跳跃的长度，后者影响跳跃的概率，二者共同影响平均分块大小。一般而言，在平均分块大小固定时，$W_x[k-A_x, k+B_x]$ 个数增加，则单个窗口 $W_x[k-A_x, k+B_x]$ 中至少部分数据满足预定条件的概率也增加，例如在去重服务器103上预设有规则，所述规则为：为潜在分割点 k 确定11个窗口 $W_x[k-A_x, k+B_x]$，x分别为1到11连续的自然数，11个窗口中任一个窗口 $W_x[k-A_x, k+B_x]$ 中至少部分数据满足预设条件的概率为1/2。而在去重服务器103上预设的另一组规则为：为潜在分割点 k 确定24个窗口 $W_x[k-A_x, k+B_x]$，x分别为1到24连续的自然数，24个窗口中任一个窗口 $W_x[k-A_x, k+B_x]$ 中至少部分数据满足预设条件的概率3/4，具体窗口 $W_x[k-A_x, k+B_x]$ 中至少部分数据满足预设条件的概率设定可参见判断窗口 $W_x[k-A_x, k+B_x]$ 中至少部分数据是否满足预设条件部分的描述。窗口 $W_x[k-A_x, k+B_x]$ 的个数M与窗口 $W_x[k-A_x, k+B_x]$ 中至少部分数据满足的预设条件的概率这两个因素决定P(n)，P（n）表示：从数据流起始位置或者从上一数据流分割点查找n个数据流分割点最小查找单位后没找到数据流分割点的概率。关于这两个因素决定P（n）的计算过程，实际上是一个多步长Fibonacci数列，后面将具体描述。得到P（n）后，1-P(n)

即为数据流分割点的分布函数，(1-P(n))-(1-P(n-1))=P(n-1)-P(n)即为在n个数据流分割点最小查找单位找到数据流分割点概率，也就是数据流分割点的密度函数，根据数据流分割点的密度函数就可以积分 $\sum_{n=4*1024}^{12*1024} n\times(P(n-1)-P(n))$，从而求得数据流分割点的期望长度，即平均分块大小,其中，4*1024（字节）表示最小数据块长度，12*1024（字节）表示最大数据块长度。

在图3所示的数据流分割点查找的基础上，在图21所示的实施方式中，在去重服务器103上预设有规则，所述规则为：为潜在分割点k确定11个窗口$W_x[k-A_x,k+B_x]$和窗口$W_x[k-A_x,k+B_x]$对应的预定条件C_x，其中，x为1到11连续的自然数，A_x和B_x为整数。其中，A_1=169，B_1=0；A_2=170，B_2=-1；A_3=171，B_3=-2；A_4=172，B_4=-3；A_5=173，B_5=-4；A_6=174，B_6=-5；A_7=175，B_7=-6；A_8=176，B_8=-7；A_9=177，B_9=-8；A_{10}=178，B_{10}=-9；A_{11}=179，B_{11}=-10，并且$C_1=C_2=C_3=C_4=C_5=C_6=C_7=C_8=C_9=C_{10}=C_{11}$，则11个窗口分别为$W_1$[k-169, k]、$W_2$[k-170, k-1]、$W_3$[k-171, k-2]、$W_4$[k-172, k-3]、$W_5$[k-173, k-4]、$W_6$[k-174, k-5]、$W_7$[k-175, k-6]、$W_8$[k-176, k-7]、$W_9$[k-177, k-8]、$W_{10}$[k-178, k-9]和$W_{11}$[k-179, k-10]。$k_a$为数据流分割点，图21中所示数据流分割点查找方向为从左向右，从数据流分割点k_a跳过最小数据块4KB后，最小数据块4KB结束位置作为下一个潜在分割点k_i，根据为去重服务器103预设的规则，为潜在分割点k_i确定窗口$W_{ix}[k_i-A_x,k_i+B_x]$,在本实施例中，x分别为1到11连续的自然数。在图21所示的实施方式中，为潜在分割点k_i确定的窗口为11个，分别为$W_{i1}[k_i-169,k_i]$、$W_{i2}[k_i-170,k_i-1]$、$W_{i3}[k_i-171,k_i-2]$、$W_{i4}[k_i-172,k_i-3]$、$W_{i5}[k_i-173,k_i-4]$、$W_{i6}[k_i-174,k_i-5]$、$W_{i7}[k_i-175,k_i-6]$、$W_{i8}[k_i-176,k_i-7]$、$W_{i9}[k_i-177,k_i-8]$、$W_{i10}[k_i-178,k_i-9]$和$W_{i11}[k_i-179,k_i-10]$。判断$W_{i1}[k_i-169,k_i]$中至

少部分数据是否满足预定条件C_1、判断$W_{i2}[k_i\text{-}170,k_i\text{-}1]$中至少部分数据是否满足预定条件$C_2$、判断$W_{i3}[k_i\text{-}171,k_i\text{-}2]$中至少部分数据是否满足预定条件$C_3$、判断$W_{i4}[k_i\text{-}172,k_i\text{-}3]$中至少部分数据是否满足预定条件$C_4$、判断$W_{i5}[k_i\text{-}173,k_i\text{-}4]$中至少部分数据是否满足预定条件$C_5$、判断$W_{i6}[k_i\text{-}174,k_i\text{-}5]$中至少部分数据是否满足预定条件$C_6$、判断$W_{i7}[k_i\text{-}175,k_i\text{-}6]$中至少部分数据是否满足预定条件$C_7$、判断$W_{i8}[k_i\text{-}176,k_i\text{-}7]$中至少部分数据是否满足预定条件$C_8$、判断$W_{i9}[k_i\text{-}177,k_i\text{-}8]$中至少部分数据是否满足预定条件$C_9$、判断$W_{i10}[k_i\text{-}178,k_i\text{-}9]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{i11}[k_i\text{-}179,k_i\text{-}10]$中至少部分数据是否满足预定条件$C_{11}$。当判断窗口$W_{i1}$中至少部分数据满足预定条件$C_1$、窗口$W_{i2}$中至少部分数据满足预定条件$C_2$、窗口$W_{i3}$中至少部分数据满足预定条件$C_3$、窗口$W_{i4}$中至少部分数据满足预定条件$C_4$、窗口$W_{i5}$中至少部分数据满足预定条件$C_5$、窗口$W_{i6}$中至少部分数据满足预定条件$C_6$、窗口$W_{i7}$中至少部分数据满足预定条件$C_7$、窗口$W_{i8}$中至少部分数据满足预定条件$C_8$、窗口$W_{i9}$中至少部分数据满足预定条件$C_9$、窗口$W_{i10}$中至少部分数据满足预定条件$C_{10}$和窗口$W_{i11}$中至少部分数据满足预定条件$C_{11}$时，则当前潜在分割点$k_i$为数据流分割点。当11个窗口中任一个窗口中至少部分数据不满足对应的预定条件时，如图22所示，$W_{i5}[k_i\text{-}173,k_i\text{-}4]$，则从潜在分割点$k_i$沿着数据流分割点查找方向跳跃N个字节，其中N个字节不大于$\| B_5 \| + \max_x(\| A_x \|)$，在图22所示的实施方式中，跳跃N个字节不大于183个字节，在本实施例中，N=7，得到新的潜在分割点，为与潜在分割点k_i区别，这里将新的潜在分割点表示为k_j。根据图21所示的实施方式中，在去重服务器103上预设有规则，所述规则为：为潜在分割点k_j确定窗口$W_{jx}[k_j\text{-}A_x,k_j+B_x]$，

在本实施例中， x分别为1到11连续的自然数。为潜在分割点k_j确定的窗口为11个，分别为$W_{j1}[k_j\text{-}169,k_j]$、$W_{j2}[k_j\text{-}170,k_j\text{-}1]$、$W_{j3}[k_j\text{-}171,k_j\text{-}2]$、$W_{j4}[k_j\text{-}172,k_j\text{-}3]$、$W_{j5}[k_j\text{-}173,k_j\text{-}4]$、$W_{j6}[k_j\text{-}174,k_j\text{-}5]$、$W_{j7}[k_j\text{-}175,k_j\text{-}6]$、$W_{j8}[k_j\text{-}176,k_j\text{-}7]$、$W_{j9}[k_j\text{-}177,k_j\text{-}8]$、$W_{j10}[k_j\text{-}178,k_j\text{-}9]$和$W_{j11}[k_j\text{-}179,k_j\text{-}10]$。如图22所示，为潜在分割点确定的第11个窗口$W_{j11}[k_j\text{-}179,k_j\text{-}10]$，在保证潜在分割点$k_i$与潜在分割点$k_j$之间的范围都在判断范围之内，则在本实施方式中，必须保证窗口$W_{j11}[k_j\text{-}179,k_j\text{-}10]$的左边界与窗口$W_{i5}[k_i\text{-}173,k_i\text{-}4]$的右边界（$k_i\text{-}4$）重合，或者位于窗口$W_{i5}[k_i\text{-}173,k_i\text{-}4]$范围之内，所述窗口$W_{j11}[k_j\text{-}179,k_j\text{-}10]$是根据所述规则，为所述潜在分割点$k_j$确定的M个窗口按照数据流查找方向获得的序列中排序第一的窗口。因此，在这一限定内，当窗口$W_{i5}[k_i\text{-}173,k_i\text{-}4]$中至少部分数据不满足预定条件$C_5$，从潜在分割点$k_i$沿数据流分割点查找方向跳跃的距离不大于$\| B_5 \| + \max_x(\| A_x \|)$。判断$W_{j1}[k_j\text{-}169,k_j]$中至少部分数据是否满足预定条件$C_1$、判断$W_{j2}[k_j\text{-}170,k_j\text{-}1]$中至少部分数据是否满足预定条件$C_2$、判断$W_{j3}[k_j\text{-}171,k_j\text{-}2]$中至少部分数据是否满足预定条件$C_3$、判断$W_{j4}[k_j\text{-}172,k_j\text{-}3]$中至少部分数据是否满足预定条件$C_4$、判断$W_{j5}[k_j\text{-}173,k_j\text{-}4]$中至少部分数据是否满足预定条件$C_5$、判断$W_{j6}[k_j\text{-}174,k_j\text{-}5]$中至少部分数据是否满足预定条件$C_6$、判断$W_{j7}[k_j\text{-}175,k_j\text{-}6]$中至少部分数据是否满足预定条件$C_7$、判断$W_{j8}[k_j\text{-}176,k_j\text{-}7]$中至少部分数据是否满足预定条件$C_8$、判断$W_{j9}[k_j\text{-}177,k_j\text{-}8]$中至少部分数据是否满足预定条件$C_9$、判断$W_{j10}[k_j\text{-}178,k_j\text{-}9]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{j11}[k_j\text{-}179,k_j\text{-}10]$中至少部分数据是否满足预定条件$C_{11}$。当判断窗口$W_{j1}$中至少部分数据满足预定

条件C_1、窗口W_{j2}中至少部分数据满足预定条件C_2、窗口W_{j3}中至少部分数据满足预定条件C_3、窗口W_{j4}中至少部分数据满足预定条件C_4、窗口W_{j5}中至少部分数据满足预定条件C_5、窗口W_{j6}中至少部分数据满足预定条件C_6、窗口W_{j7}中至少部分数据满足预定条件C_7、窗口W_{j8}中至少部分数据满足预定条件C_8、窗口W_{j9}中至少部分数据满足预定条件C_9、窗口W_{j10}中至少部分数据满足预定条件C_{10}和窗口W_{j11}中至少部分数据满足预定条件C_{11}时，则当前潜在分割点k_i为数据流分割点，k_j与k_a之间的数据构成1个数据块，同时按照与k_a相同的方式跳过最小分块大小4KB，获得下一个潜在分割点，并按照在去重服务器103上预设的规则，判断下一个潜在分割点是否为数据流分割点。当判断潜在分割点k_j不是数据流分割点时，按照与k_i相同的方式获得下一个潜在分割点，并按照在去重服务器103上预设的规则及上述方法判断下一个潜在分割点是否为数据流分割点。当超过设定的最大数据块仍然没有找到数据流分割点时，则从最大数据块的结束位置作为强制分割点。

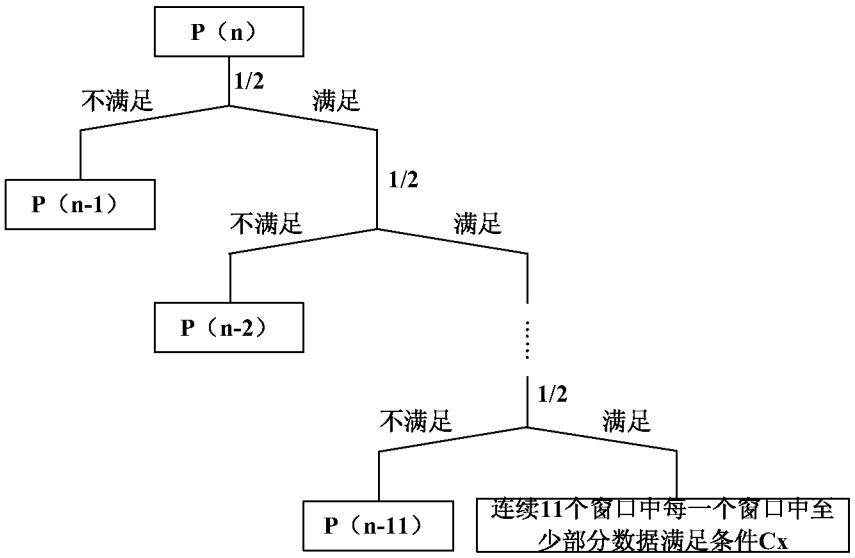
在如图21所示的实施方式中，按照在去重服务器103上预设的规则，从判断$W_{i1}[k_i\text{-}169,k_i]$中至少部分数据是否满足预定条件$C_1$开始，当判断$W_{i1}[k_i\text{-}169,k_i]$、$W_{i2}[k_i\text{-}170,k_i\text{-}1]$、$W_{i3}[k_i\text{-}171,k_i\text{-}2]$和$W_{i4}[k_i\text{-}172,k_i\text{-}3]$中至少部分数据中至少部分数据分别满足预定条件$C_1$、$C_2$、$C_3$和$C_4$，判断$W_{i5}[k_i\text{-}173,k_i\text{-}4]$中至少部分数据不满足预定条件$C_5$时，从潜在分割点$k_i$沿着数据流分割点查找方向跳跃6个字节，在第6个字节的结束位置获得新的潜在分割点，为与其他潜在分割点区别，这里表示为k_g，按照在去重服务器103上预设的规则，为潜在分割点k_g确定11个窗口，分别为$W_{g1}[k_g\text{-}169,k_g]$、$W_{g2}[k_g\text{-}170,k_g\text{-}1]$、$W_{g3}[k_g\text{-}171,k_g\text{-}2]$、

$W_{g4}[k_g\text{-}172,k_g\text{-}3]$、$W_{g5}[k_g\text{-}173,k_g\text{-}4]$、$W_{g6}[k_g\text{-}174,k_g\text{-}5]$、$W_{g7}[k_g\text{-}175,k_g\text{-}6]$、$W_{g8}[k_g\text{-}176,k_g\text{-}7]$、$W_{g9}[k_g\text{-}177,k_g\text{-}8]$、$W_{g10}[k_g\text{-}178,k_g\text{-}9]$和$W_{g11}[k_g\text{-}179,k_g\text{-}10]$。判断$W_{g1}[k_g\text{-}169,k_g]$中至少部分数据是否满足预定条件$C_1$、判断$W_{g2}[k_g\text{-}170,k_g\text{-}1]$中至少部分数据是否满足预定条件$C_2$、判断$W_{g3}[k_g\text{-}171,k_g\text{-}2]$中至少部分数据是否满足预定条件$C_3$、判断$W_{g4}[k_g\text{-}172,k_g\text{-}3]$中至少部分数据是否满足预定条件$C_4$、判断$W_{g5}[k_g\text{-}173,k_g\text{-}4]$中至少部分数据是否满足预定条件$C_5$、判断$W_{g6}[k_g\text{-}174,k_g\text{-}5]$中至少部分数据是否满足预定条件$C_6$、判断$W_{g7}[k_g\text{-}175,k_g\text{-}6]$中至少部分数据是否满足预定条件$C_7$、判断$W_{g8}[k_g\text{-}176,k_g\text{-}7]$中至少部分数据是否满足预定条件$C_8$、判断$W_{g9}[k_g\text{-}177,k_g\text{-}8]$中至少部分数据是否满足预定条件$C_9$、判断$W_{g10}[k_g\text{-}178,k_g\text{-}9]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{g11}[k_g\text{-}179,k_g\text{-}10]$中至少部分数据是否满足预定条件$C_{11}$。窗口$W_{g11}[k_g\text{-}179,k_g\text{-}10]$与窗口$W_{i5}[k_i\text{-}173,k_i\text{-}4]$重合，并且$C_5=C_{11}$，因此，当判断$W_{i5}[k_i\text{-}173,k_i\text{-}4]$中至少部分数据不满足预定条件$C_5$时，从潜在分割点$k_i$沿着数据流分割点查找方向跳跃T个字节，获得的潜在分割点k_g仍然不符合作为数据流分割点的条件。因此，如果从潜在分割点k_i沿着数据流分割点查找方向跳跃6个字节会存在重复计算，因此，从潜在分割点k_i沿着数据流分割点查找方向跳跃7个字节可以减少重复计算，效率更高。因此提高了查找数据流分割点的速度。当预设规定中窗口W_x[k-A_x,k+B_x]中至少部分数据满足预定条件C_x的概率为1/2时，即是说以1/2的概率执行跳跃，每次最多可以跳跃$\|B_{11}\|+\|A_{11}\|$=189个字节。

在本实施方式中，预定规则为:为潜在分割点k确定11个窗口W_x[k-A_x,k+B_x]及窗口W_x[k-A_x,k+B_x]中至少部分数据满足预设条件C_x，

其中W_x[k-A_x,k+B_x]中至少部分数据满足预设条件C_x的概率为1/2，x分别为1到11连续的自然数并且A_x和B_x为整数。其中，A_1=169，B_1=0；A_2=170，B_2=-1；A_3=171，B_3=-2；A_4=172，B_4=-3；A_5=173，B_5=-4；A_6=174，B_6=-5；A_7=175，B_7=-6；A_8=176，B_8=-7；A_9=177，B_9=-8；A_{10}=178，B_{10}=-9；A_{11}=179，B_{11}=-10，并且C_1=C_2=C_3=C_4=C_5=C_6=C_7=C_8=C_9=C_{10}=C_{11}。即为潜在分割点k选择11个窗口，并且为连续11个窗口，通过这两个因素可以计算P(n)。11个窗口的选择方式及判断11个窗口中的每一个窗口中至少部分数据满足预定条件C_x遵循在去重服务器103上预设的规则，因此是否存在连续11个窗口中每一个窗口中至少部分数据满足预定条件C_x就决定潜在分割点k是否为数据流分割点。我们称两个字节之间的间隙为一个点。P（n）表示：连续的n个窗口内不存在连续的11个满足条件的窗口的概率，即不存在数据流分割点的概率。从文件头/上一分割点跳跃最小分块大小4KB后，向数据流分割点查找反方向回退10个字节，找到第4086个点，在该点处不存在数据流分割点，所以P（4086）=1，依次类推，P（4087）=1，……P（4095）=1。在第4096个点处，即在最小分块大小处，以（1/2）^11的概率这11个窗口中每一个窗口中至少部分数据满足预定条件C_x，因此以（1/2）^11的概率存在数据流分割点，以1-（1/2）^11的概率不存在数据流分割点，所以P（4096）=1-（1/2）^11。

在第n个窗口处，可以分为12种情况来递推P（n）。



情况1：第n个窗口中至少部分数据以1/2的概率不满足预定条件，此时第n个窗口前面的n-1个窗口以P（n-1）的概率不存在连续的11个窗口中每一个窗口至少部分数据均满足预定条件，因此P(n)包含1/2*P（n-1）。第n个窗口中至少部分数据不满足预定条件，并且且第n个点前面的n-1个窗口存在连续的11个窗口每一个窗口中至少部分数据均满足预定条件的情况与P（n）无关。

情况2：第n个窗口中至少部分数据以1/2的概率满足预定条件，第n-1个窗口中至少部分数据以1/2的概率不满足预定条件，此时第n-1个窗口前面的n-2个窗口以P（n-2）的概率不存在连续的11个窗口中每一个窗口中至少部分数据均满足预定条件，因此P(n)包含1/2*1/2*P（n-2）。第n个窗口中至少部分数据满足预定条件，第n-1个点窗口中至少部分数据不满足预定条件，并且第n-1个窗口前面的n-2个窗口存在连续的11个窗口中每一个窗口至少部分数据满足预定条件的情况与P（n）无关。

依照上述描述，情况11：第n至n-9个窗口中至少部分数据以(1/2)^10的概率满足预定条件，第n-10个窗口中至少部分数据以1/2的概率

不满足预定条件，此时第n-10个窗口前面的n-11个窗口以P（n-11）的概率不存在连续的11个窗口中每一个窗口中至少部分数据均满足预定条件，因此P（n）包含（1/2）^10*1/2*P（n-11）。第n至n-9个窗口中至少部分数据均满足预定条件，第n-10个窗口中至少部分数据不满足预定条件，并且第n-10个窗口前面的n-11个窗口存在连续的11个窗口中每一个窗口中至少部分数据均满足预定条件的情况与P（n）无关。

情况12：第n至n-10个的窗口中至少部分数据以（1/2）^11的概率满足预定条件，该情况与P（n）无关。

因此，P（n）=1/2* P（n-1）+（1/2）^2* P（n-2）+……+（1/2）^11*P（n-11）。另一种预设规则：为潜在分割点k确定24个窗口$W_x[k-A_x,k+B_x]$和窗口$W_x[k-A_x,k+B_x]$对应的预定条件C_x，其中，x为1到11连续的自然数，A_1=169，B_1=0；A_2=170，B_2=-1；A_3=171，B_3=-2；A_4=172，B_4=-3；A_5=173，B_5=-4；A_6=174，B_6=-5；A_7=175，B_7=-6；A_8=176，B_8=-7；A_9=177，B_9=-8；A_{10}=178，B_{10}=-9；A_{11}=179，B_{11}=-10，…A_{24}=192，B_{24}=-23,并且$C_1=C_2=C_3=C_4=C_5=C_6=C_7=C_8=C_9=\ldots=C_{24}$，窗口$W_x[k-A_x,k+B_x]$中至少部分数据满足预定条件$C_x$的概率为3/4，通过这两个因素可以计算P(n)。

因此是否存在连续24个窗口中的每一个窗口中至少部分数据均满足预定条件C_x就决定潜在分割点k是否为数据流分割点，可以通过下面的公式计算：

P（1）=1，P（2）……P（23）=1，P（24）=1-（3/4）^24，

P（n）=1/4* P（n-1）+1/4*（3/4）* P（n-2）+……+1/4*（3/4）^23*P（n-24）。

经过计算，P(5*1024)=0.78，P(11*1024)=0.17，P(12*1024)=0.13，即从数据流起始位置/上一数据流分割点查找到12KB后以13%的概率

仍未找到数据流分割点，强制进行分割。通过这个概率，求得数据流分割点的密度函数，经过积分求得大约平均在从数据流起始位置/上一数据流分割点查找7.6KB时找到数据流分割点，即平均分块长度大约为7.6KB。与连续的11个窗口中至少部分数据以1/2的概率满足预定条件不同，传统CDC算法采用一个窗口以1/2^12的概率满足条件时，方可达到平均分块长度7.6KB的效果。

