

[19] 中华人民共和国国家知识产权局



[12] 发明专利申请公开说明书

[21] 申请号 200380107207.9

[51] Int. Cl.

H04K 1/00 (2006.01)

H04L 9/00 (2006.01)

H04L 9/32 (2006.01)

G06F 11/30 (2006.01)

G06F 12/14 (2006.01)

[43] 公开日 2006 年 4 月 5 日

[11] 公开号 CN 1757188A

[22] 申请日 2003.11.6

[21] 申请号 200380107207.9

[30] 优先权

[32] 2002.11.6 [33] US [31] 60/424,240

[86] 国际申请 PCT/US2003/035607 2003.11.6

[87] 国际公布 WO2004/045123 英 2004.5.27

[85] 进入国家阶段日期 2005.6.21

[71] 申请人 国际商业机器公司

地址 美国纽约

[72] 发明人 杰弗里·詹姆斯·乔纳斯

[74] 专利代理机构 中国国际贸易促进委员会专利商标事务所

代理人 李德山

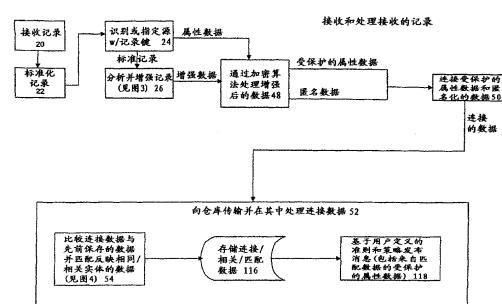
权利要求书 15 页 说明书 12 页 附图 7 页

[54] 发明名称

机密数据共享和匿名实体解析度

[57] 摘要

公开了用于处理数据的方法、程序和系统。数据包括多个实体的标识符。该方法、程序和系统包括步骤：(a)接收一条或多条记录，每条记录都具有多个标识符，每条记录对应于至少一个实体 [20]，(b)利用加密算法处理记录中多个标识符中的至少两个 [48]，(c)在将处理过的记录发送到独立的系统或数据库以后的某时，比较处理过的记录与先前存储的数据 [54]，(d)匹配处理过的记录与确定为反映实体的先前存储的数据，确定为反映实体的先前存储的数据包括至少两个确定为反映实体的先前存储的数据的至少一部分，先前存储的数据包括至少两条先前接收的记录的至少一部分，和/或基于其它标识符 [54]；和/或(e)关联处理过的记录与确定为反映与实体关系的先前存储的数据 [54]。



1、一种处理数据的方法，包括步骤：

接收具有多个标识符的记录，该记录对应于实体；

5 利用加密算法处理记录中多个标识符中的至少两个，以形成处理过的记录；

比较处理过的记录与先前存储的数据；及

匹配处理过的记录与确定为反映实体的先前存储的数据，确定为反映实体的先前存储的数据包括至少两个先前接收的记录的至少一部分。

10 2、如权利要求 1 所述的方法，还包括向记录指定或识别源的步骤。

15 3、如权利要求 2 所述的方法，还包括在匹配处理过的记录与确定为反映实体的先前存储的数据的步骤之后解译处理过的记录的至少一部分的步骤。

4、如权利要求 3 所述的方法，其中解译后的处理过的记录的至少一部分是源。

5、如权利要求 1 所述的方法，还包括向记录加盐的步骤。

6、如权利要求 1 所述的方法，还包括接收多个所接收的记录。

20 7、如权利要求 1 所述的方法，还包括在使用加密算法处理记录的至少一部分的步骤之前分析记录的步骤。

8、如权利要求 7 所述的方法，其中分析记录的步骤包括比较多多个标识符中的至少一个与以下中的一个的步骤：

用户定义的准则；及

二级数据库和列表中的一个中的数据集。

25 9、如权利要求 8 所述的方法，还包括增强记录的步骤。

10、如权利要求 9 所述的方法，其中增强记录的步骤包括根据用户定义的标准格式化多个标识符中的至少一个。

11、如权利要求 9 所述的方法，其中增强记录的步骤包括对多个

标识符中的至少一个产生变体并将该变体包括在记录中。

12、如权利要求 9 所述的方法，其中增强记录的步骤包括：

查询数据集寻找关于记录的二级标识符；及

用该二级标识符补充记录。

5 13、如权利要求 1 所述的方法，还包括将记录转换成标准化消息格式的步骤。

14、如权利要求 13 所述的方法，