在图3所示的数据流分割点查找的基础上，在图23所示的实施方式中，在去重服务器103上预设有规则，所述规则为：为潜在分割点k确定11个窗口$W_x[k-A_x,k+B_x]$和窗口$W_x[k-A_x,k+B_x]$对应的预定条件C_x，其中，x为1到11连续的自然数，A_x和B_x为整数。其中，窗口$W_x[k-A_x,k+B_x]$中至少部分数据满足预定条件C_x的概率为1/2，A_1=171，B_1=-2；A_2=172，B_2=-3；A_3=173，B_3=-4；A_4=174，B_4=-5；A_5=175，B_5=-6；A_6=176，B_6=-7；A_7=177，B_7=-8；A_8=178，B_8=-9；A_9=179，B_9=-10；A_{10}=170，B_{10}=-1；A_{11}=169，B_{11}=0，并且$C_1=C_2=C_3=C_4=C_5=C_6=C_7=C_8=C_9=C_{10}=C_{11}$。$k_a$为数据流分割点，图23中所示数据流分割点查找方向为从左向右，从数据流分割点k_a跳过最小数据块4KB后，在最小数据块4KB结束位置作为下一个潜在分割点k_i，根据在去重服务器103上预设的规则，为潜在分割点k_i确定W_x[k-A_x,k+B_x]及窗口W_x[k-A_x,k+B_x]对应的预设条件C_x，其中x为1到11连续的自然数。确定的11个窗口分别为$W_{i1}[k_i-171,k_i-2]$、$W_{i2}[k_i-172,k_i-3]$、$W_{i3}[k_i-173,k_i-4]$、$W_{i4}[k_i-174,k_i-5]$、$W_{i5}[k_i-175,k_i-6]$、$W_{i6}[k_i-176,k_i-7]$、$W_{i7}[k_i-177,k_i-8]$、$W_{i8}[k_i-178,k_i-9]$、$W_{i9}[k_i-179,k_i-10]$、$W_{i10}[k_i-170,k_i-1]$和$W_{i11}[k_i-169,k_i]$。判断$W_{i1}[k_i-171,k_i-2]$中至少部分数据是否满足预定条件C_1、判断$W_{i2}[k_i-172,k_i-3]$中至少部分数据是否满足预定条件C_2、判断$W_{i3}[k_i-173,k_i-4]$中至少部分数据是否满足预定条件C_3、判断$W_{i4}[k_i-174,k_i-5]$中至少部分数据是否满足预定条件C_4、判

断$W_{i5}[k_i\text{-}175,k_i\text{-}6]$中至少部分数据是否满足预定条件$C_5$、判断$W_{i6}[k_i\text{-}176,k_i\text{-}7]$中至少部分数据是否满足预定条件$C_6$、判断$W_{i7}[k_i\text{-}177,k_i\text{-}8]$中至少部分数据是否满足预定条件$C_7$、判断$W_{i8}[k_i\text{-}178,k_i\text{-}9]$中至少部分数据是否满足预定条件$C_8$、判断$W_{i9}[k_i\text{-}179,k_i\text{-}10]$中至少部分数据是否满足预定条件$C_9$、判断$W_{i10}[k_i\text{-}170,k_i\text{-}1]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{i11}[k_i\text{-}169,k_i]$中至少部分数据是否满足预定条件$C_{11}$。当判断窗口$W_{i1}$中至少部分数据满足预定条件$C_1$、窗口$W_{i2}$中至少部分数据满足预定条件$C_2$、窗口$W_{i3}$中至少部分数据满足预定条件$C_3$、窗口$W_{i4}$中至少部分数据满足预定条件$C_4$、窗口$W_{i5}$中至少部分数据满足预定条件$C_5$、窗口$W_{i6}$中至少部分数据满足预定条件$C_6$、窗口$W_{i7}$中至少部分数据满足预定条件$C_7$、窗口$W_{i8}$中至少部分数据满足预定条件$C_8$、窗口$W_{i9}$中至少部分数据满足预定条件$C_9$、窗口$W_{i10}$中至少部分数据满足预定条件$C_{10}$和窗口$W_{i11}$中至少部分数据满足预定条件$C_{11}$时，则当前潜在分割点$k_i$为数据流分割点。当11个窗口中任一个窗口中至少部分数据不满足对应的预定条件时，如图24所示，$W_{i3}[p_{i3}\text{-}169,p_{i3}]$中至少部分数据不满足预定条件$C_3$，点$p_{i3}$沿着数据流分割点查找方向跳跃11个字节为例进行描述。如图24所示，当判断W_3不满足预定条件C_3时，以k_i为起始点，沿着数据流分割点查找方向跳跃N个字节，其中N个字节不大于$\| B_3 \| + \max_x(\| A_x \|)$，在本实施例中，N=7，在第7个字节的结束位置，获得下一个潜在分割点，为与潜在分割点k_i区别，这里将新的潜在分割点表示为k_j，根据在去重服务器103上预设的规则，为潜在分割点k_j确定11个窗口$W_{jx}[k_j\text{-}A_x,k_j+B_x]$，分别为$W_{j1}[k_j\text{-}171,\ k_j\text{-}2]$、$W_{j2}[k_j\text{-}172,k_j\text{-}3]$、$W_{j3}[k_j\text{-}173,k_j\text{-}4]$、$W_{j4}[k_j\text{-}174,k_j\text{-}5]$、$W_{j5}[k_j\text{-}175,k_j\text{-}6]$、$W_{j6}[k_j\text{-}176,k_j\text{-}7]$、$W_{j7}[k_j\text{-}177,k_j\text{-}8]$、$W_{j8}[k_j\text{-}178,$

k_j-9]、$W_{j9}[k_j$-179,k_j-10]、$W_{j10}[k_j$-170,k_j-1]和$W_{j11}[k_j$-169,k_j]。判断W_{j1}[k_j-171, k_j-2]中至少部分数据是否满足预定条件C_1、判断$W_{j2}[k_j$-172,k_j-3]中至少部分数据是否满足预定条件C_2、判断$W_{j3}[k_j$-173,k_j-4]中至少部分数据是否满足预定条件C_3、判断$W_{j4}[k_j$-174,k_j-5]中至少部分数据是否满足预定条件C_4、判断$W_{j5}[k_j$-175,k_j-6]中至少部分数据是否满足预定条件C_5、判断$W_{j6}[k_j$-176,k_j-7]中至少部分数据是否满足预定条件C_6、判断$W_{j7}[k_j$-177,k_j-8]中至少部分数据是否满足预定条件C_7、判断$W_{j8}[k_j$-178,k_j-9]中至少部分数据是否满足预定条件C_8、判断W_{j9}[k_j-179,k_j-10]中至少部分数据是否满足预定条件C_9、判断$W_{j10}[k_j$-170,k_j-1]中至少部分数据是否满足预定条件C_{10}和判断$W_{j11}[k_j$-169,k_j]中至少部分数据是否满足预定条件C_{11}。当然在本发明实施例中，判断潜在分割点k_a是否为数据流分割点时也遵循该原则，具体实现不再描述，可以参照判断潜在分割点k_i的描述。当判断窗口W_{j1}中至少部分数据满足预定条件C_1、窗口W_{j2}中至少部分数据满足预定条件C_2、窗口W_{j3}中至少部分数据满足预定条件C_3、窗口W_{j4}中至少部分数据满足预定条件C_4、窗口W_{j5}中至少部分数据满足预定条件C_5、窗口W_{j6}中至少部分数据满足预定条件C_6、窗口W_{j7}中至少部分数据满足预定条件C_7、窗口W_{j8}中至少部分数据满足预定条件C_8、窗口W_{j9}中至少部分数据满足预定条件C_9、窗口W_{j10}中至少部分数据满足预定条件C_{10}和窗口W_{j11}中至少部分数据满足预定条件C_{11}时，则当前潜在分割点k_j为数据流分割点，k_j与k_a之间的数据构成1个数据块，同时按照与k_a相同的方式跳过最小分块大小4KB，获得下一个潜在分割点，并按照在去重服务器103上预设的规则，判断下一个潜在分割点是否为数据流分割

点。当判断潜在分割点k_j不是数据流分割点时，按照与k_i相同的方式获得下一个潜在分割点，并按照在去重服务器103上预设的规则及上述方法判断下一个潜在分割点是否为数据流分割点。当超过设定的最大数据块仍然没有找到数据流分割点时，则从最大数据块的结束位置作为强制分割点。当然该方法的实施受最大数据块长度和构成该数据流的文件的大小约束，在此不再赘述。

在图3所示的数据流分割点查找的基础上，在图25所示的实施方式中，在去重服务器103上预设有规则，所述规则为：为潜在分割点k确定11个窗口$W_x[k-A_x,k+B_x]$和窗口$W_x[k-A_x,k+B_x]$对应的预定条件C_x，其中x为1到11连续自然数，A_1=166，B_1=3；A_2=167，B_2=2；A_3=168，B_3=1；A_4=169，B_4=0；A_5=170，B_5=-1；A_6=171，B_6=-2；A_7=172，B_7=-3；A_8=173，B_8=-4；A_9=174，B_9=-5；A_{10}=175，B_{10}=-6；A_{11}=176，B_{11}=-7；并且$C_1=C_2=C_3=C_4=C_5=C_6=C_7=C_8=C_9=C_{10}=C_{11}$，则11个窗口分别为$W_1$[k-166, k+3]、$W_2$[k-167, k+2]、$W_3$[k-168, k+1]、$W_4$[k-169, k]、$W_5$[k-170, k-1]、$W_6$[k-171, k-2]、$W_7$[k-172, k-3]、$W_8$[k-173, k-4]、$W_9$[k-174, k-5]、$W_{10}$[k-175, k-6]和$W_{11}$[k-176, k-7]。$k_a$为数据流分割点，图25中所示数据流分割点查找方向为从左向右，从数据流分割点k_a跳过最小数据块4KB后，最小数据块4KB结束位置作为下一个潜在分割点k_i，在本实施例中，根据在去重服务器103上预设的规则，为潜在分割点k_i确定11个窗口W_{ix}[k-A_x,k+B_x]及窗口W_{ix}[k-A_x,k+B_x]对应的预定条件C_x，x分别为1到11连续的自然数。在图25所示的实施方式中，为潜在分割点k_i确定11个窗口，分别为$W_{i1}[k_i-166, k_i+3]$、$W_{i2}[k_i-167, k_i+2]$、$W_{i3}[k_i-168,k_i+1]$、$W_{i4}[k_i-169,k_i]$、$W_{i5}[k_i-170,k_i-1]$、$W_{i6}[k_i-171,k_i-2]$、$W_{i7}[k_i-172,k_i-3]$、$W_{i8}[k_i-173,k_i-4]$、$W_{i9}[k_i-174,k_i-5]$、$W_{i10}[k_i-175,k_i-6]$和$W_{i11}[k_i-176,k_i-7]$。判断$W_{i1}[k_i-166, k_i+3]$中至少部分数据是否满足预定条件C_1、判断W_{i2}

$[k_i\text{-}167,\ k_i+2]$中至少部分数据是否满足预定条件C_2、判断$W_{i3}[k_i\text{-}168,k_i+1]$中至少部分数据是否满足预定条件$C_3$、判断$W_{i4}[k_i\text{-}169,k_i]$中至少部分数据是否满足预定条件$C_4$、判断$W_{i5}[k_i\text{-}170,k_i\text{-}1]$中至少部分数据是否满足预定条件$C_5$、判断$W_{i6}[k_i\text{-}171,k_i\text{-}2]$中至少部分数据是否满足预定条件$C_6$、判断$W_{i7}$ $W_{i7}[k_i\text{-}172,k_i\text{-}3]$中至少部分数据是否满足预定条件$C_7$、判断$W_{i8}[k_i\text{-}173,k_i\text{-}4]$中至少部分数据是否满足预定条件$C_8$、判断$W_{i9}[k_i\text{-}174,k_i\text{-}5]$中至少部分数据是否满足预定条件$C_9$、判断$W_{i10}[k_i\text{-}175,k_i\text{-}6]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{i11}[k_i\text{-}176,k_i\text{-}7]$中至少部分数据是否满足预定条件$C_{11}$。当判断窗口$W_{i1}$中至少部分数据满足预定条件$C_1$、窗口$W_{i2}$中至少部分数据满足预定条件$C_2$、窗口$W_{i3}$中至少部分数据满足预定条件$C_3$、窗口$W_{i4}$中至少部分数据满足预定条件$C_4$、窗口$W_{i5}$中至少部分数据满足预定条件$C_5$、窗口$W_{i6}$中至少部分数据满足预定条件$C_6$、窗口$W_{i7}$中至少部分数据满足预定条件$C_7$、窗口$W_{i8}$中至少部分数据满足预定条件$C_8$、窗口$W_{i9}$中至少部分数据满足预定条件$C_9$、窗口$W_{i10}$中至少部分数据满足预定条件$C_{10}$和窗口$W_{i11}$中至少部分数据满足预定条件$C_{11}$时，则当前潜在分割点$k_i$为数据流分割点。当11个窗口中任一个窗口中至少部分数据不满足对应的预定条件时，如图26所示，$W_{i7}[k_i\text{-}172,k_i\text{-}3]$，则从潜在分割点$k_i$沿着数据流分割点查找方向跳跃N个字节，其中N个字节不大于$\| B_7 \| + \max_x(\| A_x \|)$，在图26所示的实施方式中，跳跃N个字节不大于185个字节，在本实施例中，N=5，得到新的潜在分割点，为与潜在分割点k_i区别，这里将新的潜在分割点表示为k_j，根据图25所示的实施方式中在去重服务器103上预设的规则，为潜在分割点k_j确定的窗口为11个，分别为$W_{j1}[k_j\text{-}166,k_j+3]$、$W_{j2}[k_j\text{-}167,k_j+2]$、$W_{j3}[k_j\text{-}168,k_j+1]$、

$W_{j4}[k_j\text{-}169,k_j]$、$W_{j5}[k_j\text{-}170,k_j\text{-}1]$、$W_{j6}[k_j\text{-}171,k_j\text{-}2]$、$W_{j7}[k_j\text{-}172,k_j\text{-}3]$、$W_{j8}[k_j\text{-}173,k_j\text{-}4]$、$W_{j9}[k_j\text{-}174,k_j\text{-}5]$、$W_{j10}[k_j\text{-}175,k_j\text{-}6]$和$W_{j11}[k_j\text{-}176,k_j\text{-}7]$。判断$W_{j1}[k_j\text{-}166,k_j+3]$中至少部分数据是否满足预定条件$C_1$、判断$W_{j2}[k_j\text{-}167,k_j+2]$中至少部分数据是否满足预定条件$C_2$、判断$W_{j3}[k_j\text{-}168,k_j+1]$中至少部分数据是否满足预定条件$C_3$、判断$W_{j4}[k_j\text{-}169,k_j]$中至少部分数据是否满足预定条件$C_4$、判断$W_{j5}[k_j\text{-}170,k_j\text{-}1]$中至少部分数据是否满足预定条件$C_5$、判断$W_{j6}[k_j\text{-}171,k_j\text{-}2]$中至少部分数据是否满足预定条件$C_6$、判断$W_{j7}[k_j\text{-}172,k_j\text{-}3]$中至少部分数据是否满足预定条件$C_7$、判断$W_{j8}[k_j\text{-}173,k_j\text{-}4]$中至少部分数据是否满足预定条件$C_8$、判断$W_{j9}[k_j\text{-}174,k_j\text{-}5]$中至少部分数据是否满足预定条件$C_9$、判断$W_{j10}[k_j\text{-}175,k_j\text{-}6]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{j11}[k_j\text{-}176,k_j\text{-}7]$中至少部分数据是否满足预定条件$C_{11}$。当然在本发明实施例中，判断潜在分割点$k_a$是否为数据流分割点时也遵循该原则，具体实现不再描述，可以参照判断潜在分割点k_i的描述。当判断窗口W_{j1}中至少部分数据满足预定条件C_1、窗口W_{j2}中至少部分数据满足预定条件C_2、窗口W_{j3}中至少部分数据满足预定条件C_3、窗口W_{j4}中至少部分数据满足预定条件C_4、窗口W_{j5}中至少部分数据满足预定条件C_5、窗口W_{j6}中至少部分数据满足预定条件C_6、窗口W_{j7}中至少部分数据满足预定条件C_7、窗口W_{j8}中至少部分数据满足预定条件C_8、窗口W_{j9}中至少部分数据满足预定条件C_9、窗口W_{j10}中至少部分数据满足预定条件C_{10}和窗口W_{j11}中至少部分数据满足预定条件C_{11}时，则当前潜在分割点k_j为数据流分割点，k_j与k_a之间的数据构成1个数据块，同时按照与k_a相同的方式跳过最小分块大小4KB，获得下一个潜在分割点，并按照在去

重服务器103上预设的规则，判断下一个潜在分割点是否为数据流分割点。当判断潜在分割点k_j不是数据流分割点时，按照与k_i相同的方式获得下一个潜在分割点，并按照在去重服务器103上预设的规则及上述方法判断下一个潜在分割点是否为数据流分割点。当超过设定的最大数据块仍然没有找到数据流分割点时，则从最大数据块的结束位置作为强制分割点。

在图 3 所示的数据流分割点查找的基础上，在图 27 所示的实施方式中，在去重服务器 103 上预设有规则，所述规则为：为潜在分割点k确定 11 个窗口W_x[k-A_x, k+B_x]和窗口W_x[k-A_x, k+B_x]对应的预定条件C_x，其中 x 为 1 到 11 连续的自然数，A_1=169，B_1=0；A_2=170，B_2=-1；A_3=171，B_3=-2；A_4=172，B_4=-3；A_5=173，B_5=-4；A_6=174，B_6=-5；A_7=175，B_7=-6；A_8=176，B_8=-7；A_9=177，B_9=-8；A_{10}=168，B_{10}=1；A_{11}=179，B_{11}=3；并且C_1=C_2=C_3=C_4=C_5=C_6=C_7=C_8=C_9=C_{10}≠C_{11}，则 11 个窗口分别为W_1[k-169, k]、W_2[k-170, k-1]、W_3[k-171, k-2]、W_4[k-172, k-3]、W_5[k-173, k-4]、W_6[k-174, k-5]、W_7[k-175, k-6]、W_8[k-176, k-7]、W_9[k-177, k-8]、W_{10}[k-168, k+1]和W_{11}[k-179, k+3]。k_a为数据流分割点，图 27 中所示数据流分割点查找方向为从左向右，从数据流分割点k_a跳过最小数据块 4KB 后，最小数据块 4KB 结束位置作为下一个潜在分割点k_i，在本实施例中，根据在去重服务器 103 上预设的规则，为潜在分割点k_i确定窗口W_{ix}[k_i-A_x，k_i+B_x]，x 分别为 1 到 11 连续的自然数，在图 27 所示的实施方式中，为潜在分割点k_i确定 11 个窗口分别为W_{i1}[k_i-169，k_i]、W_{i2}[k_i-170，k_i-1]、W_{i3}[k_i-171，k_i-2]、W_{i4}[k_i-172，k_i-3]、W_{i5}[k_i-173，k_i-4]、W_{i6}[k_i-174，k_i-5]、W_{i7}[k_i-175，k_i-6]、W_{i8}[k_i-176，k_i-7]、W_{i9}[k_i

-177, k_i-8]、W_{i10}[k_i-168, k_i+1]和W_{i11}[k_i-179, k_i+3]。判断W_{i1}[k_i-169, k_i]中至少部分数据是否满足预定条件C_1、判断W_{i2}[k_i-170, k_i-1]中至少部分数据是否满足预定条件C_2、判断W_{i3}[k_i-171, k_i-2]中至少部分数据是否满足预定条件C_3、判断W_{i4}[k_i-172, k_i-3]中至少部分数据是否满足预定条件C_4、判断W_{i5}[k_i-173, k_i-4]中至少部分数据是否满足预定条件C_5、判断W_{i6}[k_i-174, k_i-5]中至少部分数据是否满足预定条件C_6、判断W_{i7}[k_i-175, k_i-6]中至少部分数据是否满足预定条件C_7、判断W_{i8}[k_i-176, k_i-7]中至少部分数据是否满足预定条件C_8、判断W_{i9}[k_i-177, k_i-8]中至少部分数据是否满足预定条件C_9、判断W_{i10}[k_i-168, k_i+1]中至少部分数据是否满足预定条件C_{10}和判断W_{i11}[k_i-179, k_i+3]中至少部分数据是否满足预定条件C_{11}。当判断窗口W_{i1}中至少部分数据满足预定条件C_1、窗口W_{i2}中至少部分数据满足预定条件C_2、窗口W_{i3}中至少部分数据满足预定条件C_3、窗口W_{i4}中至少部分数据满足预定条件C_4、窗口W_{i5}中至少部分数据满足预定条件C_5、窗口W_{i6}中至少部分数据满足预定条件C_6、窗口W_{i7}中至少部分数据满足预定条件C_7、窗口W_{i8}中至少部分数据满足预定条件C_8、窗口W_{i9}中至少部分数据满足预定条件C_9、窗口W_{i10}中至少部分数据满足预定条件C_{10}和窗口W_{i11}中至少部分数据满足预定条件C_{11}时，则当前潜在分割点k_i为数据流分割点。当判断窗口W_{i11}中至少部分数据不满足预定条件C_{11}时，则从潜在分割点k_i沿着数据流分割点查找方向跳跃 1 个字节，得到新的潜在分割点，为与潜在分割点k_i区别，这里将新的潜在分割点表示为k_j。当W_{i1}、W_{i2}、W_{i3}、W_{i4}、W_{i5}、W_{i6}、W_{i7}、W_{i8}、W_{i9}和W_{i10}10 个窗口中任一个窗口中至少部分数据不满足对应的预定条件时，如图 28 所示，W_{i4}[k_i

-172, k_i-3]，则从点k_i沿着数据流分割点查找方向跳跃 N 个字节，其中 N 个字节不大于 $\| B_4 \| + \max_x(\| A_x \|)$，在图 28 所示的实施方式中，跳跃 N 个字节不大于 182 个字节，在本实施例中，N=6，得到新的潜在分割点，为与潜在分割点k_i区别，这里将新的潜在分割点表示为k_j，根据图 27 所示的实施方式中在去重服务器 103 上预设的规则，为潜在分割点k_j确定的窗口分别为W_{j1}[k_j-169, k_j]、W_{j2}[k_j-170, k_j-1]、W_{j3}[k_j-171, k_j-2]、W_{j4}[k_j-172, k_j-3]、W_{j5}[k_j-173, k_j-4]、W_{j6}[k_j-174, k_j-5]、W_{j7}[k_j-175, k_j-6]、W_{j8}[k_j-176, k_j-7]、W_{j9}[k_j-177, k_j-8]、W_{j10}[k_j-168, k_j+1]和W_{j11}[k_j-179, k_j+3]。判断W_{j1}[k_j-169, k_j]中至少部分数据是否满足预定条件C_1、判断W_{j2}[k_j-170, k_j-1]中至少部分数据是否满足预定条件C_2、判断W_{j3}[k_j-171, k_j-2]中至少部分数据是否满足预定条件C_3、判断W_{j4}[k_j-172, k_j-3]中至少部分数据是否满足预定条件C_4、判断W_{j5}[k_j-173, k_j-4]中至少部分数据是否满足预定条件C_5、判断W_{j6}[k_j-174, k_j-5]中至少部分数据是否满足预定条件C_6、判断W_{j7}[k_j-175, k_j-6]中至少部分数据是否满足预定条件C_7、判断W_{j8}[k_j-176, k_j-7]中至少部分数据是否满足预定条件C_8、判断W_{j9}[k_j-177, k_j-8]中至少部分数据是否满足预定条件C_9、判断W_{j10}[k_j-168, k_j+1]中至少部分数据是否满足预定条件C_{10}和判断W_{j11}[k_j-179, k_j+3]中至少部分数据是否满足预定条件C_{11}。当然在本发明实施例中，判断潜在分割点k_a是否为数据流分割点时也遵循该原则，具体实现不再描述，可以参照判断潜在分割点k_i的描述。当判断窗口W_{j1}中至少部分数据满足预定条件C_1、窗口W_{j2}中至少部分数据满足预定条件C_2、窗口W_{j3}中至少部分数据

满足预定条件C_3、窗口W_{j4}中至少部分数据满足预定条件C_4、窗口W_{j5}中至少部分数据满足预定条件C_5、窗口W_{j6}中至少部分数据满足预定条件C_6、窗口W_{j7}中至少部分数据满足预定条件C_7、窗口W_{j8}中至少部分数据满足预定条件C_8、窗口W_{j9}中至少部分数据满足预定条件C_9、窗口W_{j10}中至少部分数据满足预定条件C_{10}和窗口W_{j11}中至少部分数据满足预定条件C_{11}时，则当前潜在分割点k_j为数据流分割点，k_j与k_a之间的数据构成 1 个数据块，同时按照与k_a相同的方式跳过最小分块大小 4KB，获得下一个潜在分割点，并按照在去重服务器 103 上预设的规则，判断下一个潜在分割点是否为数据流分割点。当判断潜在分割点k_j不是数据流分割点时，按照与k_i相同的方式获得下一个潜在分割点，并按照在去重服务器 103 上预设的规则及上述方法判断下一个潜在分割点是否为数据流分割点。当超过设定的最大数据块仍然没有找到数据流分割点时，则从最大数据块的结束位置作为强制分割点。

在图3所示的数据流分割点查找的基础上，在图29所示的实施方式中，在去重服务器103上预设有规则，所述规则为：为潜在分割点k确定11个窗口$W_x[p_x-A_x, p_x+B_x]$和窗口$W_x[p_x-A_x, p_x+B_x]$对应的预定条件C_x，x分别为1到11连续的自然数，其中，窗口$W_x[p_x-A_x, p_x+B_x]$中至少部分数据满足预定条件的概率为1/2，A_1=169，B_1=0；A_2=171，B_2=-2；A_3=173，B_3=-4；A_4=175，B_4=-6；A_5=177，B_5=-8；A_6=179，B_6=-10；A_7=181，B_7=-12；A_8=183，B_8=-14；A_9=185，B_9=-16；A_{10}=187，B_{10}=-18；A_{11}=189，B_{11}=-20；并且$C_1=C_2=C_3=C_4=C_5=C_6=C_7=C_8=C_9=C_{10}=C_{11}$，则11个窗口分别为$W_1$[k-169, k]、$W_2$[k-171, k-2]、$W_3$[k-173, k-4]、$W_4$[k-175, k-6]、$W_5$[k-177, k-8]、$W_6$[k-179, k-10]、$W_7$[k-181, k-12]、$W_8$[k-183, k-14]、$W_9$[k-185, k-16]、

W_{10}[k-187, k-18]和W_{11}[k-189, k-20]。k_a为数据流分割点，图29中所示数据流分割点查找方向为从左向右，从数据流分割点k_a跳过最小数据块4KB后，在最小数据块4KB结束位置作为下一个潜在分割点k_i，为潜在分割点k_i确定点p_{ix}，在本实施例中，根据在去重服务器103上预设的规则，x分别为1到11连续的自然数。在图29所示的实施方式中，依据预定规则，为潜在分割点k_i确定的11个窗口分别为W_{i1}[k_i-169, k_i]、W_{i2}[k_i-171, k_i-2]、W_{i3}[k_i-173, k_i-4]、W_{i4}[k_i-175, k_i-6]、W_{i5}[k_i-177, k_i-8]、W_{i6}[k_i-179, k_i-10]、W_{i7}[k_i-181, k_i-12]、W_{i8}[k_i-183, k_i-14]、W_{i9}[k_i-185, k_i-16]、W_{i10}[k_i-187, k_i-18]和W_{i11}[k_i-189, k_i-20]。判断W_{i1}[k_i-169, k_i]中至少部分数据是否满足预定条件C_1、判断W_{i2}[k_i-171, k_i-2]中至少部分数据是否满足预定条件C_2、判断W_{i3}[k_i-173, k_i-4]中至少部分数据是否满足预定条件C_3、判断W_{i4}[k_i-175, k_i-6]中至少部分数据是否满足预定条件C_4、判断W_{i5}[k_i-177, k_i-8]中至少部分数据是否满足预定条件C_5、判断W_{i6}[k_i-179, k_i-10]中至少部分数据是否满足预定条件C_6、判断W_{i7}[k_i-181, k_i-12]中至少部分数据是否满足预定条件C_7、判断W_{i8}[k_i-183, k_i-14]中至少部分数据是否满足预定条件C_8、判断W_{i9}[k_i-185, k_i-16]中至少部分数据是否满足预定条件C_9、判断W_{i10}[k_i-187, k_i-18]中至少部分数据是否满足预定条件C_{10}和判断W_{i11}[k_i-189, k_i-20]中至少部分数据是否满足预定条件C_{11}。当判断窗口W_{i1}中至少部分数据满足预定条件C_1、窗口W_{i2}中至少部分数据满足预定条件C_2、窗口W_{i3}中至少部分数据满足预定条件C_3、窗口W_{i4}中至少部分数据满足预定条件C_4、窗口W_{i5}中至少部分数据满足预定条件C_5、窗口W_{i6}中至少部分数据满足预定条件C_6、窗口W_{i7}中至少部分数据满足预定条件C_7、窗口W_{i8}中至少部分数据满足预定条件C_8、窗口W_{i9}中至

少部分数据满足预定条件C_9、窗口W_{i10}中至少部分数据满足预定条件C_{10}和窗口W_{i11}中至少部分数据满足预定条件C_{11}时，则当前潜在分割点k_i为数据流分割点。当11个窗口中任一个窗口中至少部分数据不满足对应的预定条件时，如图30所示，$W_{i4}[k_i\text{-}175, k_i\text{-}6]$中至少部分数据不满足预定条件$C_4$，则选择下一个潜在分割点，为与潜在分割点$k_i$区别，这里表示为$k_j$，$k_j$位于$k_i$右边，并且$k_j$与$k_i$间距1个字节。如图30所示，依为去重服务器103预设的规则，为潜在分割点k_j确定11个窗口分别为$W_{j1}[k_j\text{-}169, k_j]$、$W_{j2}[k_j\text{-}171, k_j\text{-}2]$、$W_{j3}[k_j\text{-}173, k_j\text{-}4]$、$W_{j4}[k_j\text{-}175, k_j\text{-}6]$、$W_{j5}[k_j\text{-}177, k_j\text{-}8]$、$W_{j6}[k_j\text{-}179, k_j\text{-}10]$、$W_{j7}[k_j\text{-}181, k_j\text{-}12]$、$W_{j8}[k_j\text{-}183, k_j\text{-}14]$、$W_{j9}[k_j\text{-}185, k_j\text{-}16]$、$W_{j10}[k_j\text{-}187, k_j\text{-}18]$和$W_{j11}[k_j\text{-}189, k_j\text{-}20]$，并且$C_1=C_2=C_3=C_4=C_5=C_6=C_7=C_8=C_9=C_{10}=C_{11}$。判断$W_{j1}[k_j\text{-}169, k_j]$中至少部分数据是否满足预定条件$C_1$、判断$W_{j2}[k_j\text{-}171, k_j\text{-}2]$中至少部分数据是否满足预定条件$C_2$、判断$W_{j3}[k_j\text{-}173, k_j\text{-}4]$中至少部分数据是否满足预定条件$C_3$、判断$W_{j4}[k_j\text{-}175, k_j\text{-}6]$中至少部分数据是否满足预定条件$C_4$、判断$W_{j5}[k_j\text{-}177, k_j\text{-}8]$中至少部分数据是否满足预定条件$C_5$、判断$W_{j6}[k_j\text{-}179, k_j\text{-}10]$中至少部分数据是否满足预定条件$C_6$、判断$W_{j7}[k_j\text{-}181, k_j\text{-}12]$中至少部分数据是否满足预定条件$C_7$、判断$W_{j8}[k_j\text{-}183, k_j\text{-}14]$中至少部分数据是否满足预定条件$C_8$、判断$W_{j9}[k_j\text{-}185, k_j\text{-}16]$中至少部分数据是否满足预定条件$C_9$、判断$W_{j10}[k_j\text{-}187, k_j\text{-}18]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{j11}[k_j\text{-}189, k_j\text{-}20]$中至少部分数据是否满足预定条件$C_{11}$。当判断窗口$W_{j1}$中至少部分数据满足预定条件$C_1$、窗口$W_{j2}$中至少部分数据满足预定条件$C_2$、窗口$W_{j3}$中至少部分数据满足预定条件$C_3$、窗口$W_{j4}$中至

少部分数据满足预定条件C_4、窗口W_{j5}中至少部分数据满足预定条件C_5、窗口W_{j6}中至少部分数据满足预定条件C_6、窗口W_{j7}中至少部分数据满足预定条件C_7、窗口W_{j8}中至少部分数据满足预定条件C_8、窗口W_{i9}中至少部分数据满足预定条件C_9、窗口W_{j10}中至少部分数据满足预定条件C_{10}和窗口W_{j11}中至少部分数据满足预定条件C_{11}时，则当前潜在分割点k_j为数据流分割点。当判断窗口W_{j1}、W_{j2}、W_{j3}、W_{j4}、W_{j5}、W_{j6}、W_{j7}、W_{j8}、W_{j9}、W_{j10}和W_{j11}中任一个窗口中至少部分数据不满足预定条件时，如图31所示，$W_{j3}[k_j\text{-}173, k_j\text{-}4]$中至少部分数据不满足预定条件$C_3$时，$k_j$位于$k_i$右边从$k_i$沿着数据流分割点查找方向跳跃N个字节，其中N个字节不大于$\| B_4 \| + \max_x(\| A_x \|)$，在图28所示的实施方式中，N个字节不大于195个字节，在本实施例中，N=15，获得下一个潜在分割点，为与潜在分割点k_i、k_j相区别，表示为k_l。根据图29所实施方式中为去重服务器103预设的规则，为潜在分割点k_l确定11个窗口分别为$W_{l1}[k_l\text{-}169, k_l]$、$W_{l2}[k_l\text{-}171, k_l\text{-}2]$、$W_{l3}[k_l\text{-}173, k_l\text{-}4]$、$W_{l4}[k_l\text{-}175, k_l\text{-}6]$、$W_{l5}[k_l\text{-}177, k_l\text{-}8]$、$W_{l6}[k_l\text{-}179, k_l\text{-}10]$、$W_{l7}[k_l\text{-}181, k_l\text{-}12]$、$W_{l8}[k_l\text{-}183, k_l\text{-}14]$、$W_{l9}[k_l\text{-}185, k_l\text{-}16]$、$W_{l10}[k_l\text{-}187, k_l\text{-}18]$和$W_{l11}[k_l\text{-}189, k_l\text{-}20]$。判断$W_{l1}[k_l\text{-}169, k_l]$中至少部分数据是否满足预定条件$C_1$、判断$W_{l2}[k_l\text{-}171, k_l\text{-}2]$中至少部分数据是否满足预定条件$C_2$、判断$W_{l3}[k_l\text{-}173, k_l\text{-}4]$中至少部分数据是否满足预定条件$C_3$、判断$W_{l4}[k_l\text{-}175, k_l\text{-}6]$中至少部分数据是否满足预定条件$C_4$、判断$W_{l5}[k_l\text{-}177, k_l\text{-}8]$中至少部分数据是否满足预定条件$C_5$、判断$W_{l6}[k_l\text{-}179, k_l\text{-}10]$中至少部分数据是否满足预定条件$C_6$、判断$W_{l7}[k_l\text{-}181, k_l\text{-}12]$中至少部分数据是否满足预定条件$C_7$、判断$W_{l8}[k_l\text{-}183, k_l\text{-}14]$中至少部分数据是否满足预定条件$C_8$、判断$W_{l9}[k_l\text{-}185, k_l\text{-}16]$中至少

部分数据是否满足预定条件C_9、判断$W_{l10}[k_l\text{-}187,\ k_l\text{-}18]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{l11}[k_l\text{-}189,\ k_l\text{-}20]$中至少部分数据是否满足预定条件$C_{11}$。当判断窗口$W_{l1}$中至少部分数据满足预定条件$C_1$、窗口$W_{l2}$中至少部分数据满足预定条件$C_2$、窗口$W_{l3}$中至少部分数据满足预定条件$C_3$、窗口$W_{l4}$中至少部分数据满足预定条件$C_4$、窗口$W_{l5}$中至少部分数据满足预定条件$C_5$、窗口$W_{l6}$中至少部分数据满足预定条件$C_6$、窗口$W_{l7}$中至少部分数据满足预定条件$C_7$、窗口$W_{l8}$中至少部分数据满足预定条件$C_8$、窗口$W_{l9}$中至少部分数据满足预定条件$C_9$、窗口$W_{l10}$中至少部分数据满足预定条件$C_{10}$和窗口$W_{l11}$中至少部分数据满足预定条件$C_{11}$时，则当前潜在分割点$k_l$为数据流分割点。当窗口$W_{l1}$、$W_{l2}$、$W_{l3}$、$W_{l4}$、$W_{l5}$、$W_{l6}$、$W_{l7}$、$W_{l8}$、$W_{l9}$、$W_{l10}$和$W_{l11}$中任一窗口中至少部分数据不满足预定条件时，选择下一个潜在分割点，为与潜在分割点k_i、k_j和k_l区别，表示为k_m，k_m位于k_l右边，并且k_m与k_l间距1个字节。根据图29所示实施例为去重服务器103预设的规则，为潜在分割点k_m确定的11个窗口分别为$W_{m1}[k_m\text{-}169,\ k_m]$、$W_{m2}[k_m\text{-}171,\ k_m\text{-}2]$、$W_{m3}[k_m\text{-}173,\ k_m\text{-}4]$、$W_{m4}[k_m\text{-}175,\ k_m\text{-}6]$、$W_{m5}[k_m\text{-}177,\ k_m\text{-}8]$、$W_{m6}[k_m\text{-}179,\ k_m\text{-}10]$、$W_{m7}[k_m\text{-}181,\ k_m\text{-}12]$、$W_{m8}[k_m\text{-}183,\ k_m\text{-}14]$、$W_{m9}[k_m\text{-}185,\ k_m\text{-}16]$、$W_{m10}[k_m\text{-}187,\ k_m\text{-}18]$和$W_{m11}[k_m\text{-}189,\ k_m\text{-}20]$。判断$W_{m1}[k_m\text{-}169,\ k_m]$中至少部分数据是否满足预定条件$C_1$、判断$W_{m2}[k_m\text{-}171,\ k_m\text{-}2]$中至少部分数据是否满足预定条件$C_2$、判断$W_{m3}[k_m\text{-}173,\ k_m\text{-}4]$中至少部分数据是否满足预定条件$C_3$、判断$W_{m4}[k_m\text{-}175,\ k_m\text{-}6]$中至少部分数据是否满足预定条件$C_4$、判断$W_{m5}[k_m\text{-}177,\ k_m\text{-}8]$中至少部分数据是否满足预定条件$C_5$、判断$W_{m6}[k_m\text{-}179,\ k_m\text{-}10]$中至少部分数据是否满足预定条件$C_6$、判断$W_{m7}[k_m\text{-}181,\ k_m\text{-}12]$中至少部分数据是否满足预定条

件C_7、判断$W_{m8}[k_m\text{-}183,\ k_m\text{-}14]$中至少部分数据是否满足预定条件$C_8$、判断$W_{m9}[k_m\text{-}185,\ k_m\text{-}16]$中至少部分数据是否满足预定条件$C_9$、判断$W_{m10}[k_m\text{-}187,\ k_m\text{-}18]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{m11}[k_m\text{-}189,\ k_m\text{-}20]$中至少部分数据是否满足预定条件$C_{11}$。当判断窗口$W_{m1}$中至少部分数据满足预定条件$C_1$、窗口$W_{m2}$中至少部分数据满足预定条件$C_2$、窗口$W_{m3}$中至少部分数据满足预定条件$C_3$、窗口$W_{m4}$中至少部分数据满足预定条件$C_4$、窗口$W_{m5}$中至少部分数据满足预定条件$C_5$、窗口$W_{m6}$中至少部分数据满足预定条件$C_6$、窗口$W_{m7}$中至少部分数据满足预定条件$C_7$、窗口$W_{m8}$中至少部分数据满足预定条件$C_8$、窗口$W_{m9}$中至少部分数据满足预定条件$C_9$、窗口$W_{m10}$中至少部分数据满足预定条件$C_{10}$和窗口$W_{m11}$中至少部分数据满足预定条件$C_{11}$时，则当前潜在分割点$k_m$为数据流分割点。当任一个窗口中至少部分数据不满足预定条件时，则按照前面描述的方案执行跳跃，以获得下一个潜在分割点并判断是否为数据流分割点。

本发明实施例提供了一种判断窗口$W_{iz}[k_i-A_z,\ k_i+B_z]$中至少部分数据是否满足预定条件$C_z$的方法，本实施例中使用随机函数判断窗口$W_{iz}[k_i-A_z,\ k_i+B_z]$中至少部分数据是否满足预定条件$C_z$，以图 21 所示的实施方式为例，根据在去重服务器 103 上预设的规则，为潜在分割点k_i确定窗口$W_{i1}[k_i\text{-}169,\ k_i]$，判断$W_{i1}[k_i\text{-}169,\ k_i]$中至少部分数据是否满足预定的条件$C_1$，如图 32 所示，$W_{i1}$表示窗口$W_{i1}[k_i\text{-}169,\ k_i]$，为判断$W_{i1}[k_i\text{-}169,\ k_i]$中至少部分数据是否满足预定条件$C_1$，选择 5 个字节，图 32 中“■”表示选择的 1 个字节，相邻两个选择的字节之间相差 42 个字节。将选择的 5 字节数据反复利用 51 次，共获得 255 字节，以增加随机性。其中每个字节由

8位组成，记为$a_{m,1}\ldots a_{m,8}$,表示255个字节中第m个字节的第1到第8位，因此，255个字节对应的位可以表示为：$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,8} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ a_{255,1} & a_{255,2} & \cdots & a_{255,8} \end{pmatrix}$，当$a_{m,n}$=1时，$V_{am,n}$ =1，当$a_{m,n}$=0时，$V_{am,n}$=-1，其中$a_{m,n}$表示$a_{m,1}\ldots a_{m,8}$中的任一个，255个字节对应的位按照$a_{m,n}$与$V_{am,n}$的转换关系得到矩阵V_a，可以表示为：$\begin{pmatrix} V_{a1,1} & V_{a1,2} & \cdots & V_{a1,8} \\ V_{a2,1} & V_{a2,2} & \cdots & V_{a2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{a255,1} & V_{a255,2} & \cdots & V_{a255,8} \end{pmatrix}$。选取大量随机数，组成矩阵，由随机数据组成的矩阵一旦组成，保持不变，如从服从特定分布（这里以正态分布为例）的随机数中选择255*8个随机数组成矩阵R:$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,8} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{255,1} & h_{255,2} & \cdots & h_{255,8} \end{pmatrix}$,将矩阵$V_a$的第m行与矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为$S_{am}=V_{am,1}*h_{m,1}+V_{am,2}*h_{m,2}+\ldots+V_{am,8}*h_{m,8}$。根据该方法，获得$S_{a1}$、$S_{a2}$...到$S_{a255}$，统计$S_{a1}$、$S_{a2}$...到$S_{a255}$中满足特定条件（这里以大于0为例）的值的个数K。由于矩阵R服从正态分布，则S_{am}与矩阵R一样，仍然服从正态分布，根据概率论，正态分布随机数大于0的概率为1/2，在S_{a1}、S_{a2}...到S_{a255}中，每个值大于0的概率为1/2，所以K满足二项分布：$P(k=n)=C_{255}^{n}(\frac{1}{2})^{n}(\frac{1}{2})^{255-n}=C_{255}^{n}(\frac{1}{2})^{255}$。根据统计结果，判断$S_{a1}$、$S_{a2}$...到$S_{a255}$的值大于0的个数K是否为偶数，二项分布的随机数为偶数的概率为1/2，所以K以1/2的概率满足条件。当K为偶数时，

表明$W_{i1}[k_i\text{-}169,\ k_i]$中至少部分数据满足预定条件$C_1$；当 K 为奇数时，表明$W_1[k_i\text{-}169,\ k_i]$中至少部分数据不满足预定条件$C_1$,这里$C_1$即指根据上述方式获得的$S_{a1}$、$S_{a2}$...到$S_{a255}$的值大于 0 的个数 K 为偶数。在图 21 所示的实施方式中，在$W_{i1}[k_i\text{-}169,k_i]$、$W_{i2}[k_i\text{-}170,k_i\text{-}1]$、$W_{i3}[k_i\text{-}171,k_i\text{-}2]$、$W_{i4}[k_i\text{-}172,k_i\text{-}3]$、$W_{i5}[k_i\text{-}173,k_i\text{-}4]$、$W_{i6}[k_i\text{-}174,k_i\text{-}5]$、$W_{i7}[k_i\text{-}175,k_i\text{-}6]$、$W_{i8}[k_i\text{-}176,k_i\text{-}7]$、$W_{i9}[k_i\text{-}177,k_i\text{-}8]$、$W_{i10}[k_i\text{-}178,k_i\text{-}9]$和$W_{i11}[k_i\text{-}179,k_i\text{-}10]$中，各窗口大小相同，即窗口大小均为 169 字节，同时判断窗口中至少部分数据是否满足预定条件的方式也相同,具体见上述判断$W_{i1}[k_i\text{-}169,\ k_i]$中至少部分数据是否满足预定条件$C_1$的描述。因此，如图 32 所示，“▨”表示判断窗口$W_{i2}[k_i\text{-}170,\ k_i\text{-}1]$中至少部分数据是否满足预定条件$C_2$时选择的 1 个字节，相邻两个选择的字节之间相差 42 个字节。将选择的 5 字节数据反复利用 51 次，共获得 255 字节，以增加随机性。其中每个字节由 8 位组成，记为$b_{m,1}...b_{m,8}$,表示 255 个字节中第 m 个字节的第 1 到第 8 位，因此，255 个字节对应的位可以表示为：$\begin{pmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,8} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ b_{255,1} & b_{255,2} & \cdots & b_{255,8} \end{pmatrix}$，当$b_{m,n}$=1 时，$V_{bm,n}$=1，当$b_{m,n}$=0 时，$V_{bm,n}$=-1，其中$b_{m,n}$表示$b_{m,1}...b_{m,8}$中的任一个，255 个字节对应的位按照$b_{m,n}$与$V_{bm,n}$的转换关系得到矩阵$V_b$，可以表示为：$\begin{pmatrix} V_{b1,1} & V_{b1,2} & \cdots & V_{b1,8} \\ V_{b2,1} & V_{b2,2} & \cdots & V_{b2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{b255,1} & V_{b255,2} & \cdots & V_{b255,8} \end{pmatrix}$。

判断$W_{i1}[k_i\text{-}169,\ k_i]$中至少部分数据是否满足预定条件的方式与判断窗口$W_{i2}[k_i\text{-}170,\ k_i\text{-}1]$中至少部分数据是否满足预定条件的方式

相同，因此使用矩阵 R: $\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,8} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{255,1} & h_{255,2} & \cdots & h_{255,8} \end{pmatrix}$,将矩阵 V_b 的第 m 行与矩阵 R 的第 m 行的随机数相乘，然后求和得到一个值，具体表示为 $S_{bm}=V_{bm,1}*h_{m,1}+V_{bm,2}*h_{m,2}+\ldots+V_{bm,8}*h_{m,8}$。根据该方法，获得 S_{b1}、S_{b2}...到 S_{b255}，统计 S_{b1}、S_{b2}...到 S_{b255} 中满足特定条件（这里以大于 0 为例）的值的个数 K。由于矩阵 R 服从正态分布，则 S_{bm} 与矩阵 R 一样，仍然服从正态分布，根据概率论，正态分布随机数大于 0 的概率为 1/2，在 S_{b1}、S_{b2}...到 S_{b255} 中，每个值大于 0 的概率为 1/2，所以 K 满足二项分布: $P(k=n)=C_{255}^{n}(\frac{1}{2})^{n}(\frac{1}{2})^{255-n}=C_{255}^{n}(\frac{1}{2})^{255}$。根据统计结果，判断 S_{b1}、S_{b2}...到 S_{b255} 的值大于 0 的个数 K 是否为偶数，二项分布的随机数为偶数的概率为为 1/2，所以 K 以 1/2 的概率满足条件。当 K 为偶数时，表明 $W_{i2}[k_i\text{-}170,\ k_i\text{-}1]$中至少部分数据满足预定条件 C_2；当 K 为奇数时，表明 $W_{i2}[k_i\text{-}170,\ k_i\text{-}1]$中至少部分数据不满足预定条件 C_2,这里 C_2 即指根据上述方式获得的 S_{b1}、S_{b2}...到 S_{b255} 的值大于 0 的个数 K 为偶数。图 21 所示的实施方式中，$W_{i2}[k_i\text{-}170,\ k_i\text{-}1]$中至少部分数据满足预定条件 C_2。

因此，如图32所示，“▨”表示判断窗口 $W_{i3}[k_i\text{-}171,k_i\text{-}2]$中至少部分数据是否满足预定条件 C_3 时选择的1个字节，相邻两个选择的字节之间相差42个字节。将选择的5字节数据反复利用51次，共获得255字节，以增加随机性。然后使用判断窗口 $W_{i1}[k_i\text{-}169,k_i]$和 $W_{i2}[k_i\text{-}170,k_i\text{-}1]$中至少部分数据是否满足预定条件的方法，判断 $W_{i3}[k_i\text{-}171,k_i\text{-}2]$中至少数据是否满足预定条件 C_3。图21所示的实施方式中，$W_{i3}[k_i\text{-}171,k_i\text{-}2]$中至少部分数据满足预定条件。如图32所示，“▨” 表示判断窗

口 $W_{i4}[k_i\text{-}172,k_i\text{-}3]$中至少部分数据是否满足预定条件$C_4$时选择的1个字节，相邻两个选择的字节之间相差42个字节。将选择的5字节数据反复利用51次，共获得255字节，以增加随机性。然后使用判断窗口 $W_{i1}[k_i\text{-}169,k_i]$、$W_{i2}[k_i\text{-}170,k_i\text{-}1]$和 $W_{i3}[k_i\text{-}171,k_i\text{-}2]$中至少部分数据是否满足预定条件的方法，判断 $W_{i4}[k_i\text{-}172,k_i\text{-}3]$中至少部分数据是否满足预定条件$C_4$。图21所示的实施方式中，$W_{i4}[k_i\text{-}172,k_i\text{-}3]$中至少部分数据满足预定条件$C_4$。如图32所示，“▦”表示判断窗口 $W_{i5}[k_i\text{-}173,k_i\text{-}4]$中至少部分数据是否满足预定条件$C_5$时选择的1个字节，相邻两个选择的字节之间相差42个字节。将选择的5字节数据反复利用51次，共获得255字节，以增加随机性。然后使用判断窗口 $W_{i1}[k_i\text{-}169,k_i]$、$W_{i2}[k_i\text{-}170,k_i\text{-}1]$、$W_{i3}[k_i\text{-}171,k_i\text{-}2]$和 $W_{i4}[k_i\text{-}172,k_i\text{-}3]$中至少部分数据是否满足预定条件的方法，判断 $W_{i5}[k_i\text{-}173,k_i\text{-}4]$中至少数据是否满足预定条件$C_5$。图21所示的实施方式中，$W_{i5}[k_i\text{-}173,k_i\text{-}4]$中至少部分数据不满足预定条件$C_5$。

当 $W_{i5}[k_i\text{-}173,k_i\text{-}4]$中至少部分数据不满足预定条件时$C_5$，从点 p_{i5} 沿着数据流分割点查找方向跳跃7个字节，在第7个字节的结束位置获得下一个潜在分割点 k_j，如图22所示，根据为去重服务器103预设的规则，为潜在分割点 k_j 确定窗口 $W_{j1}[k_j\text{-}169,k_j]$，判断窗口 $W_{j1}[k_j\text{-}169,k_j]$中至少部分数据是否满足预定条件$C_1$的方式与判断窗口 $W_{i1}[k_i\text{-}169,k_i]$中至少部分数据是否满足预定条件$C_1$的方式相同，因此如图33所示，$W_{j1}$表示窗口，为判断中至少部分数据是否满足预定条件$C_1$，选择5个字节，图33中“■”表示选择的1个字节，相邻两个选择的字节之间相差42个字节。将选择的5字节数据反复利用51次，共获得255字节，以增加随机性。其中每个字节由8位组成，记为 $a_{m,1}'\ldots a_{m,8}'$,表

示255个字节中第m个字节的第1到第8位，因此，255个字节对应的位可以表示为：$\begin{pmatrix} a_{1,1}' & a_{1,2}' & \cdots & a_{1,8}' \\ a_{2,1}' & a_{2,2}' & \cdots & a_{2,8}' \\ \vdots & \vdots & \vdots & \vdots \\ a_{255,1}' & a_{255,2}' & \cdots & a_{255,8}' \end{pmatrix}$，当 $a_{m,n}'$=1时，$V_{am,n}'$ =1，当 $a_{m,n}'$ =0时，$V_{am,n}'$=-1，其中 $a_{m,n}'$表示 $a_{m,1}'\ldots a_{m,8}'$中的任一个，255个字节对应的位按照 $a_{m,n}'$与 $V_{am,n}'$的转换关系得到矩阵 V_a'，可以表示为：$\begin{pmatrix} V_{a1,1}' & V_{a1,2}' & \cdots & V_{a1,8}' \\ V_{a2,1}' & V_{a2,2}' & \cdots & V_{a2,8}' \\ \vdots & \vdots & \vdots & \vdots \\ V_{a255,1}' & V_{a255,2}' & \cdots & V_{a255,8}' \end{pmatrix}$。判断窗口中至少部分数据是否满足预定的条件与判断窗口 $W_{i1}[k_i\text{-}169,k_i]$中至少部分数据是否满足预定的条件的方式相同，因此使用矩阵R：$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,8} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{255,1} & h_{255,2} & \cdots & h_{255,8} \end{pmatrix}$,将矩阵 V_a'的第m行与矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为 $S_{am}'=V_{am,1}'*h_{m,1}+V_{am,2}'*h_{m,2}+\ldots+V_{am,8}'*h_{m,8}$。根据该方法，获得 S_{a1}'、S_{a2}'...到 S_{a255}'，统计 S_{a1}'、 S_{a2}'...到 S_{a255}'中满足特定条件（这里以大于0为例）的值的个数K。由于矩阵R服从正态分布，则 S_{am}'与矩阵R一样，仍然服从正态分布，根据概率论，正态分布随机数大于0的概率为1/2，在 S_{a1}'、 S_{a2}'...到 S_{a255}'中，每个值大于0的概率为1/2，所以K满足二项分布：$P(k=n)=C_{255}^{n}(\frac{1}{2})^{n}(\frac{1}{2})^{255-n}=C_{255}^{n}(\frac{1}{2})^{255}$。根据统计结果，判断 S_{a1}'、 S_{a2}'...到 S_{a255}'的值大于0的个数K是否为偶数，二项分布的随机数为偶数的概率为1/2，所以K以1/2的概率满足条件。当K为偶数时，表明 $W_{j1}[k_j\text{-}169,k_j]$中至少部分数据满足预定条件 C_1；当K为奇数时，表明 $W_{j1}[k_j\text{-}169,k_j]$中至少部分数据不满足预定条件 C_1。

判断$W_{i2}[k_i\text{-}170,k_i\text{-}1]$中至少部分数据是否满足预定条件$C_2$的方式和判断$W_{j2}[k_j\text{-}170,k_j\text{-}1]$中至少部分数据是否满足预定条件$C_2$的方式相同，因此，如图33所示，“▨”表示判断窗口$W_{j2}[k_j\text{-}170,k_j\text{-}1]$中至少部分数据是否满足预定条件$C_2$时选择的1个字节，相邻两个选择的字节之间相差42个字节。将选择的5字节数据反复利用51次，共获得255字节，以增加随机性。其中每个字节由8位组成，记为$b_{m,1}'\ldots b_{m,8}'$，表示255个字节中第m个字节的第1到第8位，因此，255个字节对应的位可以表示为：$\begin{pmatrix} b_{1,1}' & b_{1,2}' & \cdots & b_{1,8}' \\ b_{2,1}' & b_{2,2}' & \cdots & b_{2,8}' \\ \vdots & \vdots & \vdots & \vdots \\ b_{255,1}' & b_{255,2}' & \cdots & b_{255,8}' \end{pmatrix}$，当$b_{m,n}'$=1时，$V_{bm,n}'$=1，当$b_{m,n}'$=0时，$V_{bm,n}'$=-1，其中$b_{m,n}'$表示$b_{m,1}'\ldots b_{m,8}'$中的任一个，255个字节对应的位按照$b_{m,n}'$与$V_{bm,n}'$的转换关系得到矩阵$V_b'$，可以表示为：

$\begin{pmatrix} V_{b1,1} & V_{b1,2} & \cdots & V_{b1,8} \\ V_{b2,1} & V_{b2,2} & \cdots & V_{b2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{b255,1} & V_{b255,2} & \cdots & V_{b255,8} \end{pmatrix}$。判断窗口$W_{i2}[k_i\text{-}170,k_i\text{-}1]$中至少部分数据是否满足预定条件$C_1$和$W_{j2}[k_j\text{-}170,k_j\text{-}1]$中至少部分数据是否满足预定条件$C_1$的方式相同，因此仍使用矩阵R：$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,8} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{255,1} & h_{255,2} & \cdots & h_{255,8} \end{pmatrix}$，将矩阵$V_b'$的第m行与矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为$S_{bm}'=V_{bm,1}'*h_{m,1}+V_{bm,2}'*h_{m,2}+\ldots+V_{bm,8}'*h_{m,8}$。根据该方法，获得$S_{b1}'$、$S_{b2}'$...到$S_{b255}'$，统计$S_{b1}'$、$S_{b2}'$...到$S_{b255}'$中满足特定条件（这里以大于0为例）的值的个数K。由于矩阵R服从正态分布，则S_{bm}'与矩阵R一样，仍然服从正态分布，根据概率论，正态分布随机数大于0

的概率为1/2，在S_{b1}'、S_{b2}'...到S_{b255}'中，每个值大于0的概率为1/2，所以K满足二项分布：$P(k=n)=C_{255}^{n}(\frac{1}{2})^{n}(\frac{1}{2})^{255-n}=C_{255}^{n}(\frac{1}{2})^{255}$。根据统计结果，判断$S_{b1}$'、$S_{b2}$'...到$S_{b255}$'的值大于0的个数K是否为偶数，二项分布的随机数为偶数的概率为为1/2，所以K以1/2的概率满足条件。当K为偶数时，表明中至少部分数据满足预定条件C_2；当K为奇数时，表明W_{j2}[k_j-170,k_j-1]中至少部分数据不满足预定条件C_2。同理，判断W_{i3}[k_i-171,k_i-2]中至少部分数据是否满足预定条件C_3的方式与判断W_{j3}[k_j-171,k_j-2]中至少部分数据是否满足预定条件C_3的方式相同，同理，判断W_{j4}[k_j-172,k_j-3]中至少部分数据是否满足预定条件C_4、判断W_{j5}[k_j-173,k_j-4]中至少部分数据是否满足预定条件C_5、判断W_{j6}[k_j-174,k_j-5]中至少部分数据是否满足预定条件C_6、判断W_{j7}[k_j-175,k_j-6]中至少部分数据是否满足预定条件C_7、判断W_{j8}[k_j-176,k_j-7]中至少部分数据是否满足预定条件C_8、判断W_{j9}[k_j-177,k_j-8]中至少部分数据是否满足预定条件C_9、判断W_{j10}[k_j-178,k_j-9]中至少部分数据是否满足预定条件C_{10}和判断W_{j11}[k_j-179,k_j-10]中至少部分数据是否满足预定条件C_{11}，在此不再赘述。

本实施例中使用随机函数判断窗口W_{iz}[k_i-A_z，k_i+B_z]中至少部分数据是否满足预定条件C_z，仍然以图21所示实施方式为例，根据在去重服务器103上预设的规则，为潜在分割点k_i确定窗口W_{i1}[k_i-169,k_i]，判断W_{i1}[k_i-169，k_i]中至少部分数据是否满足预定的条件C_1，如图32所示，W_{i1}表示窗口W_{i1}[k_i-169，k_i]，为判断W_{i1}[k_i-169，k_i]中至少部分数据是否满足预定条件C_1，选择5个字节，图32中“■”表示选择的1个字节，相邻两个选择“■”的字节之间相差42个字节。其中一

种实现方式为使用HASH函数计算选择的5个字节，使用HASH函数计算得到的数值是一个固定均匀分布，如果使用HASH函数计算得到的数值为偶数，则判断$W_{i1}[k_i\text{-}169,\ k_i]$中至少部分数据满足预定条件$C_1$，即$C_1$表示根据上述方式使用HASH函数计算得到的数值为偶数。因此，$W_{i1}[k_i\text{-}169,\ k_i]$中至少部分数据是否满足预定条件的概率为1/2。在图21所示的实施方式中，使用Hash函数判断$W_{i2}[k_i\text{-}170,k_i\text{-}1]$中至少部分数据是否满足预定条件$C_2$、$W_{i3}[k_i\text{-}171,k_i\text{-}2]$中至少部分数据是否满足预定条件$C_3$、$W_{i4}[k_i\text{-}172,k_i\text{-}3]$中至少部分数据是否满足预定条件$C_4$和$W_{i5}[k_i\text{-}173,k_i\text{-}4]$中至少部分数据是否满足预定条件$C_5$，具体实现可参考描述图21所示实施方式使用Hash函数判断$W_{i1}[k_i\text{-}169,k_i]$中至少部分数据是否满足预定条件的方式$C_1$，在此不再赘述。

当$W_{i5}[k_i\text{-}173,k_i\text{-}4]$中至少部分数据不满足预定条件$C_5$时，从潜在分割点$k_i$沿着数据流分割点查找方向跳跃7个字节，在第7个字节的结束位置获得当前潜在分割点k_j，如图22所示，根据为去重服务器103预设的规则，为潜在分割点k_j确定窗口$W_{j1}[k_j\text{-}169,k_j]$，判断窗口$W_{j1}[k_j\text{-}169,k_j]$中至少部分数据是否满足预定条件$C_1$的方式与判断窗口$W_{i1}[k_i\text{-}169,k_i]$中至少部分数据是否满足预定条件$C_1$的方式相同，因此如图33所示，$W_{j1}$表示窗口$W_{j1}[k_j\text{-}169,k_j]$，为判断$W_{j1}[k_j\text{-}169,k_j]$中至少部分数据是否满足预定条件$C_1$，选择5个字节，图33中“■”表示选择的1个字节，相邻两个选择的字节“■”之间相差42个字节。使用Hash函数计算从窗口$W_{j1}[k_j\text{-}169,k_j]$中选取的5个字节，如果得到的数值为偶数，则$W_{j1}[k_j\text{-}169,k_j]$中至少部分数据满足预定条件$C_1$。图33中，判断$W_{i2}[k_i\text{-}170,k_i\text{-}1]$中至少部分数据是否满足预定条件$C_2$的方式和判断$W_{j2}[k_j\text{-}170,k_j\text{-}1]$中至少部分数据是否满足预定条件$C_2$的方式相同，

因此，如图33所示，“ ”表示判断窗口$W_{j2}[k_j\text{-}170,k_j\text{-}1]$中至少部分数据是否满足预定条件$C_2$时选择的1个字节，相邻两个选择的字节“ ”之间相差42个字节。使用Hash函数计算选择的5个字节，如果得到的数值为偶数，则$W_{j2}[k_j\text{-}170,k_j\text{-}1]$中至少部分数据满足预定条件$C_2$。图33中，判断$W_{i3}[k_i\text{-}171,k_i\text{-}2]$中至少部分数据是否满足预定条件$C_3$的方式与判断$W_{j3}[k_j\text{-}171,k_j\text{-}2]$中至少部分数据是否满足预定条件$C_3$的方式相同，因此，如图33所示，“ ”表示判断窗口$W_{j3}[k_j\text{-}171,k_j\text{-}2]$中至少部分数据是否满足预定条件$C_3$时选择的1个字节，相邻两个选择的字节“ ”之间相差42个字节。使用Hash函数计算选择的5个字节，得到的数值为偶数，则$W_{j3}[k_j\text{-}171,k_j\text{-}2]$中至少部分数据满足预定条件$C_3$。图33中，判断$W_{j4}[k_j\text{-}172,k_j\text{-}3]$中至少部分数据是否满足预定条件$C_4$的方式和判断窗口$W_{i4}[k_i\text{-}172,\ k_i\text{-}3]$中至少部分数据是否满足预定条件$C_4$的方式，因此，如图33所示，“ ”表示判断窗口$W_{j4}[k_j\text{-}172,k_j\text{-}3]$中至少部分数据是否满足预定条件$C_4$时选择的1个字节，相邻两个选择的字节“ ”之间相差42个字节。使用Hash函数计算选择的5个字节，得到的数值为偶数，则$W_{j4}[k_j\text{-}172,k_j\text{-}3]$中至少部分数据满足预定条件$C_4$。根据上述方法，判断$W_{j5}[k_j\text{-}173,k_j\text{-}4]$中至少部分数据是否满足预定条件$C_5$、判断$W_{j6}[k_j\text{-}174,k_j\text{-}5]$中至少部分数据是否满足预定条件$C_6$、判断$W_{j7}[k_j\text{-}175,k_j\text{-}6]$中至少部分数据是否满足预定条件$C_7$、判断$W_{j8}[k_j\text{-}176,k_j\text{-}7]$中至少部分数据是否满足预定条件$C_8$、判断$W_{j9}[k_j\text{-}177,k_j\text{-}8]$中至少部分数据是否满足预定条件$C_9$、判断$W_{j10}[k_j\text{-}178,k_j\text{-}9]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{j11}[k_j\text{-}179,k_j\text{-}10]$中至少部分数据是否满足预定条件$C_{11}$，在此不再赘述。

本实施例中使用随机函数判断窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 中至少部分数据是否满足预定条件 C_z，以图21所示的实施方式为例，根据在去重服务器103上预设的规则，为潜在分割点 k_i 确定窗口 $W_{i1}[k_i\text{-}169, k_i]$，判断 $W_{i1}[k_i\text{-}169, k_i]$ 中至少部分数据是否满足预定条件 C_1，如图32所示，W_{i1} 表示窗口 $W_{i1}[k_i\text{-}169, k_i]$，为判断 $W_{i1}[k_i\text{-}169, k_i]$ 中至少部分数据是否满足预定条件 C_1，选择5个字节，图32中序号为169、127、85、43和1的字节“■”分别表示选择的1个字节，相邻两个选择的字节之间相差42个字节。将序号为169、127、85、43和1的字节“■”分别转换成一个十进制数值，分别表示为 a_1、a_2、a_3、a_4 和 a_5。因为1个字节由8位组成，所以每个字节“■”作为一个数值，则 a_1、a_2、a_3、a_4 和 a_5 中的任一个 a_r 均满足 $0\leq a_r\leq 255$。a_1、a_2、a_3、a_4 和 a_5 组成1*5的矩阵。从服从二项分布的随机数中选择256*5个随机数，组成矩阵R，表示为：

$$\begin{pmatrix} h_{0,1} & h_{0,2} & \cdots & h_{0,5} \\ h_{1,1} & h_{1,2} & \cdots & h_{1,5} \\ \vdots & \vdots & \vdots & \vdots \\ h_{255,5} & h_{255,5} & \cdots & h_{255,5} \end{pmatrix},$$

根据 a_1 的值和所在的列，从矩阵R中查找对应的值，如 a_1=36，a_1 位于第1列，则查找 $h_{36,1}$ 对应的值；根据 a_2 的值和所在的列，从矩阵R中查找对应的值，如 a_2=48，a_2 位于第2列，则查找 $h_{48,2}$ 对应的值；根据 a_3 的值和所在的列，从矩阵R中查找对应的值，如 a_3=26，a_3 位于第3列，则查找 $h_{26,3}$ 对应的值；根据 a_4 的值和所在的列，从矩阵R中查找对应的值，如 a_4=26，a_4 位于第4列，则查找 $h_{26,4}$ 对应的值；根据 a_5 的值和所在的列，从矩阵R中查找对应的值，如 a_5=88，a_5 位于第5列，则查找 $h_{88,5}$ 对应的值。$S_1=h_{36,1}+h_{48,2}+h_{26,3}+h_{26,4}+h_{88,5}$，因为矩阵R服从二项分布，因此，$S_1$ 也服从二项分布。当 S_1 为偶数，则 $W_{i1}[k_i\text{-}169,$

k_i]中至少部分数据满足预定条件C_1，当S_1为奇数，则W_{i1}[k_i-169，k_i]中至少部分数据不满足预定条件C_1，S_1为偶数的概率为1/2，C_1表示按上述方式计算S_1为偶数。在图21所示实施例中，W_{i1}[k_i-169，k_i]中至少部分数据满足预定条件C_1。如图32所示，“▨”表示判断窗口W_{i2}[k_i-170,k_i-1]中至少部分数据是否满足预定条件C_2时分别选择的1个字节，在图32中，分别用序号170、128、86、44和2表示，相邻两个选择的字节之间相差42个字节。将序号170、128、86、44和2的字节“▨”分别转换成一个十进制数值，分别表示为b_1、b_2、b_3、b_4和b_5。因为1个字节由8位组成，所以每个字节“▨”作为一个数值，则b_1、b_2、b_3、b_4和b_5中的任一个b_r均满足$0\leq b_r\leq 255$。b_1、b_2、b_3、b_4和b_5组成1*5的矩阵。本实施方式中，判断W_{i1}和W_{i2}中至少部分数据是否满足预定条件的方式相同，因此仍然使用矩阵R，根据b_1的值和所在的列，从矩阵R中查找对应的值，如b_1=66，b_1位于第1列，则查找$h_{66,1}$对应的值；根据b_2的值和所在的列，从矩阵R中查找对应的值，如b_2=48，b_2位于第2列，则查找$h_{48,2}$对应的值；根据b_3的值和所在的列，从矩阵R中查找对应的值，如b_3=99，b_3位于第3列，则查找$h_{99,3}$对应的值；根据b_4的值和所在的列，从矩阵R中查找对应的值，如b_4=26，b_4位于第4列，则查找$h_{26,4}$对应的值；根据b_5的值和所在的列，从矩阵R中查找对应的值，如b_5=90，b_5位于第5列，则查找$h_{90,5}$对应的值。$S_2=h_{66,1}+h_{48,2}+h_{99,3}+h_{26,4}+h_{90,5}$,因为矩阵R服从二项分布，因此，$S_2$也服从二项分布。当$S_2$为偶数，则$W_{i2}$[$k_i$-170,$k_i$-1]中至少部分数据满足预定条件$C_2$，当$S_2$为奇数，则$W_{i2}$[$k_i$-170,$k_i$-1]中至少部分数据不满足预定条件$C_2$，$S_2$为偶数的概率为1/2。在图21所示实施例中，$W_{i2}$[$k_i$-170,$k_i$-1]中至少部分数据满足预定条件$C_2$。使用同样的规则，分别判断$W_{i3}$[$k_i$-171,

k_i-2]中至少部分数据是否满足预定条件C_3、判断$W_{i4}[k_i\text{-}172,k_i\text{-}3]$中至少部分数据是否满足预定条件$C_4$、判断$W_{i5}[k_i\text{-}173,k_i\text{-}4]$中至少部分数据是否满足预定条件$C_5$、判断$W_{i6}[k_i\text{-}174,k_i\text{-}5]$中至少部分数据是否满足预定条件$C_6$、判断$W_{i7}[k_i\text{-}175,k_i\text{-}6]$中至少部分数据是否满足预定条件$C_7$、判断$W_{i8}[k_i\text{-}176,k_i\text{-}7]$中至少部分数据是否满足预定条件$C_8$、判断$W_{i9}[k_i\text{-}177,k_i\text{-}8]$中至少部分数据是否满足预定条件$C_9$、判断$W_{i10}[k_i\text{-}178,k_i\text{-}9]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{i11}[k_i\text{-}179,k_i\text{-}10]$中至少部分数据是否满足预定条件$C_{11}$。图21所示的实施方式中，$W_{i5}[k_i\text{-}173,k_i\text{-}4]$中至少部分数据不满足预定条件$C_5$，从潜在分割点$k_i$沿着数据流分割点查找方向跳跃7个字节，在第7个字节的结束位置获得当前潜在分割点k_j，如图22所示，根据为去重服务器103预设的规则，为潜在分割点k_j确定窗口$W_{j1}[k_j\text{-}169,k_j]$，判断窗口$W_{j1}[k_j\text{-}169,k_j]$中至少部分数据是否满足预定条件$C_1$的方式与判断窗口$W_{i1}[k_i\text{-}169,k_i]$中至少部分数据是否满足预定条件$C_1$的方式相同，因此如图33所示，$W_{j1}$表示窗口$W_{j1}[k_j\text{-}169,k_j]$，为判断$W_{j1}[k_j\text{-}169,k_j]$中至少部分数据是否满足预定条件$C_1$，图33中序号为169、127、85、43和1的字节“■”分别表示选择的1个字节，相邻两个选择的字节之间相差42个字节。将序号为169、127、85、43和1的字节“■”分别转换成一个十进制数值，分别表示为a_1'、a_2'、a_3'、a_4'和a_5'。因为1个字节由8位组成，所以每个字节“■”作为一个数值，则a_1'、a_2'、a_3'、a_4'和a_5'中的任一个a_r'均满足$0\leq a_r'\leq 255$。a_1'、a_2'、a_3'、a_4'和a_5'组成1*5的矩阵。判断窗口$W_{j1}[k_j\text{-}169,k_j]$中至少部分数据是否满足预定条件$C_1$的方式与判断窗口$W_{i1}[k_i\text{-}169,k_i]$中至少部分数据是否满足预定条件$C_1$的方

式相同，因此，仍然使用矩阵R，表示为：$\begin{pmatrix} h_{0,1} & h_{0,2} & \cdots & h_{0,5} \\ h_{1,1} & h_{1,2} & \cdots & h_{1,5} \\ \vdots & \vdots & \vdots & \vdots \\ h_{255,5} & h_{255,5} & \cdots & h_{255,5} \end{pmatrix}$，

根据a_1'的值和所在的列，从矩阵R中查找对应的值，如a_1'=16，a_1'位于第1列，则查找$h_{16,1}$对应的值；根据a_2'的值和所在的列，从矩阵R中查找对应的值，如a_2'=98，a_2'位于第2列，则查找$h_{98,2}$对应的值；根据a_3'的值和所在的列，从矩阵R中查找对应的值，如a_3'=56，a_3'位于第3列，则查找$h_{56,3}$对应的值；根据a_4'的值和所在的列，从矩阵R中查找对应的值，如a_4'=36，a_4'位于第4列，则查找$h_{36,4}$对应的值；根据a_5'的值和所在的列，从矩阵R中查找对应的值，如a_5'=99，a_5'位于第5列，则查找$h_{99,5}$对应的值。$S_1'=h_{16,1}+h_{98,2}+h_{56,3}+h_{36,4}+h_{99,5}$，因为矩阵R服从二项分布，因此，$S_1'$也服从二项分布。当$S_1'$为偶数，则$W_{j1}[k_j\text{-}169,k_j]$中至少部分数据满足预定条件$C_1$，当$S_1'$为奇数，则$W_{j1}[k_j\text{-}169,k_j]$中至少部分数据不满足预定条件$C_1$，$S_1'$为偶数的概率为1/2。

判断$W_{i2}[k_i\text{-}170,k_i\text{-}1]$中至少部分数据是否满足预定条件$C_2$的方式和判断$W_{j2}[k_j\text{-}170,k_j\text{-}1]$中至少部分数据是否满足预定条件$C_2$的方式相同，因此，如图33所示，“▨”表示判断窗口$W_{j2}[k_j\text{-}170,k_j\text{-}1]$中至少部分数据是否满足预定条件$C_2$时选择的1个字节，相邻两个选择的字节之间相差42个字节，分别用序号170、128、86、44和2表示，相邻两个选择的字节之间相差42个字节。将序号170、128、86、44和2的字节“▨”分别转换成一个十进制数值，分别表示为b_1'、b_2'、b_3'、b_4'和b_5'。因为1个字节由8位组成，所以每个字节“▨”作为一个数值，则b_1'、b_2'、b_3'、b_4'和b_5'中的任一个b_r'均满足$0\leq b_r'\leq 255$。b_1'、b_2'、b_3'、

b_4'和b_5'组成1*5的矩阵。与判断窗口$W_{i2}[k_i\text{-}170,k_i\text{-}1]$中至少部分数据是否满足预定条件$C_2$使用相同的矩阵R，根据$b_1'$的值和所在的列，从矩阵R中查找对应的值，如$b_1'$=210，$b_1'$位于第1列，则查找$h_{210,1}$对应的值;根据$b_2'$的值和所在的列,从矩阵R中查找对应的值,如$b_2'$=156，$b_2'$位于第2列，则查找$h_{156,2}$对应的值；根据$b_3'$的值和所在的列，从矩阵R中查找对应的值，如$b_3'$=144，$b_3'$位于第3列，则查找$h_{144,3}$对应的值;根据$b_4'$的值和所在的列,从矩阵R中查找对应的值,如$b_4'$=60，$b_4'$位于第4列，则查找$h_{60,4}$对应的值；根据$b_5'$的值和所在的列，从矩阵R中查找对应的值，如$b_5'$=90，$b_5'$位于第5列，则查找$h_{90,5}$对应的值。$S_2'=h_{210,1}+h_{156,2}+h_{144,3}+h_{60,4}+h_{90,5}$，与$S_2$的判断条件相同，当$S_2'$为偶数，则$W_{j2}[k_j\text{-}170,k_j\text{-}1]$中至少部分数据满足预定条件$C_2$，当$S_2'$为奇数，则$W_{j2}[k_j\text{-}170,k_j\text{-}1]$中至少部分数据不满足预定条件$C_2$，$S_2'$为偶数的概率为1/2。

同理，判断$W_{i3}[k_i\text{-}171,k_i\text{-}2]$中至少部分数据是否满足预定条件$C_3$的方式与判断$W_{j3}[k_j\text{-}171,k_j\text{-}2]$中至少部分数据是否满足预定条件$C_3$的方式相同，同理，判断$W_{j4}[k_j\text{-}172,k_j\text{-}3]$中至少部分数据是否满足预定条件$C_4$、判断$W_{j5}[k_j\text{-}173,k_j\text{-}4]$中至少部分数据是否满足预定条件$C_5$、判断$W_{j6}[k_j\text{-}174,k_j\text{-}5]$中至少部分数据是否满足预定条件$C_6$、判断$W_{j7}[k_j\text{-}175,k_j\text{-}6]$中至少部分数据是否满足预定条件$C_7$、判断$W_{j8}[k_j\text{-}176,k_j\text{-}7]$中至少部分数据是否满足预定条件$C_8$、判断$W_{j9}[k_j\text{-}177,k_j\text{-}8]$中至少部分数据是否满足预定条件$C_9$、判断$W_{j10}[k_j\text{-}178,k_j\text{-}9]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{j11}[k_j\text{-}179,k_j\text{-}10]$中至少部分数据是否满足预定条件$C_{11}$，在此不再赘述。

本实施例中使用随机函数判断窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部

分数据是否满足预定条件C_z，以图21所示的实施方式为例，根据在去重服务器103上预设的规则，为潜在分割点k_i确定窗口$W_{i1}[k_i\text{-}169, k_i]$，判断$W_{i1}[k_i\text{-}169, k_i]$中至少部分数据是否满足预定的条件$C_1$，如图32所示，$W_{i1}$表示窗口$W_{i1}[k_i\text{-}169, k_i]$，为判断$W_{i1}[k_i\text{-}169, k_i]$中至少部分数据是否满足预定条件$C_1$，选择5个字节，图32中序号为169、127、85、43和1的字节“■”分别表示选择的1个字节，相邻两个选择的字节之间相差42个字节。将序号为169、127、85、43和1的字节“■”分别转换成一个十进制数值，分别表示为a_1、a_2、a_3、a_4和a_5。因为1个字节由8位组成，所以每个字节“■”作为一个数值，则a_1、a_2、a_3、a_4和a_5中的任一个a_s均满足$0\leq a_s\leq 255$。a_1、a_2、a_3、a_4和a_5组成1*5的矩阵。从服从二项分布的随机数中选择256*5个随机数，组成矩阵R，表示为：$\begin{pmatrix} h_{0,1} & h_{0,2} & \cdots & h_{0,5} \\ h_{1,1} & h_{1,2} & \cdots & h_{1,5} \\ \vdots & \vdots & \vdots & \vdots \\ h_{255,5} & h_{255,5} & \cdots & h_{255,5} \end{pmatrix}$，从服从二项分布的随机数中选择256*5个随机数，组成矩阵G，表示为：$\begin{pmatrix} g_{0,1} & g_{0,2} & \cdots & g_{0,5} \\ g_{1,1} & g_{1,2} & \cdots & g_{1,5} \\ \vdots & \vdots & \vdots & \vdots \\ g_{255,5} & g_{255,5} & \cdots & g_{255,5} \end{pmatrix}$。

根据a_1的值和所在的列，如a_1=36，a_1位于第1列，则从矩阵R中查找查找$h_{36,1}$对应的值，从矩阵G中查找$g_{36,1}$对应的值；根据a_2的值和所在的列，如a_2=48，a_2位于第2列，则从矩阵R中查$h_{48,2}$对应的值，从矩阵G中查找$g_{48,2}$对应的值；根据a_3的值和所在的列，如a_3=26，a_3位于第3列，则从矩阵R中查找$h_{26,3}$对应的值，从矩阵G中查找$g_{26,3}$对应的值；根据a_4的值和所在的列，如a_4=26，a_4位于第4列，则从矩阵R中查找$h_{26,4}$对应的值，从矩阵G中查找$g_{26,4}$对应的值；根据a_5的值和

所在的列，如a_5=88，a_5位于第5列，则从矩阵R中查找$h_{88,5}$对应的值，从矩阵G中查找$g_{88,5}$对应的值。$S_{1h}=h_{36,1}+h_{48,2}+h_{26,3}+h_{26,4}+h_{88,5}$,因为矩阵R服从二项分布，因此，$S_{1h}$也服从二项分布;$S_{1g}=g_{36,1}+g_{48,2}+g_{26,3}+g_{26,4}+g_{88,5}$，因为矩阵G服从二项分布，因此$S_{1g}$也服从二项分布。当$S_{1h}$和$S_{1g}$中有1个为偶数，则$W_{i1}[k_i\text{-}169,\ k_i]$中至少部分数据满足预定条件$C_1$，当$S_{1h}$和$S_{1g}$均为奇数，则$W_{i1}[k_i\text{-}169,\ k_i]$中至少部分数据不满足预定条件$C_1$，$C_1$表述按照上述方法获得的$S_{1h}$和$S_{1g}$中有1个为偶数。因为$S_{1h}$和$S_{1g}$均服从二项分布，因此$S_{1h}$为偶数的概率为1/2，$S_{1g}$为偶数的概率为1/2，$S_{1h}$和$S_{1g}$中有1个为偶数的概率为1-1/4=3/4，因此，$W_{i1}[k_i\text{-}169,\ k_i]$中至少部分数据满足预定条件$C_1$的概率为3/4。在图21所示实施例中，$W_{i1}[k_i\text{-}169,\ k_i]$中至少部分数据满足预定条件$C_1$。在图21所示的实施方式中，在$W_{i1}[k_i\text{-}169,k_i]$、$W_{i2}[k_i\text{-}170,k_i\text{-}1]$、$W_{i3}[k_i\text{-}171,k_i\text{-}2]$、$W_{i4}[k_i\text{-}172,k_i\text{-}3]$、$W_{i5}[k_i\text{-}173,k_i\text{-}4]$、$W_{i6}[k_i\text{-}174,k_i\text{-}5]$、$W_{i7}[k_i\text{-}175,k_i\text{-}6]$、$W_{i8}[k_i\text{-}176,k_i\text{-}7]$、$W_{i9}[k_i\text{-}177,k_i\text{-}8]$、$W_{i10}[k_i\text{-}178,k_i\text{-}9]$和$W_{i11}[k_i\text{-}179,k_i\text{-}10]$中，各窗口大小相同,即窗口大小均为169字节，同时判断窗口中至少部分数据是否满足预定条件的方式也相同,具体见上述判断$W_{i1}[k_i\text{-}169,k_i]$中至少部分数据是否满足预定条件$C_1$的描述。因此，如图32所示，“▨”表示判断窗口$W_{i2}[k_i\text{-}170,k_i\text{-}1]$中至少部分数据是否满足预定条件$C_2$时分别选择的1个字节，在图32中，分别用序号170、128、86、44和2表示，相邻两个选择的字节之间相差42个字节。将序号170、128、86、44和2的字节“▨”分别转换成一个十进制数值，分别表示为b_1、b_2、b_3、b_4和b_5。因为1个字节由8位组成，所以每个字节“▨”作为一个数值，则b_1、b_2、b_3、b_4和b_5中的任一个b_s均满足$0\le b_s\le 255$。b_1、b_2、b_3、b_4和b_5组成1*5的矩阵。本实施方式中，判断各窗口中至

少部分数据是否满足预定条件的方式相同，因此仍然使用相同矩阵R和G。根据b_1的值和所在的列，如b_1=66，b_1位于第1列，则从矩阵R中查找$h_{66,1}$对应的值，从矩阵G中查找$g_{66,1}$对应的值；根据b_2的值和所在的列，如b_2=48，b_2位于第2列，则从矩阵R中查找$h_{48,2}$对应的值，从矩阵G中查找$g_{48,2}$对应的值；根据b_3的值和所在的列，如b_3=99，b_3位于第3列，则从矩阵R中查找$h_{99,3}$对应的值，从矩阵G中查找$g_{99,3}$对应的值；根据b_4的值和所在的列，如b_4=26，b_4位于第4列，则从矩阵R中查找$h_{26,4}$对应的值，从矩阵G中查找$g_{26,4}$对应的值；根据b_5的值和所在的列，如b_5=90，b_5位于第5列，则从矩阵R中查找$h_{90,5}$对应的值，从矩阵G中查找$g_{90,5}$对应的值。$S_{2h}=h_{66,1}+h_{48,2}+h_{99,3}+h_{26,4}+h_{90,5}$,因为矩阵R服从二项分布，因此，$S_{2h}$也服从二项分布。$S_{2g}=g_{66,1}+g_{48,2}+g_{99,3}+g_{26,4}+g_{90,5}$，因为矩阵G服从二项分布，因此，$S_{2g}$也服从二项分布。当$S_{2h}$和$S_{2g}$中有1个为偶数，则$W_{i2}[k_i\text{-}170,k_i\text{-}1]$中至少部分数据满足预定条件$C_2$，当$S_{2h}$和$S_{2g}$均为奇数，则$W_{i2}[k_i\text{-}170,k_i\text{-}1]$中至少部分数据不满足预定条件$C_2$，$S_{2h}$和$S_{2g}$中有1个为偶数的概率为3/4。在图21所示实施例中，$W_{i2}[k_i\text{-}170,k_i\text{-}1]$中至少部分数据满足预定条件$C_2$。使用同样的规则，分别判断$W_{i3}[k_i\text{-}171,k_i\text{-}2]$中至少部分数据是否满足预定条件$C_3$、判断$W_{i4}[k_i\text{-}172,k_i\text{-}3]$中至少部分数据是否满足预定条件$C_4$、判断$W_{i5}[k_i\text{-}173,k_i\text{-}4]$中至少部分数据是否满足预定条件$C_5$、判断$W_{i6}[k_i\text{-}174,k_i\text{-}5]$中至少部分数据是否满足预定条件$C_6$、判断$W_{i7}[k_i\text{-}175,k_i\text{-}6]$中至少部分数据是否满足预定条件$C_7$、判断$W_{i8}[k_i\text{-}176,k_i\text{-}7]$中至少部分数据是否满足预定条件$C_8$、判断$W_{i9}[k_i\text{-}177,k_i\text{-}8]$中至少部分数据是否满足预定条件$C_9$、判断$W_{i10}[k_i\text{-}178,k_i\text{-}9]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{i11}[k_i\text{-}179,k_i\text{-}10]$中至少部分数据是否满足预定条

件C_{11}。图21所示的实施方式中，$W_{i5}[k_i\text{-}173,k_i\text{-}4]$中至少部分数据不满足预定条件$C_5$，从潜在分割点$k_i$沿着数据流分割点查找方向跳跃7个字节，在第7个字节的结束位置获得当前潜在分割点k_j，如图22所示，根据为去重服务器103预设的规则，为潜在分割点k_j确定窗口$W_{j1}[k_j\text{-}169,k_j]$，判断窗口$W_{j1}[k_j\text{-}169,k_j]$中至少部分数据是否满足预定条件$C_1$的方式与判断窗口$W_{i1}[k_i\text{-}169,k_i]$中至少部分数据是否满足预定条件$C_1$的方式相同，因此如图33所示，$W_{j1}$表示窗口$W_{j1}[k_j\text{-}169,k_j]$，为判断$W_{j1}[k_j\text{-}169,k_j]$中至少部分数据是否满足预定条件$C_1$，图33中序号为169、127、85、43和1的字节“■”分别表示选择的1个字节，相邻两个选择的字节之间相差42个字节。将序号为169、127、85、43和1的字节“■”分别转换成一个十进制数值，分别表示为a_1'、a_2'、a_3'、a_4'和a_5'。因为1个字节由8位组成，所以每个字节“■”作为一个数值，则a_1'、a_2'、a_3'、a_4'和a_5'中的任一个a_s'均满足$0\leq a_s'\leq 255$。a_1'、a_2'、a_3'、a_4'和a_5'组成1*5的矩阵。使用与判断窗口$W_{i1}[k_i\text{-}169,k_i]$中至少部分数据是否满足预定条件$C_1$相同的矩阵R和G，分别表示为：

$$\begin{pmatrix} h_{0,1} & h_{0,2} & \cdots & h_{0,5} \\ h_{1,1} & h_{1,2} & \cdots & h_{1,5} \\ \vdots & \vdots & \vdots & \vdots \\ h_{255,5} & h_{255,5} & \cdots & h_{255,5} \end{pmatrix} \text{和} \begin{pmatrix} g_{0,1} & g_{0,2} & \cdots & g_{0,5} \\ g_{1,1} & g_{1,2} & \cdots & g_{1,5} \\ \vdots & \vdots & \vdots & \vdots \\ g_{255,5} & g_{255,5} & \cdots & g_{255,5} \end{pmatrix}。$$

根据a_1'的值和所在的列，如a_1'=16，a_1'位于第1列，则从矩阵R中查找$h_{16,1}$对应的值，从矩阵G中查找$g_{16,1}$对应的值；根据a_2'的值和所在的列，如a_2'=98，a_2'位于第2列，则从矩阵R中查找$h_{98,2}$对应的值，从矩阵G中查找$g_{98,2}$对应的值；根据a_3'的值和所在的列，如a_3'=56，a_3'位于第3列，则从矩阵R中查找$h_{56,3}$对应的值，从矩阵G中查找$g_{56,3}$对应的值；根据a_4'的值和所在的列，如a_4'=36，a_4'位于第4列，则从矩阵R

中查找$h_{36,4}$对应的值，从矩阵G中查找$g_{36,4}$对应的值；根据a_5'的值和所在的列，如a_5'=99，a_5'位于第5列，则从矩阵R中查找$h_{99,5}$对应的值，从矩阵G中查找$g_{99,5}$对应的值。$S_{1h}'=h_{16,1}+h_{98,2}+h_{56,3}+h_{36,4}+h_{99,5}$，因为矩阵R服从二项分布，因此，$S_{1h}'$也服从二项分布；$S_{1g}'=g_{16,1}+g_{98,2}+g_{56,3}+g_{36,4}+g_{99,5}$，因为矩阵G服从二项分布，因此$S_{1g}'$也服从二项分布。当$S_{1h}'$和$S_{1g}'$中有1个为偶数，则$W_{j1}[k_j\text{-}169,k_j]$中至少部分数据满足预定条件$C_1$，当$S_{1h}'$和$S_{1g}'$均为奇数，则$W_{j1}[k_j\text{-}169,k_j]$中至少部分数据不满足预定条件$C_1$，$S_{1h}'$和$S_{1g}'$有1个为偶数的概率为3/4。

判断$W_{i2}[k_i\text{-}170,k_i\text{-}1]$中至少部分数据是否满足预定条件$C_2$的方式和判断$W_{j2}[k_j\text{-}170,k_j\text{-}1]$中至少部分数据是否满足预定条件$C_2$的方式相同，因此，如图33所示，“▨”表示判断窗口$W_{j2}[k_j\text{-}170,k_j\text{-}1]$中至少部分数据是否满足预定条件$C_2$时选择的1个字节，相邻两个选择的字节之间相差42个字节。在图33中，分别用序号170、128、86、44和2表示，相邻两个选择的字节之间相差42个字节。将序号170、128、86、44和2的字节“▨”分别转换成一个十进制数值，分别表示为b_1'、b_2'、b_3'、b_4'和b_5'。因为1个字节由8位组成，所以每个字节“▨”作为一个数值，则b_1'、b_2'、b_3'、b_4'和b_5'中的任一个b_s'均满足$0\le b_s'\le 255$。b_1'、b_2'、b_3'、b_4'和b_5'组成1*5的矩阵。使用与判断窗口$W_{j2}[k_j\text{-}170,k_j\text{-}1]$中至少部分数据是否满足预定条件$C_2$相同的矩阵R和G，根据$b_1'$的值和所在的列，如$b_1'$=210，$b_1'$位于第1列，则从矩阵R中查找$h_{210,1}$对应的值，从矩阵G中查找$g_{210,1}$对应的值；根据$b_2'$的值和所在的列，如$b_2'$=156，$b_2'$位于第2列，则从矩阵R中查找$h_{156,2}$对应的值，从矩阵G中查找$g_{156,2}$对应的值；根据$b_3'$的值和所在的列，如$b_3'$=144，$b_3'$位于第3列，则从矩阵R中查找$h_{144,3}$对应的值，从矩阵G中查找$g_{144,3}$对应的值；

根据b_4'的值和所在的列，如b_4'=60，b_4'位于第4列，则从矩阵R中查找$h_{60,4}$对应的值，从矩阵G中查找$g_{60,4}$对应的值；根据b_5'的值和所在的列，如b_5'=90，b_5'位于第5列，则从矩阵R中查找$h_{90,5}$对应的值，从矩阵G中查找$g_{90,5}$对应的值。$S_{2h}'=h_{210,1}+h_{156,2}+h_{144,3}+h_{60,4}+h_{90,5}$，$S_{2g}'=g_{210,1}+g_{156,2}+g_{144,3}+g_{60,4}+g_{90,5}$。当$S_{2h}$'和$S_{2g}$'中有1个为偶数，则$W_{j2}[k_j\text{-}170,k_j\text{-}1]$中至少部分数据满足预定条件$C_2$，当$S_{2h}$'和$S_{2g}$'均为奇数，则$W_{j2}[k_j\text{-}170,k_j\text{-}1]$中至少部分数据不满足预定条件$C_2$，$S_{2h}$'和$S_{2g}$'中有1个为偶数的概率为3/4。

同理，判断$W_{i3}[k_i\text{-}171,k_i\text{-}2]$中至少部分数据是否满足预定条件$C_3$的方式与判断$W_{j3}[k_j\text{-}171,k_j\text{-}2]$中至少部分数据是否满足预定条件$C_3$的方式相同，同理，判断$W_{j4}[k_j\text{-}172,k_j\text{-}3]$中至少部分数据是否满足预定条件$C_4$、判断$W_{j5}[k_j\text{-}173,k_j\text{-}4]$中至少部分数据是否满足预定条件$C_5$、判断$W_{j6}[k_j\text{-}174,k_j\text{-}5]$中至少部分数据是否满足预定条件$C_6$、判断$W_{j7}[k_j\text{-}175,k_j\text{-}6]$中至少部分数据是否满足预定条件$C_7$、判断$W_{j8}[k_j\text{-}176,k_j\text{-}7]$中至少部分数据是否满足预定条件$C_8$、判断$W_{j9}[k_j\text{-}177,k_j\text{-}8]$中至少部分数据是否满足预定条件$C_9$、判断$W_{j10}[k_j\text{-}178,k_j\text{-}9]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{j11}[k_j\text{-}179,k_j\text{-}10]$中至少部分数据是否满足预定条件$C_{11}$，在此不再赘述。

本实施例中使用随机函数判断窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据是否满足预定条件C_z，以图21所示的实施方式为例，根据在去重服务器103上预设的规则，为潜在分割点k_i确定窗口$W_{i1}[k_i\text{-}169,k_i]$，判断$W_{i1}[k_i\text{-}169,k_i]$中至少部分数据是否满足预定的条件$C_1$，如图32所示，$W_{i1}$表示窗口$W_{i1}[k_i\text{-}169,k_i]$，为判断$W_{i1}[k_i\text{-}169,k_i]$中至少部分数据是否满足预定条件$C_1$，选择5个字节，图32中序号为169、127、

85、43和1的字节“■”分别表示选择的1个字节，相邻两个选择的字节之间相差42个字节。将序号为169、127、85、43和1的字节“■”依次看成40个位，分别表示为 a_1、a_2、a_3、$a_4 \ldots a_{40}$。a_1、a_2、a_3、$a_4 \ldots a_{40}$ 中的任一 a_t，当 a_t=0时，V_{at} =-1，当 a_t=1时，V_{at}=1，根据 a_t 与 V_{at} 对应关系，生成 V_{a1}、V_{a2}、V_{a3}、$V_{a4} \ldots V_{a40}$。从服从正态分布的随机数中选择40个随机数，分别表示为：h_1、h_2、h_3、$h_4 \ldots h_{40}$。$S_a = V_{a1} * h_1 + V_{a2} * h_2 + V_{a3} * h_3 + V_{a4} * h_4 + \ldots + V_{a40} * h_{40}$。因为 h_1、h_2、h_3、$h_4 \ldots h_{40}$ 服从正态分布，因此，S_a 也服从正态分布。当 S_a 为正数，则 $W_{i1}[k_i\text{-}169, k_i]$ 中至少部分数据满足预定条件 C_1，当 S_a 为负数或0，则 $W_{i1}[k_i\text{-}169, k_i]$ 中至少部分数据不满足预定条件 C_1，S_a 为正数的概率为1/2。在图21所示实施例中，$W_{i1}[k_i\text{-}169, k_i]$ 中至少部分数据满足预定条件 C_1。如图32所示，“▨”表示判断窗口 $W_{i2}[k_i\text{-}170, k_i\text{-}1]$ 中至少部分数据是否满足预定条件 C_2 时分别选择的1个字节，在图32中，分别用序号170、128、86、44和2表示，相邻两个选择的字节之间相差42个字节。将序号170、128、86、44和2的字节“▨”依次看成40个位，分别表示为 b_1、b_2、b_3、$b_4 \ldots b_{40}$。b_1、b_2、b_3、$b_4 \ldots b_{40}$ 中的任一 b_t，当 b_t=0时，V_{bt}=-1，当 b_t=1时，V_{bt}=1，根据 b_t 与 V_{bt} 对应关系，生成 V_{b1}、V_{b2}、V_{b3}、$V_{b4} \ldots V_{b40}$。判断窗口 $W_{i1}[k_i\text{-}169, k_i]$ 中至少部分数据是否满足预定条件 C_1 的方式与判断窗口 $W_{i2}[k_i\text{-}170, k_i\text{-}1]$ 中至少部分数据是否满足预定条件 C_2 的方式相同，因此，使用相同的随机数：h_1、h_2、h_3、$h_4 \ldots h_{40}$，$S_b = V_{b1} * h_1 + V_{b2} * h_2 + V_{b3} * h_3 + V_{b4} * h_4 + \ldots + V_{b40} * h_{40}$。因为 h_1、h_2、h_3、$h_4 \ldots h_{40}$ 服从正态分布，因此，S_b 也服从正态分布。当 S_b 为正数，则 $W_{i2}[k_i\text{-}170, k_i\text{-}1]$ 中至少部分数据满足预定条件 C_2，当 S_b 为负数或0，则 $W_{i2}[k_i\text{-}170, k_i\text{-}1]$ 中至少部分数据不满足预定条件 C_2，S_b 为正数的概率为1/2。在

图21所示实施例中，$W_{i2}[k_i\text{-}170, k_i\text{-}1]$中至少部分数据满足预定条件$C_2$。使用同样的规则，分别判断$W_{i3}[k_i\text{-}171,k_i\text{-}2]$中至少部分数据是否满足预定条件$C_3$、判断$W_{i4}[k_i\text{-}172,k_i\text{-}3]$中至少部分数据是否满足预定条件$C_4$、判断$W_{i5}[k_i\text{-}173,k_i\text{-}4]$中至少部分数据是否满足预定条件$C_5$、判断$W_{i6}[k_i\text{-}174,k_i\text{-}5]$中至少部分数据是否满足预定条件$C_6$、判断$W_{i7}[k_i\text{-}175,k_i\text{-}6]$中至少部分数据是否满足预定条件$C_7$、判断$W_{i8}[k_i\text{-}176,k_i\text{-}7]$中至少部分数据是否满足预定条件$C_8$、判断$W_{i9}[k_i\text{-}177,k_i\text{-}8]$中至少部分数据是否满足预定条件$C_9$、判断$W_{i10}[k_i\text{-}178,k_i\text{-}9]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{i11}[k_i\text{-}179,k_i\text{-}10]$中至少部分数据是否满足预定条件$C_{11}$。图21所示的实施方式中，$W_{i5}[k_i\text{-}173,k_i\text{-}4]$中至少部分数据不满足预定条件$C_5$，从潜在分割点$k_i$沿着数据流分割点查找方向跳跃7个字节，在第7个字节的结束位置获得当前潜在分割点k_j，如图22所示，根据为去重服务器103预设的规则，为潜在分割点k_j确定窗口$W_{j1}[k_j\text{-}169,k_j]$，判断窗口$W_{j1}[k_j\text{-}169,k_j]$中至少部分数据是否满足预定条件$C_1$的方式与判断窗口$W_{i1}[k_i\text{-}169,k_i]$中至少部分数据是否满足预定条件$C_1$的方式相同，因此如图33所示，$W_{j1}$表示窗口$W_{j1}[k_j\text{-}169,k_j]$，为判断$W_{j1}[k_j\text{-}169,k_j]$中至少部分数据是否满足预定条件$C_1$，选择5个字节，图33中序号为169、127、85、43和1的字节“■”分别表示选择的1个字节，相邻两个选择的字节之间相差42个字节。将序号为169、127、85、43和1的字节“■”依次看成40个位，分别表示为a_1'、a_2'、a_3'、$a_4'\ldots a_{40}'$。a_1'、a_2'、a_3'、$a_4'\ldots a_{40}'$中的任一a_t'，当a_t'=0时，V_{at}' =-1，当a_t'=1时，V_{at}'=1，根据a_t'与V_{at}'对应关系，生成V_{a1}'、V_{a2}'、V_{a3}'、$V_{a4}'\ldots V_{a40}'$。判断窗口$W_{j1}[k_j\text{-}169,k_j]$中至少部分数据是否满足预定条件$C_1$的方式与判断窗口$W_{i1}[k_i\text{-}169,k_i]$中至少部分数据是

否满足预定条件C_1的方式相同，因此使用相同的随机数：h_1、h_2、h_3、h_4...h_{40}。$S_a'=V_{a1}'*h_1+V_{a2}'*h_2+V_{a3}'*h_3+V_{a4}'*h_4+...+V_{a40}'*h_{40}$。因为$h_1$、$h_2$、$h_3$、$h_4$...$h_{40}$服从正态分布，因此，$S_a'$也服从正态分布。当$S_a'$为正数，则$W_{j1}[k_j\text{-}169,k_j]$中至少部分数据满足预定条件$C_1$，当$S_a'$为负数或0，则$W_{j1}[k_j\text{-}169,k_j]$中至少部分数据不满足预定条件$C_1$，$S_a'$为正数的概率为1/2。

判断$W_{i2}[k_i\text{-}170,k_i\text{-}1]$中至少部分数据是否满足预定条件$C_2$的方式和判断$W_{j2}[k_j\text{-}170,k_j\text{-}1]$中至少部分数据是否满足预定条件$C_2$的方式相同，因此，如图33所示，“▨”表示判断窗口$W_{j2}[k_j\text{-}170,k_j\text{-}1]$中至少部分数据是否满足预定条件$C_2$时选择的1个字节，相邻两个选择的字节之间相差42个字节。在图33中，分别用序号170、128、86、44和2表示，相邻两个选择的字节之间相差42个字节。将序号170、128、86、44和2的字节“▨”依次看成40个位，分别表示为b_1'、b_2'、b_3'、b_4'...b_{40}'。b_1'、b_2'、b_3'、b_4'...b_{40}'中的任一b_t'，当b_t'=0时，V_{bt}'=-1，当b_t'=1时，V_{bt}' =1，根据b_t'与V_{bt}'对应关系，生成V_{b1}'、V_{b2}'、V_{b3}'、V_{b4}'...V_{b40}'。判断$W_{i2}[k_i\text{-}170,k_i\text{-}1]$中至少部分数据是否满足预定条件$C_2$的方式和判断$W_{j2}[k_j\text{-}170,k_j\text{-}1]$中至少部分数据是否满足预定条件$C_2$的方式相同，因此，使用相同的随机数：$h_1$、$h_2$、$h_3$、$h_4$...$h_{40}$，$S_b'=V_{b1}'*h_1+V_{b2}'*h_2+V_{b3}'*h_3+V_{b4}'*h_4+...+V_{b40}'*h_{40}$。因为$h_1$、$h_2$、$h_3$、$h_4$...$h_{40}$服从正态分布，因此，$S_b'$也服从正态分布。当$S_b'$为正数，则$W_{j2}[k_j\text{-}170,k_j\text{-}1]$中至少部分数据满足预定条件$C_2$，当$S_b'$为负数或0，则$W_{j2}[k_j\text{-}170,k_j\text{-}1]$中至少部分数据不满足预定条件$C_2$，$S_b'$为正数的概率为1/2。

同理，判断$W_{i3}[k_i\text{-}171,k_i\text{-}2]$中至少部分数据是否满足预定条件$C_3$的方式与判断$W_{j3}[k_j\text{-}171,k_j\text{-}2]$中至少部分数据是否满足预定条件

C_3的方式相同，同理，判断$W_{j4}[k_j\text{-}172,k_j\text{-}3]$中至少部分数据是否满足预定条件$C_4$、判断$W_{j5}[k_j\text{-}173,k_j\text{-}4]$中至少部分数据是否满足预定条件$C_5$、判断$W_{j6}[k_j\text{-}174,k_j\text{-}5]$中至少部分数据是否满足预定条件$C_6$、判断$W_{j7}[k_j\text{-}175,k_j\text{-}6]$中至少部分数据是否满足预定条件$C_7$、判断$W_{j8}[k_j\text{-}176,k_j\text{-}7]$中至少部分数据是否满足预定条件$C_8$、判断$W_{j9}[k_j\text{-}177,k_j\text{-}8]$中至少部分数据是否满足预定条件$C_9$、判断$W_{j10}[k_j\text{-}178,k_j\text{-}9]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{j11}[k_j\text{-}179,k_j\text{-}10]$中至少部分数据是否满足预定条件$C_{11}$，在此不再赘述。

本实施例中使用随机函数判断窗口$W_{iz}[k_i-A_z,k_i+B_z]$中至少部分数据是否满足预定条件C_z，仍然以图21所示实施方式为例，根据在去重服务器103上预设的规则，为潜在分割点k_i确定窗口$W_{i1}[k_i\text{-}169,k_i]$，判断$W_{i1}[k_i\text{-}169,k_i]$中至少部分数据是否满足预定的条件$C_1$，如图32所示，$W_{i1}$表示窗口$W_{i1}[k_i\text{-}169,k_i]$，为判断$W_{i1}[k_i\text{-}169,k_i]$中至少部分数据是否满足预定条件$C_1$，选择5个字节，图32中序号为169、127、85、43和1的字节“■”分别表示选择的1个字节，相邻两个选择的字节之间相差42个字节。将序号为169、127、85、43和1的字节“■”转换成1个十进制数，范围为0-（2^40-1），使用均匀分布随机数生成器为0-（2^40-1）中的每一个十进制数生成1个指定值，记录0-（2^40-1）中的每一个十进制数与指定值之间的对应关系R，一旦指定则该十进制数对应的指定值就不变，该指定值服从均匀分布，如果该指定值为偶数，则$W_{i1}[k_i\text{-}169,k_i]$中至少部分数据满足预定条件$C_1$，如果该指定值为奇数，则$W_{i1}[k_i\text{-}169,k_i]$中至少部分数据不满足预定条件$C_1$，$C_1$表示按照上述方法获得的指定值为偶数。因为均匀分布的随机数为偶数的概率为1/2，因此，$W_{i1}[k_i\text{-}169,k_i]$中至少部分数据满足预定条件$C_1$

的概率为1/2。在图21所示的实施方式中，使用同样的规则，分别判断$W_{i2}[k_i\text{-}170,k_i\text{-}1]$中至少部分数据是否满足预定条件$C_2$，判断$W_{i3}[k_i\text{-}171,k_i\text{-}2]$中至少部分数据是否满足预定条件$C_3$、判断$W_{i4}[k_i\text{-}172,k_i\text{-}3]$中至少部分数据是否满足预定条件$C_4$、判断$W_{i5}[k_i\text{-}173,k_i\text{-}4]$中至少部分数据是否满足预定条件$C_5$，在此不再赘述。

当$W_{i5}[k_i\text{-}173,k_i\text{-}4]$中至少部分数据不满足预定条件$C_5$，从潜在分割点$k_i$沿着数据流分割点查找方向跳跃7个字节，在第7个字节的结束位置获得当前潜在分割点k_j，如图22所示，根据为去重服务器103预设的规则，为潜在分割点k_j确定窗口$W_{j1}[k_j\text{-}169,k_j]$，判断窗口$W_{j1}[k_j\text{-}169,k_j]$中至少部分数据是否满足预定条件$C_1$的方式与判断窗口$W_{i1}[k_i\text{-}169,k_i]$中至少部分数据是否满足预定条件$C_1$的方式相同，因此，使用相同的0-（2^40-1）中的每一个十进制数与指定值之间的对应关系R，如图33所示，W_{j1}表示窗口，为判断$W_{j1}[k_j\text{-}169,k_j]$中至少部分数据是否满足预定条件$C_1$，选择5个字节，图33中“■”表示选择的1个字节，相邻两个选择的字节“■”之间相差42个字节。将序号为169、127、85、43和1的字节“■”转换成1个十进制数，在R查找该十进制数对应的指定值，如果该指定值为偶数，则$W_{j1}[k_j\text{-}169,k_j]$中至少部分数据满足预定条件$C_1$，如果该指定值为奇数，则$W_{j1}[k_j\text{-}169,k_j]$中至少部分数据不满足预定条件$C_1$，因为均匀分布的随机数为偶数的概率为1/2，因此，$W_{j1}[k_j\text{-}169,k_j]$中至少部分数据满足预定条件$C_1$的概率为1/2。同理，判断$W_{i2}[k_i\text{-}170,k_i\text{-}1]$中至少部分数据是否满足预定条件$C_2$的方式和判断$W_{j2}[k_j\text{-}170,k_j\text{-}1]$中至少部分数据是否满足预定条件$C_2$的方式相同，判断$W_{i3}[k_i\text{-}171,k_i\text{-}2]$中至少部分数据是否满足预定条件$C_3$的方式与判断$W_{j3}[k_j\text{-}171,k_j\text{-}2]$中至少部分数据是否满足预定条件

C_3的方式相同，同理，判断$W_{j4}[k_j\text{-}172,k_j\text{-}3]$中至少部分数据是否满足预定条件$C_4$、判断$W_{j5}[k_j\text{-}173,k_j\text{-}4]$中至少部分数据是否满足预定条件$C_5$、判断$W_{j6}[k_j\text{-}174,k_j\text{-}5]$中至少部分数据是否满足预定条件$C_6$、判断$W_{j7}[k_j\text{-}175,k_j\text{-}6]$中至少部分数据是否满足预定条件$C_7$、判断$W_{j8}[k_j\text{-}176,k_j\text{-}7]$中至少部分数据是否满足预定条件$C_8$、判断$W_{j9}[k_j\text{-}177,k_j\text{-}8]$中至少部分数据是否满足预定条件$C_9$、判断$W_{j10}[k_j\text{-}178,k_j\text{-}9]$中至少部分数据是否满足预定条件$C_{10}$和判断$W_{j11}[k_j\text{-}179,k_j\text{-}10]$中至少部分数据是否满足预定条件$C_{11}$，在此不再赘述。

图1所示的本发明实施例中的去重服务器103，是指能够实现本发明实施例所描述的技术方案的装置，如图18所示，通常包括中央处理单元、主存储器以及输入输出接口。中央处理单元、主存储器与输入输出接口之间相互通信，主存储器存储可执行指令，中央处理单元执行主存储器中存储的可执行指令，从而执行特定的功能，使去重服务器103具备特定功能，如本发明实施例图20至图33所描述的查找数据流分割点。因此，如图19所示，根据20至图33所示的本发明实施例，去重服务器103，在去重服务器103上预设有规则，所述规则为：为潜在分割点k确定M个窗口$W_x[k-A_x, k+B_x]$和窗口$W_x[k-A_x, k+B_x]$对应的预定条件C_x，其中，x为1到M连续的自然数，M≥2，A_x和B_x为整数；

去重服务器103包括确定单元1901和判断处理单元1902。其中，确定单元1901用于执行步骤a）：

a）依据所述规则为当前潜在分割点k_i确定对应的窗口$W_{iz}[k_i\text{-}A_z, k_i+B_z]$，i和z为整数，并且1≤z≤M；

判断处理单元1902，用于判断所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少

部分数据是否满足预定条件C_z；

当所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据不满足所述预定条件C_z，从所述当前潜在分割点k_i沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，N*U不大于$\| B_z \| + \max_x(\| A_x \|)$，获得新的潜在分割点，则所述确定单元1901为所述新的潜在分割点执行步骤a）；

当所述当前潜在分割点k_i的M个窗口中的每一个窗口$W_{ix}[k_i-A_x, k_i+B_x]$中至少部分数据满足预定条件C_x，则所述当前潜在分割点k_i为数据流分割点。

进一步地，所述规则还包括：至少两个窗口$W_{ie}[k_i-A_e, k_i+B_e]$与$W_{if}[k_i-A_f, k_i+B_f]$，满足条件：$|A_e+B_e|=|A_f+B_f|$，$C_e=C_f$。进一步地，所述规则还包括：A_e和A_f为正整数。进一步地，所述规则还包括：$A_e-1=A_f$，$B_e+1=B_f$。

进一步地，判断处理单元1902具体用于使用随机函数判断窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据是否满足预定条件C_z。更进一步地，判断处理单元1902具体使用hash函数判断窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据是否满足预定条件C_z。

进一步地，判断处理单元1902用于当所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据不满足所述预定条件C_z，从所述当前潜在分割点k_i沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，获得所述新的潜在分割点，所述确定单元1901为所述新的潜在分割点执行步骤a），根据所述规则，为所述新的潜在分割点确定的窗口$W_{ic}[k_i-A_c, k_i+B_c]$的左边界与所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$的右边界重合或者为所述新的潜在分割点确定的所述窗口$W_{ic}[k_i-A_c, k_i+B_c]$的左边界位

于所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 范围之内；其中，为所述新的潜在分割点确定的所述窗口 $W_{ic}[k_i-A_c, k_i+B_c]$ 是根据所述规则，为所述新的潜在分割点确定的M个窗口按照数据流查找方向获得的序列中排序第一的窗口。

进一步地，判断处理单元1902使用随机函数判断所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 中至少部分数据是否满足所述预定条件 C_z，具体包括：

在所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 中选择F个字节，将所述F个字节反复利用H次，共获得F*H个字节，其中每个字节由8位组成，记为 $a_{m,1}\ldots a_{m,8}$，表示所述F*H个字节中第m个字节的第1到第8位，所述F*H个字节对应的位可以表示为：$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,8} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ a_{F*H,1} & a_{F*H,2} & \cdots & a_{F*H,8} \end{pmatrix}$，当 $a_{m,n}$=1时，$V_{am,n}$=1，当 $a_{m,n}$=0时，$V_{am,n}$=-1，其中 $a_{m,n}$ 表示 $a_{m,1}\ldots a_{m,8}$ 中的任一个，所述F*H个字节对应的位按照 $a_{m,n}$ 与 $V_{am,n}$ 的转换关系得到矩阵 V_a，所述矩阵 V_a 表示为：$\begin{pmatrix} V_{a1,1} & V_{a1,2} & \cdots & V_{a1,8} \\ V_{a2,1} & V_{a2,2} & \cdots & V_{a2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{aF*H,1} & V_{aF*H,2} & \cdots & V_{aF*H,8} \end{pmatrix}$，从服务正态分布的随机数中选择F*H*8个随机数组成矩阵R，所述矩阵R表示为：

$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,8} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{F*H,1} & h_{F*H,2} & \cdots & h_{F*H,8} \end{pmatrix}$，将所述矩阵 V_a 的第m行与所述矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为 $S_{am}=V_{am,1}*h_{m,1}+V_{am,2}*h_{m,2}+\ldots+V_{am,8}*h_{m,8}$，同理，获得 S_{a1}、S_{a2} …到 S_{aF*H}，统计 S_{a1}、S_{a2} …到 S_{aF*H} 中满足大于0的值的个数K，当K为偶数，则所述窗口 $W_{iz}[k_i-A_z,$

$k_i + B_z$]中至少部分数据满足所述预定条件C_z。

根据20至图33所示的本发明实施例提供的基于服务器查找数据流分割点的方法中，为潜在分割点k_i确定窗口W_{ix}[k_i-A_x，k_i+B_x]，其中，x分别为1到M连续的自然数，M≥2，可以并行判断M个窗口中每一个窗口中至少部分数据是否满足预定条件C_x，或者依次判断窗口中至少部分数据是否满足预定条件，也可以依次窗口W_{i1} [k_i-A_1，k_i+B_1]，中至少部分数据满足预定条件C_1时，再判断W_{i2}[k_i-A_2，k_i+B_2]中至少部分数据满足预定条件C_2时，直到判断W_{im} [k_i-A_m，k_i+B_m]中至少部分数据满足预定条件C_m。实施例中其他窗口的判断与此相同，不再赘述。

另外，根据20至图33所示的本发明实施例，在去重服务器103上预设有规则，所述规则：为潜在分割点k确定M个窗口W_x[k-A_x, k + B_x]和窗口W_x[k - A_x, k + B_x]对应的预定条件C_x，x分别为1到M连续的自然数，M≥2，在该预设规则中，A_1、A_2、A_3…A_m可以不全部相等，B_1、B_2、B_3…B_m可以不全部相等，C_1、C_2、C_3…C_M也可以不全部相同。在图21所示的实施方式中，在W_{i1}[k_i-169,k_i]、W_{i2}[k_i-170,k_i-1]、W_{i3}[k_i-171,k_i-2]、W_{i4}[k_i-172,k_i-3]、W_{i5}[k_i-173,k_i-4]、W_{i6}[k_i-174,k_i-5]、W_{i7}[k_i-175,k_i-6]、W_{i8}[k_i-176,k_i-7]、W_{i9}[k_i-177,k_i-8]、W_{i10}[k_i-178,k_i-9]和W_{i11}[k_i-179,k_i-10]中，各窗口大小相同,即窗口大小均为169字节，同时判断窗口中至少部分数据是否满足预定条件的方式也相同,具体见上述判断W_{i1}[k_i-169,k_i]中至少部分数据是否满足预定条件C_1的描述，但在图11所示的实施方式中，W_{i1}[k_i-169, k_i]、W_{i2}[k_i-170, k_i-1]、W_{i3}[k_i-171, k_i-2]、W_{i4}[k_i-172, k_i-3]、W_{i5}[k_i-173, k_i-4]、W_{i6}[k_i-174, k_i-5]、

$W_{i7}[k_i\text{-}175,\ k_i\text{-}6]$、$W_{i8}[k_i\text{-}176,\ k_i\text{-}7]$、$W_{i9}[k_i\text{-}177,\ k_i\text{-}8]$、$W_{i10}[k_i\text{-}168,\ k_i+1]$和$W_{i11}[k_i\text{-}179,\ k_i+3]$窗口大小可以不相同,同时判断窗口中至少部分数据是否满足预定条件的方式也可以不相同。在所有实施例中，根据为去重服务器103预设的规则，判断窗口W_{i1}中至少部分数据是否满足预定条件C_1的方式与判断窗口W_{j1}中至少部分数据是否满足预定条件C_1的方式必然相同，判断W_{i2}中至少部分数据是否满足预定条件C_2的方式与判断W_{j2}中至少部分数据是否满足预定条件C_2的方式必然相同...判断窗口W_{iM}中至少部分数据是否满足预定条件C_M的方式与判断窗口W_{jM}中至少部分数据是否满足预定条件C_M的方式必然相同。在此不再赘述。

根据20至图33所示的本发明实施例，在去重服务器103上预设有规则，k_a、k_i、k_j、k_l和k_m为沿着数据流分割点查找方向查找分割点时获得的潜在分割点，k_a、k_i、k_j、k_l和k_m都依据该规则。本发明实施例中的窗口$W_x[\text{k-}A_x, \text{k}+B_x]$表示一个特定范围，在该特定范围选择数据以判断这些数据是否满足预定条件C_x，具体地，可以在该特定范围内选择部分数据，也可以选择全部数据以判断这些数据是否满足预定条件C_x。本发明实施例中具体使用的窗口概念可参照窗口$W_x[\text{k-}A_x, \text{k}+B_x]$，在此不再赘述。

窗口$W_x[\text{k-}A_x, \text{k}+B_x]$中，（$\text{k}-A_x$）和（$\text{k}+B_x$）表示该窗口$W_x[\text{k-}A_x, \text{k}+B_x]$的两个边界，其中（$\text{k}-A_x$）表示窗口$W_x[\text{k-}A_x, \text{k}+B_x]$相对于潜在分割点k 位于数据流分割点查找反方向的边界，（$\text{k}+B_x$）表示窗口$W_x[\text{k-}A_x, \text{k}+B_x]$相对于潜在分割点 k 位于数据流分割点查找方向的边界。具体地，在本发明实施例中，在图20至图33所示的数据流分割点查找方向为从左向右，则其中（$\text{k}-A_x$）表示窗口$W_x[\text{k-}A_x, \text{k}+$

B_x]相对于潜在分割点k位于数据流分割点查找反方向的边界(即左边界），（k+B_x）表示窗口W_x[k-A_x, k+B_x]相对于潜在分割点k位于数据流分割点查找方向的边界（即右边界）。如果在图20至图33所示的数据流分割点查找方向为从右向左，则其中（k-A_x）表示窗口W_x[k-A_x, k+B_x]相对于潜在分割点k位于数据流分割点查找反方向的边界（即右边界），（k+B_x）表示窗口W_x[k-A_x, k+B_x]相对于潜在分割点k位于数据流分割点查找方向的边界（即左边界）。

本领域普通技术人员可以意识到，结合本发明实施例图20至图33描述的各示例的单元及算法步骤，本发明实施例的关键特征可以与其他技术相结合，以更为复杂的形式呈现，但仍会包含本发明的关键特征。在真实环境中可能使用备用分割点，例如一种实施方式为，根据为去重服务器103预设的规则，为潜在分割点k_i确定11个窗口W_x[k-A_x, k+B_x]及窗口W_x[k-A_x, k+B_x]对应的预定条件C_x，x为1到11连续的自然数，当11个窗口中每一个窗口W_x[k-A_x, k+B_x]中至少部分数据均满足预定条件C_x，则潜在分割点k_i为数据流分割点，当超过设定的最大数据块时，仍未查找到分割点，这时可能使用备用预设规则，备用的预设规则与在去重服务器103上预设的规则类似，备用的预设规则为：例如为潜在分割点k_i确定10个窗口W_x[k-A_x, k+B_x]及窗口W_x[k-A_x, k+B_x]对应的预定条件C_x，x为1到10连续的自然数，确定当10个窗口中每一个窗口W_x[k-A_x, k+B_x]中至少部分数据均满足预定条件C_x，则潜在分割点k_i为数据流分割点，当超过设定的最大数据块时，仍未查找到数据流分割点时，从最大数据块的结束位置作为强制分割点。

根据 20 至图 33 所示的本发明实施例，在去重服务器 103 上预设有规则，所述规则中为潜在分割点 k 确定 M 个窗口，并不一定要求先有一个潜在分割点 k，可以通过确定的 M 个窗口来判断

潜在分割点 k。

本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本发明的范围。

所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统、装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

在提供的几个实施例中，应该理解到，所公开的系统、方法，可以通过其它的方式实现。例如，以上所描述的装置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，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的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，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者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者全部单元来实现本实施例方案的目的。

另外，在本发明各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。

所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取非易失性存储介质

中。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案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非易失性存储介质中，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步骤。而前述的非易失性存储介质包括：U盘、移动硬盘、只读存储器（Read-Only Memory，ROM）、磁碟或者光盘等各种可以存储程序代码的介质。

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所述以权利要求的保护范围为准。

权利要求

1、一种基于服务器查找数据流分割点的方法，其特征在于:

在所述服务器上预设有规则，所述规则为：为潜在分割点k确定M个点p_x、点p_x对应的窗口$W_x[p_x-A_x, p_x+B_x]$和窗口$W_x[p_x-A_x, p_x+B_x]$对应的预定条件C_x，其中， x为1到M连续的自然数，M≥2，A_x和B_x为整数;

所述方法包括:

a)依据所述规则为当前潜在分割点k_i确定点p_{iz}及所述点p_{iz}对应的窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$，i和z为整数，并且1≤z≤M;

b) 判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足预定条件$C_z$;

当所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据不满足所述预定条件$C_z$，从所述点$p_{iz}$沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，N*U不大于$\| B_z \| + \max_x(\| A_x \| + \| (k_i - p_{ix}) \|)$，获得新的潜在分割点，执行步骤a)；

c)当所述当前潜在分割点k_i的M个窗口中的每一个窗口$W_{ix}[p_{ix}-A_x, p_{ix}+B_x]$中至少部分数据满足预定条件$C_x$，则所述当前潜在分割点$k_i$为数据流分割点。

2、根据权利要求1所述的方法，其特征在于，所述规则还包括:至少两个点p_e和p_f,满足条件$A_e=A_f$， $B_e=B_f$， $C_e=C_f$。

3、根据权利要求2所述的方法，其特征在于，所述规则还包括:所述至少两个点p_e和p_f,相对于所述潜在分割点k,在所述数据流分割点查找反方向上。

4、根据权利要求2或3所述的方法，其特征在于，所述规则还包

括：所述至少两个点p_e和p_f之间的距离为1个U。

5、根据权利要求1至4任一所述的方法，其特征在于，判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$，具体包括：

使用随机函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$。

6、根据权利要求5所述的方法，其特征在于，所述使用随机函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$，具体为使用hash函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$。

7、根据权利要求1至6任一所述的方法，其特征在于，当所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据不满足所述预定条件$C_z$，从所述点$p_{iz}$沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，获得所述新的潜在分割点，根据所述规则，为所述新的潜在分割点确定的点p_{ic}对应的窗口$W_{ic}[p_{ic}-A_c, p_{ic}+B_c]$的左边界与所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$的右边界重合或者为所述新的潜在分割点确定的所述点$p_{ic}$对应的所述窗口$W_{ic}[p_{ic}-A_c, p_{ic}+B_c]$的左边界位于所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$范围之内；其中，为所述新的潜在分割点确定的所述点$p_{ic}$是根据所述规则，为所述新的潜在分割点确定的M个点按照数据流查找方向获得的序列中排序第一的点。

8、根据权利要求5所述的方法，其特征在于，使用随机函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$，具体包括：

在所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中选择F个字节，将所述F个字节

反复利用H次，共获得F*H个字节，其中每个字节由8位组成，记为 $a_{m,1}\ldots a_{m,8}$，表示所述F*H个字节中第m个字节的第1到第8位，所述F*H个字节对应的位可以表示为：$\begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,8} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ a_{F*H,1} & a_{F*H,2} & \ldots & a_{F*H,8} \end{pmatrix}$，当 $a_{m,n}$=1时，$V_{am,n}$=1，当 $a_{m,n}$=0时，$V_{am,n}$=-1，其中 $a_{m,n}$ 表示 $a_{m,1}\ldots a_{m,8}$ 中的任一个，所述F*H个字节对应的位按照 $a_{m,n}$ 与 $V_{am,n}$ 的转换关系得到矩阵 V_a，所述矩阵 V_a 表示为：$\begin{pmatrix} V_{a1,1} & V_{a1,2} & \cdots & V_{a1,8} \\ V_{a2,1} & V_{a2,2} & \cdots & V_{a2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{aF*H,1} & V_{aF*H,2} & \cdots & V_{aF*H,8} \end{pmatrix}$，从服务正态分布的随机数中选择F*H*8个随机数组成矩阵R，所述矩阵R表示为：

$\begin{pmatrix} \mathrm{h}_{1,1} & h_{1,2} & \ldots & h_{1,8} \\ h_{2,1} & h_{2,2} & \ldots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{F*H,1} & h_{F*H,2} & \ldots & h_{F*H,8} \end{pmatrix}$，将所述矩阵 V_a 的第m行与所述矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为 $S_{am}=V_{am,1}*h_{m,1}+V_{am,2}*h_{m,2}+\ldots+V_{am,8}*h_{m,8}$，同理，获得 S_{a1}、S_{a2} …到 S_{aF*H}，统计 S_{a1}、S_{a2} …到 S_{aF*H} 中满足大于0的值的个数K，当K为偶数，则所述窗口 W_{iz} [$p_{iz}-A_z$，$p_{iz}+B_z$] 中至少部分数据满足所述预定条件 C_z。

9、一种基于服务器查找数据流分割点的方法，其特征在于，

在所述服务器上预设有规则，所述规则为：为潜在分割点 k 确定M个窗口 W_x [$k-A_x$，$k+B_x$] 和窗口 W_x [$k-A_x$，$k+B_x$] 对应的预定条件 C_x，其中，x为1到M连续的自然数，M≥2，A_x 和 B_x 为整数；

所述方法包括：

a）依据所述规则为当前潜在分割点 k_i 确定对应的窗口 W_{iz} [k_i-A_z，

$k_i+B_z]$，i和z为整数，并且1≤z≤M；

b) 判断所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据是否满足预定条件C_z；

当所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据不满足所述预定条件C_z，从所述当前潜在分割点k_i沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，N*U不大于‖B_z‖+$\max_x$(‖A_x‖)，获得新的潜在分割点，执行步骤a)；

c) 当所述当前潜在分割点k_i的M个窗口中的每一个窗口$W_{ix}[k_i-A_x, k_i+B_x]$中至少部分数据满足预定条件C_x，则所述当前潜在分割点k_i为数据流分割点。

10、根据权利要求9所述的方法，其特征在于，所述规则还包括：至少两个窗口$W_{ie}[k_i-A_e, k_i+B_e]$与$W_{if}[k_i-A_f, k_i+B_f]$，满足条件：$|A_e+B_e|=|A_f+B_f|$，$C_e=C_f$。

11、根据权利要求10所述的方法，其特征在于，所述规则还包括：A_e和A_f为正整数。

12、根据权利要求10或11所述的方法，其特征在于，所述规则还包括：A_e-1=A_f，B_e+1=B_f。

13、根据权利要求9至12任一所述的方法，其特征在于，判断所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据是否满足所述预定条件C_z，具体包括：

使用随机函数判断所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据是否满足所述预定条件C_z。

14、根据权利要求13所述的方法，其特征在于，所述使用随机函数判断所述$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据是否满足所述预定条

件C_z，具体为使用hash函数判断所述$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据是否满足所述预定条件C_z。

15、根据权利要求9至14任一所述的方法，其特征在于，当所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据不满足所述预定条件C_z，从所述当前潜在分割点k_i沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，获得所述新的潜在分割点，根据所述规则，为所述新的潜在分割点确定的窗口$W_{ic}[k_i-A_c, k_i+B_c]$的左边界与所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$的右边界重合或者为所述新的潜在分割点确定的所述窗口$W_{ic}[k_i-A_c, k_i+B_c]$的左边界位于所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$范围之内；其中，为所述新的潜在分割点确定的所述窗口$W_{ic}[k_i-A_c, k_i+B_c]$是根据所述规则，为所述新的潜在分割点确定的M个窗口按照数据流查找方向获得的序列中排序第一的窗口。

16、根据权利要求13所述的方法，其特征在于，使用随机函数判断所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据是否满足所述预定条件C_z，具体包括：

在所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中选择F个字节，将所述F个字节反复利用H次，共获得F*H个字节，其中每个字节由8位组成，记为$a_{m,1}\ldots a_{m,8}$，表示所述F*H个字节中第m个字节的第1到第8位，所述F*H个字节对应的位可以表示为：$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,8} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ a_{F*H,1} & a_{F*H,2} & \cdots & a_{F*H,8} \end{pmatrix}$，当$a_{m,n}$=1时，$V_{am,n}$=1，当$a_{m,n}$=0时，$V_{am,n}$=-1，其中$a_{m,n}$表示$a_{m,1}\ldots a_{m,8}$中的任一个，所述F*H个字节对应的位按照$a_{m,n}$与$V_{am,n}$的转换关系得到矩阵$V_a$，所述矩阵

V_a表示为：$\begin{pmatrix} V_{a1,1} & V_{a1,2} & \cdots & V_{a1,8} \\ V_{a2,1} & V_{a2,2} & \cdots & V_{a2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{aF*H,1} & V_{aF*H,2} & \cdots & V_{aF*H,8} \end{pmatrix}$，从服务正态分布的随机数中选择F*H*8个随机数组成矩阵R，所述矩阵R表示为：

$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,8} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{F*H,1} & h_{F*H,2} & \cdots & h_{F*H,8} \end{pmatrix}$，将所述矩阵$V_a$的第m行与所述矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为$S_{am}=V_{am,1}*h_{m,1}+V_{am,2}*h_{m,2}+\ldots+V_{am,8}*h_{m,8}$，同理，获得$S_{a1}$、$S_{a2}$...到$S_{aF*H}$，统计$S_{a1}$、$S_{a2}$...到$S_{aF*H}$中满足大于0的值的个数K，当K为偶数，则所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据满足所述预定条件$C_z$。

17、一种用于查找数据流分割点的服务器，其特征在于，所述服务器包括中央处理单元和主存储器，所述中央处理单元与所述主存储器通信，在所述服务器上预设有规则，所述规则为：为潜在分割点k确定M个点p_x、点p_x对应的窗口$W_x[p_x-A_x, p_x+B_x]$和窗口$W_x[p_x-A_x, p_x+B_x]$对应的预定条件C_x，其中，x为1到M连续的自然数，M≥2，A_x和B_x为整数；

所述主存储器用于存储可执行指令，所述中央处理单元执行所述可执行指令，以执行以下步骤：

a)依据所述规则为当前潜在分割点k_i确定点p_{iz}及所述点p_{iz}对应的窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$，i和z为整数，并且1≤z≤M；

b）判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足预定条件$C_z$；

当所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据不满足所述预定条件$C_z$，从所述点$p_{iz}$沿所述数据流分割点查找方向跳跃N个数据流分

割点最小查找单位U，N*U不大于 $\| B_z \| + \max_x (\| A_x \| + \| (k_i - p_{ix}) \|)$，获得新的潜在分割点，执行步骤a)；

c)当所述当前潜在分割点k_i的M个窗口中的每一个窗口$W_{ix}[p_{ix}-A_x, p_{ix}+B_x]$中至少部分数据满足预定条件$C_x$，则所述当前潜在分割点$k_i$为数据流分割点。

18、根据权利要求17所述服务器，其特征在于，所述规则还包括：至少两个点p_e和p_f，满足条件$A_e=A_f$，$B_e=B_f$，$C_e=C_f$。

19、根据权利要求18所述服务器，其特征在于，所述规则还包括：所述至少两个点p_e和p_f，相对于所述潜在分割点k，在所述数据流分割点查找反方向上。

20、根据权利要求18或19所述的服务器，其特征在于，所述规则还包括：所述至少两个点p_e和p_f之间的距离为1个U。

21、根据权利要求17至20任一所述的服务器，其特征在于，所述中央处理单元具体用于使用随机函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$。

22、根据权利要求21所述的服务器，其特征在于，所述中央处理单元具体用于使用hash函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$。

23、根据权利要求17至22任一所述的服务器，其特征在于，当所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据不满足所述预定条件$C_z$，从所述点$p_{iz}$沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，获得所述新的潜在分割点，根据所述规则，为所述新的潜在分割点确定的点p_{ic}对应的窗口$W_{ic}[p_{ic}-A_c, p_{ic}+B_c]$的左边界与所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$的右边界重合或者为所述新的潜在分割点确定的所述点$p_{ic}$对应的所述窗口$W_{ic}[p_{ic}-A_c, p_{ic}+B_c]$的左边界位于所述

窗口 $W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$范围之内；其中，为所述新的潜在分割点确定的所述点 p_{ic} 是根据所述规则，为所述新的潜在分割点确定的M个点按照数据流查找方向获得的序列中排序第一的点。

24、根据权利要求21所述的服务器，其特征在于，所述中央处理单元使用随机函数判断所述窗口 $W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$ 中至少部分数据是否满足所述预定条件 C_z，具体包括：

在所述窗口 $W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$ 中选择F个字节，将所述F个字节反复利用H次，共获得F*H个字节，其中每个字节由8位组成，记为 $a_{m,1}\ldots a_{m,8}$，表示所述F*H个字节中第m个字节的第1到第8位，所述F*H个字节对应的位可以表示为：$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,8} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ a_{F*H,1} & a_{F*H,2} & \cdots & a_{F*H,8} \end{pmatrix}$，当 $a_{m,n}$=1时，$V_{am,n}$=1，当 $a_{m,n}$=0时，$V_{am,n}$=-1，其中 $a_{m,n}$ 表示 $a_{m,1}\ldots a_{m,8}$ 中的任一个，所述F*H个字节对应的位按照 $a_{m,n}$ 与 $V_{am,n}$ 的转换关系得到矩阵 V_a，所述矩阵 V_a 表示为：$\begin{pmatrix} V_{a1,1} & V_{a1,2} & \cdots & V_{a1,8} \\ V_{a2,1} & V_{a2,2} & \cdots & V_{a2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{aF*H,1} & V_{aF*H,2} & \cdots & V_{aF*H,8} \end{pmatrix}$，从服务正态分布的随机数中选择F*H*8个随机数组成矩阵R，所述矩阵R表示为：

$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,8} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{F*H,1} & h_{F*H,2} & \cdots & h_{F*H,8} \end{pmatrix}$，将所述矩阵 V_a 的第m行与所述矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为 $S_{am}=V_{am,1}*h_{m,1}+V_{am,2}*h_{m,2}+\ldots+V_{am,8}*h_{m,8}$，同理，获得 S_{a1}、$S_{a2}\ldots$到 S_{aF*H}，统计 S_{a1}、$S_{a2}\ldots$到 S_{aF*H} 中满足大于0的值的个数K，当K为偶数，则所述窗口 $W_{iz}[p_{iz}-A_z,$

$p_{iz}+B_z]$中至少部分数据满足所述预定条件C_z。

25、一种用于查找数据流分割点的服务器，其特征在于，所述服务器包括中央处理单元和主存储器，所述中央处理单元与所述主存储器通信，在所述服务器上预设有规则，所述规则为：为潜在分割点k确定M个窗口$W_x[k-A_x, k+B_x]$和窗口$W_x[k-A_x, k+B_x]$对应的预定条件C_x，其中，x为1到M连续的自然数，M≥2，A_x和B_x为整数；

所述主存储器用于存储可执行指令，所述中央处理单元执行所述可执行指令，以执行以下步骤：

a)依据所述规则为当前潜在分割点k_i确定对应的窗口$W_{iz}[k_i-A_z, k_i+B_z]$，i和z为整数，并且1≤z≤M；

b)判断所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据是否满足预定条件C_z；

当所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据不满足所述预定条件C_z，从所述当前潜在分割点k_i沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，N*U不大于$\| B_z \|+\max_x(\| A_x \|)$，获得新的潜在分割点，执行步骤a）；

c)当所述当前潜在分割点k_i的M个窗口中的每一个窗口$W_{ix}[k_i-A_x, k_i+B_x]$中至少部分数据满足预定条件C_x，则所述当前潜在分割点k_i为数据流分割点。

26、根据权利要求25所述的服务器，其特征在于，所述规则还包括：至少两个窗口$W_{ie}[k_i-A_e, k_i+B_e]$与$W_{if}[k_i-A_f, k_i+B_f]$，满足条件：$|A_e+B_e|=|A_f+B_f|$，$C_e=C_f$。

27、根据权利要求26所述的服务器，其特征在于，为所述服务器预设规则，所述规则还包括：A_e和A_f为正整数。

28、根据权利要求26或27所述的服务器，其特征在于，所述规则

还包括： A_e-1=A_f， B_e+1=B_f。

29、根据权利要求25至28任一所述的服务器，其特征在于，所述中央处理单元具体用于使用随机函数判断所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据是否满足所述预定条件C_z。

30、根据权利要求29所述的服务器，其特征在于，所述中央处理单元具体用于使用hash函数判断所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据是否满足所述预定条件C_z。

31、根据权利要求25至30任一所述的服务器，其特征在于，当所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据不满足所述预定条件C_z，从所述当前潜在分割点k_i沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，获得所述新的潜在分割点，根据所述规则，为所述新的潜在分割点确定的窗口$W_{ic}[k_i-A_c, k_i+B_c]$的左边界与所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$的右边界重合或者为所述新的潜在分割点确定的所述窗口$W_{ic}[k_i-A_c, k_i+B_c]$的左边界位于所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$范围之内；其中，为所述新的潜在分割点确定的所述窗口$W_{ic}[k_i-A_c, k_i+B_c]$是根据所述规则，为所述新的潜在分割点确定的M个窗口按照数据流查找方向获得的序列中排序第一的窗口。

32、根据权利要求29所述的服务器，其特征在于，所述中央处理单元使用随机函数判断所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据是否满足所述预定条件C_z，具体包括：

在所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中选择F个字节，将所述F个字节反复利用H次，共获得F*H个字节，其中每个字节由8位组成，记为$a_{m,1}\ldots a_{m,8}$，表示所述F*H个字节中第m个字节的第1到第8位，所述F*H个字

节对应的位可以表示为：$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,8} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ a_{F*H,1} & a_{F*H,2} & \cdots & a_{F*H,8} \end{pmatrix}$，当 $a_{m,n}$=1时，$V_{am,n}$ =1，当 $a_{m,n}$=0时，$V_{am,n}$=-1，其中 $a_{m,n}$ 表示 $a_{m,1}\ldots a_{m,8}$ 中的任一个，所述F*H个字节对应的位按照 $a_{m,n}$ 与 $V_{am,n}$ 的转换关系得到矩阵 V_a，所述矩阵 V_a 表示为：$\begin{pmatrix} V_{a1,1} & V_{a1,2} & \cdots & V_{a1,8} \\ V_{a2,1} & V_{a2,2} & \cdots & V_{a2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{aF*H,1} & V_{aF*H,2} & \cdots & V_{aF*H,8} \end{pmatrix}$，从服务正态分布的随机数中选择F*H*8个随机数组成矩阵R，所述矩阵R表示为：

$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,8} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{F*H,1} & h_{F*H,2} & \cdots & h_{F*H,8} \end{pmatrix}$，将所述矩阵 V_a 的第m行与所述矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为 $S_{am}=V_{am,1}*h_{m,1}+V_{am,2}*h_{m,2}+\ldots+V_{am,8}*h_{m,8}$，同理，获得 S_{a1}、S_{a2} …到 S_{aF*H}，统计 S_{a1}、S_{a2} …到 S_{aF*H} 中满足大于0的值的个数K，当K为偶数，则所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 中至少部分数据满足所述预定条件 C_z。

33、一种用于查找数据流分割点的服务器，其特征在于，在所述服务器上预设有规则，所述规则为：为潜在分割点 k 确定M个点 p_x、点 p_x 对应的窗口 $W_x[p_x-A_x, p_x+B_x]$ 和窗口 $W_x[p_x-A_x, p_x+B_x]$ 对应的预定条件 C_x，其中，x为1到M连续的自然数，M≥2，A_x 和 B_x 为整数；

所述服务器包括：确定单元，用于执行步骤a）：a)依据所述规则为当前潜在分割点 k_i 确定点 p_{iz} 及所述点 p_{iz} 对应的窗口 $W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$，i和z为整数，并且1≤z≤M；

判断处理单元，用于判断所述窗口 $W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$ 中至少部分数据是否满足预定条件 C_z；

当所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据不满足所述预定条件$C_z$，从所述点$p_{iz}$沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，N*U不大于$\| B_z \| + \max_x(\| A_x \| + \| (k_i - p_{ix}) \|)$，获得新的潜在分割点，则所述确定单元为所述新的潜在分割点执行步骤a)；

当所述当前潜在分割点k_i的M个窗口中的每一个窗口$W_{ix}[p_{ix}-A_x, p_{ix}+B_x]$中至少部分数据满足预定条件$C_x$，则所述当前潜在分割点$k_i$为数据流分割点。

34、根据权利要求33所述的服务器，其特征在于，所述规则还包括：至少两个点p_e和p_f，满足条件$A_e=A_f$，$B_e=B_f$，$C_e=C_f$。

35、根据权利要求34所述的服务器，其特征在于，所述规则还包括：所述至少两个点p_e和p_f，相对于所述潜在分割点k，在所述数据流分割点查找反方向上。

36、根据权利要求34或35所述的服务器，其特征在于，所述规则还包括：所述至少两个点p_e和p_f之间的距离为1个U。

37、根据权利要求33至36任一所述的服务器，其特征在于，所述判断处理单元具体用于使用随机函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$。

38、根据权利要求37所述的服务器，其特征在于，所述判断处理单元具体用于使用hash函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$。

39、根据权利要求33至38任一所述的服务器，其特征在于，所述判断处理单元用于当所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据不满足所述预定条件$C_z$，从所述点$p_{iz}$沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，获得所述新的潜在分割点，所述确

定单元为所述新的潜在分割点执行步骤a），根据所述规则，为所述新的潜在分割点确定的点p_{ic}对应的窗口$W_{ic}[p_{ic}-A_c, p_{ic}+B_c]$的左边界与所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$的右边界重合或者为所述新的潜在分割点确定的所述窗口$W_{ic}[p_{ic}-A_c, p_{ic}+B_c]$的左边界位于所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$范围之内；其中，为所述新的潜在分割点确定的所述窗口$W_{ic}[p_{ic}-A_c, p_{ic}+B_c]$是根据所述规则，为所述新的潜在分割点确定的M个点按照数据流查找方向获得的序列中排序第一的点。

40、根据权利要求37所述的服务器，其特征在于，所述判断处理单元具体用于使用随机函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$，具体包括：

在所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中选择F个字节，将所述F个字节反复利用H次，共获得F*H个字节，其中每个字节由8位组成，记为$a_{m,1}\ldots a_{m,8}$，表示所述F*H个字节中第m个字节的第1到第8位，所述F*H个字节对应的位可以表示为：$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,8} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ a_{F*H,1} & a_{F*H,2} & \cdots & a_{F*H,8} \end{pmatrix}$，当$a_{m,n}$=1时，$V_{am,n}$=1，当$a_{m,n}$=0时，$V_{am,n}$=-1，其中$a_{m,n}$表示$a_{m,1}\ldots a_{m,8}$中的任一个，所述F*H个字节对应的位按照$a_{m,n}$与$V_{am,n}$的转换关系得到矩阵$V_a$，所述矩阵$V_a$表示为：$\begin{pmatrix} V_{a1,1} & V_{a1,2} & \cdots & V_{a1,8} \\ V_{a2,1} & V_{a2,2} & \cdots & V_{a2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{aF*H,1} & V_{aF*H,2} & \cdots & V_{aF*H,8} \end{pmatrix}$，从服务正态分布的随机数中选择F*H*8个随机数组成矩阵R，所述矩阵R表示为：

$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,8} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{F*H,1} & h_{F*H,2} & \cdots & h_{F*H,8} \end{pmatrix}$，将所述矩阵$V_a$的第m行与所述矩阵R的第m行

的随机数相乘，然后求和得到一个值，具体表示为 $S_{am}=V_{am,1}*h_{m,1}+V_{am,2}*h_{m,2}+\ldots+V_{am,8}*h_{m,8}$，同理，获得 S_{a1}、S_{a2} …到 S_{aF*H}，统计 S_{a1}、S_{a2} …到 S_{aF*H} 中满足大于0的值的个数K，当K为偶数，则所述窗口 $W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$ 中至少部分数据满足所述预定条件 C_z。

41、一种用于查找数据流分割点的服务器，其特征在于，在所述服务器上预设有规则，所述规则为：为潜在分割点 k 确定M个窗口 $W_x[k-A_x, k+B_x]$ 和窗口 $W_x[k-A_x, k+B_x]$ 对应的预定条件 C_x，其中，x为1到M连续的自然数，M≥2，A_x 和 B_x 为整数；

所述服务器包括：确定单元，用于执行步骤a）：

a)依据所述规则为当前潜在分割点 k_i 确定对应的窗口 $W_{iz}[k_i-A_z, k_i+B_z]$，i和z为整数，并且1≤z≤M；

判断处理单元，用于判断所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 中至少部分数据是否满足预定条件 C_z；

当所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 中至少部分数据不满足所述预定条件 C_z，从所述当前潜在分割点 k_i 沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，N*U不大于 $\| B_z \|+\max_x(\| A_x \|)$，获得新的潜在分割点，则所述确定单元为所述新的潜在分割点执行步骤a）；

当所述当前潜在分割点 k_i 的M个窗口中的每一个窗口 $W_{ix}[k_i-A_x, k_i+B_x]$ 中至少部分数据满足预定条件 C_x，则所述当前潜在分割点 k_i 为数据流分割点。

42、根据权利要求41所述的服务器，其特征在于，所述规则还包括：至少两个窗口 $W_{ie}[k_i-A_e, k_i+B_e]$ 与 $W_{if}[k_i-A_f, k_i+B_f]$，满足条件：$|A_e+B_e|=|A_f+B_f|$，$C_e=C_f$。

43、根据权利要求42所述的服务器，其特征在于，所述规则还包括：A_e和A_f为正整数。

44、根据权利要求42或43所述的服务器，其特征在于，所述规则还包括：A_e-1=A_f，B_e+1=B_f。

45、根据权利要求41至44任一所述的服务器，其特征在于，所述判断处理单元具体用于

使用随机函数判断所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据是否满足所述预定条件C_z。

46、根据权利要求45所述的服务器，其特征在于，所述判断处理单元具体使用hash函数判断所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据是否满足所述预定条件C_z。

47、根据权利要求41至46任一所述的服务器，其特征在于，所述判断处理单元用于当所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据不满足所述预定条件C_z，从所述当前潜在分割点k_i沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，获得所述新的潜在分割点，所述确定单元为所述新的潜在分割点执行步骤a），根据所述规则，为所述新的潜在分割点确定的窗口$W_{ic}[k_i-A_c, k_i+B_c]$的左边界与所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$的右边界重合或者为所述新的潜在分割点确定的所述窗口$W_{ic}[k_i-A_c, k_i+B_c]$的左边界位于所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$范围之内；其中，为所述新的潜在分割点确定的所述窗口$W_{ic}[k_i-A_c, k_i+B_c]$是根据所述规则，为所述新的潜在分割点确定的M个窗口按照数据流查找方向获得的序列中排序第一的窗口。

48、根据权利要求46所述的服务器，其特征在于，所述判断处理单元使用随机函数判断所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据是否满足所述预定条件C_z，具体包括：

在所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 中选择F个字节，将所述F个字节反复利用H次，共获得F*H个字节，其中每个字节由8位组成，记为 $a_{m,1}\ldots a_{m,8}$，表示所述F*H个字节中第m个字节的第1到第8位，所述F*H个字节对应的位可以表示为：$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,8} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ a_{F*H,1} & a_{F*H,2} & \cdots & a_{F*H,8} \end{pmatrix}$，当 $a_{m,n}$=1时，$V_{am,n}$=1，当 $a_{m,n}$=0时，$V_{am,n}$=-1，其中 $a_{m,n}$ 表示 $a_{m,1}\ldots a_{m,8}$ 中的任一个，所述F*H个字节对应的位按照 $a_{m,n}$ 与 $V_{am,n}$ 的转换关系得到矩阵 V_a，所述矩阵 V_a 表示为：$\begin{pmatrix} V_{a1,1} & V_{a1,2} & \cdots & V_{a1,8} \\ V_{a2,1} & V_{a2,2} & \cdots & V_{a2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{aF*H,1} & V_{aF*H,2} & \cdots & V_{aF*H,8} \end{pmatrix}$，从服务正态分布的随机数中选择F*H*8个随机数组成矩阵R，所述矩阵R表示为：

$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,8} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{F*H,1} & h_{F*H,2} & \cdots & h_{F*H,8} \end{pmatrix}$，将所述矩阵 V_a 的第m行与所述矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为 $S_{am}=V_{am,1}*h_{m,1}+V_{am,2}*h_{m,2}+\ldots+V_{am,8}*h_{m,8}$，同理，获得 S_{a1}、S_{a2} …到 S_{aF*H}，统计 S_{a1}、S_{a2} …到 S_{aF*H} 中满足大于0的值的个数K，当K为偶数，则所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 中至少部分数据满足所述预定条件 C_z。

49、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用于存储可执行指令，服务器执行所述可执行指令以查找数据流分割点，在所述服务器上预设有规则，所述规则为：为潜在分割点 k 确定M个点 p_x、点 p_x 对应的窗口 $W_x[p_x-A_x, p_x+B_x]$ 和窗口 $W_x[p_x-A_x, p_x+B_x]$ 对应的预定条件 C_x，其中，x为1到M连续的自然数，M≥2，

A_x和B_x为整数;

当所述服务器执行所述可执行指令，以执行以下步骤:

a)依据所述规则为当前潜在分割点k_i确定p_{iz}及p_{iz}对应的窗口W_{iz}[$p_{iz}-A_z$，$p_{iz}+B_z$]，i和z为整数，并且1≤z≤M;

b) 判断所述窗口W_{iz}[$p_{iz}-A_z$，$p_{iz}+B_z$]中至少部分数据是否满足预定条件C_z;

当所述窗口W_{iz}[$p_{iz}-A_z$，$p_{iz}+B_z$]中至少部分数据不满足所述预定条件C_z，从所述点p_{iz}沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，N*U不大于$\| B_z \| + \max_x(\| A_x \| + \| (k_i - p_{ix}) \|)$，获得新的潜在分割点，执行步骤a)；

c)当所述当前潜在分割点k_i的M个窗口中的每一个窗口W_{ix}[$p_{ix}-A_x$，$p_{ix}+B_x$]中至少部分数据满足预定条件C_x，则所述当前潜在分割点k_i为数据流分割点。

50、根据权利要求49所述的计算机可读存储介质，其特征在于，所述规则还包括：至少两个点p_e和p_f，满足条件$A_e=A_f$，$B_e=B_f$，$C_e=C_f$。

51、根据权利要求50所述的计算机可读存储介质，其特征在于，所述规则还包括:所述至少两个点p_e和p_f，相对于所述潜在分割点k，在所述数据流分割点查找反方向上。

52、根据权利要求50或51所述的计算机可读存储介质，其特征在于，所述规则还包括：所述至少两个点p_e和p_f之间的距离为1个U。

53、根据权利要求49至52任一所述的计算机可读存储介质，其特征在于，所述服务器判断所述窗口W_{iz}[$p_{iz}-A_z$，$p_{iz}+B_z$]中至少部分数据是否满足所述预定条件C_z，具体包括:

所述服务器使用随机函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$。

54、根据权利要求53所述的计算机可读存储介质，其特征在于，所述服务器使用随机函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$，具体包括:

所述服务器使用hash函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$。

55、根据权利要求49至54任一所述的计算机可读存储介质，其特征在于，当所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据不满足所述预定条件$C_z$，从所述点$p_{iz}$沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，获得所述新的潜在分割点，根据所述规则，为所述新的潜在分割点确定的点p_{ic}对应的窗口$W_{ic}[p_{ic}-A_c, p_{ic}+B_c]$的左边界与所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$的右边界重合或者为所述新的潜在分割点确定的所述点$p_{ic}$对应的所述窗口$W_{ic}[p_{ic}-A_c, p_{ic}+B_c]$的左边界位于所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$范围之内；其中，为所述新的潜在分割点确定的所述点$p_{ic}$是根据所述规则，为所述新的潜在分割点确定的M个点按照数据流查找方向获得的序列中排序第一的点。

56、根据权利要求53所述的计算机可读存储介质，其特征在于，使用随机函数判断所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中至少部分数据是否满足所述预定条件$C_z$，具体包括:

在所述窗口$W_{iz}[p_{iz}-A_z, p_{iz}+B_z]$中选择F个字节，将所述F个字节反复利用H次，共获得F*H个字节，其中每个字节由8位组成，记为$a_{m,1}\ldots a_{m,8}$，表示所述F*H个字节中第m个字节的第1到第8位，所述F*H个字

节对应的位可以表示为：$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,8} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ a_{F*H,1} & a_{F*H,2} & \cdots & a_{F*H,8} \end{pmatrix}$，当 $a_{m,n}$=1时，$V_{am,n}$ =1，当 $a_{m,n}$=0时，$V_{am,n}$=-1，其中 $a_{m,n}$ 表示 $a_{m,1}\ldots a_{m,8}$ 中的任一个，所述F*H个字节对应的位按照 $a_{m,n}$ 与 $V_{am,n}$ 的转换关系得到矩阵 V_a，所述矩阵 V_a 表示为：$\begin{pmatrix} V_{a1,1} & V_{a1,2} & \cdots & V_{a1,8} \\ V_{a2,1} & V_{a2,2} & \cdots & V_{a2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{aF*H,1} & V_{aF*H,2} & \cdots & V_{aF*H,8} \end{pmatrix}$，从服务正态分布的随机数中选择F*H*8个随机数组成矩阵R，所述矩阵R表示为：

$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,8} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{F*H,1} & h_{F*H,2} & \cdots & h_{F*H,8} \end{pmatrix}$，将所述矩阵 V_a 的第m行与所述矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为 $S_{am}=V_{am,1}*h_{m,1}+V_{am,2}*h_{m,2}+\ldots+V_{am,8}*h_{m,8}$，同理，获得 S_{a1}、S_{a2} …到 S_{aF*H}，统计 S_{a1}、S_{a2} …到 S_{aF*H} 中满足大于0的值的个数K，当K为偶数，则所述窗口 W_{iz} [$p_{iz}-A_z$，$p_{iz}+B_z$] 中至少部分数据满足所述预定条件 C_z。

57、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用于存储可执行指令，服务器执行所述可执行指令以查找数据流分割点，在所述服务器上预设有规则，所述规则为：为潜在分割点 k 确定M个窗口 W_x [$k-A_x$，$k+B_x$] 和窗口 W_x [$k-A_x$，$k+B_x$] 对应的预定条件 C_x，其中，x为1到M连续的自然数，M≥2，A_x 和 B_x 为整数；

当所述服务器执行所述可执行指令，以执行以下步骤：

a) 依据所述规则为当前潜在分割点 k_i 确定对应的窗口 W_{iz} [k_i-A_z，k_i+B_z]，i和z为整数，并且1≤z≤M；

b) 判断所述窗口 W_{iz} [k_i-A_z，k_i+B_z] 中至少部分数据是否满足预

定条件C_z；

当所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据不满足所述预定条件C_z，从所述当前潜在分割点k_i沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，N*U不大于$\| B_z \| + \max_x(\| A_x \|)$，获得新的潜在分割点，执行步骤a）；

c) 当所述当前潜在分割点k_i的M个窗口中的每一个窗口$W_{ix}[k_i-A_x, k_i+B_x]$中至少部分数据满足预定条件C_x，则所述当前潜在分割点k_i为数据流分割点。

58、根据权利要求57所述的计算机可读存储介质，其特征在于，所述规则还包括：至少两个窗口$W_{ie}[k_i-A_e, k_i+B_e]$与$W_{if}[k_i-A_f, k_i+B_f]$，满足条件：$|A_e+B_e|=|A_f+B_f|$，$C_e=C_f$。

59、根据权利要求58所述的计算机可读存储介质，其特征在于，所述规则还包括：A_e和A_f为正整数。

60、根据权利要求58或59所述的计算机可读存储介质，其特征在于，所述规则还包括：A_e-1=A_f，B_e+1=B_f。

61、根据权利要求57至60任一所述的计算机可读存储介质，其特征在于，

所述服务器判断所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据是否满足所述预定条件C_z，具体包括：

使用随机函数判断所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据是否满足所述预定条件C_z。

62、根据权利要求61所述的计算机可读存储介质，其特征在于，所述服务器使用随机函数判断所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据是否满足所述预定条件C_z，具体为所述服务器使用hash函数判

断所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 中至少部分数据是否满足所述预定条件 C_z。

63、根据权利要求57至62任一所述的计算机可读存储介质，其特征在于，当所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 中至少部分数据不满足所述预定条件 C_z，从所述当前潜在分割点 k_i 沿所述数据流分割点查找方向跳跃N个数据流分割点最小查找单位U，获得所述新的潜在分割点，根据所述规则，为所述新的潜在分割点确定的窗口 $W_{ic}[k_i-A_c, k_i+B_c]$ 的左边界与所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 的右边界重合或者为所述新的潜在分割点确定的所述窗口 $W_{ic}[k_i-A_c, k_i+B_c]$ 的左边界位于所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 范围之内；其中，为所述新的潜在分割点确定的所述窗口 $W_{ic}[k_i-A_c, k_i+B_c]$ 是根据所述规则，为所述新的潜在分割点确定的M个窗口按照数据流查找方向获得的序列中排序第一的窗口。

64、根据权利要求61所述的计算机可读存储介质，其特征在于，使用随机函数判断所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 中至少部分数据是否满足所述预定条件 C_z，具体包括:

在所述窗口 $W_{iz}[k_i-A_z, k_i+B_z]$ 中选择F个字节，将所述F个字节反复利用H次，共获得F*H个字节，其中每个字节由8位组成，记为 $a_{m,1}\ldots a_{m,8}$，表示所述F*H个字节中第m个字节的第1到第8位，所述F*H个字节对应的位可以表示为: $\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,8} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ a_{F*H,1} & a_{F*H,2} & \cdots & a_{F*H,8} \end{pmatrix}$，当 $a_{m,n}$=1时，$V_{am,n}$=1，当 $a_{m,n}$=0时，$V_{am,n}$=-1，其中 $a_{m,n}$ 表示 $a_{m,1}\ldots a_{m,8}$ 中的任一个，所述F*H个字节对应的位按照 $a_{m,n}$ 与 $V_{am,n}$ 的转换关系得到矩阵 V_a，所述矩阵

V_a表示为：$\begin{pmatrix} V_{a1,1} & V_{a1,2} & \cdots & V_{a1,8} \\ V_{a2,1} & V_{a2,2} & \cdots & V_{a2,8} \\ \vdots & \vdots & \vdots & \vdots \\ V_{aF*H,1} & V_{aF*H,2} & \cdots & V_{aF*H,8} \end{pmatrix}$，从服务正态分布的随机数中选择F*H*8个随机数组成矩阵R，所述矩阵R表示为：

$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,8} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,8} \\ \vdots & \vdots & \vdots & \vdots \\ h_{F*H,1} & h_{F*H,2} & \cdots & h_{F*H,8} \end{pmatrix}$，将所述矩阵$V_a$的第m行与所述矩阵R的第m行的随机数相乘，然后求和得到一个值，具体表示为$S_{am}=V_{am,1}*h_{m,1}+V_{am,2}*h_{m,2}+\ldots+V_{am,8}*h_{m,8}$，同理，获得$S_{a1}$、$S_{a2}$...到$S_{aF*H}$，统计$S_{a1}$、$S_{a2}$...到$S_{aF*H}$中满足大于0的值的个数K，当K为偶数，则所述窗口$W_{iz}[k_i-A_z, k_i+B_z]$中至少部分数据满足所述预定条件$C_z$。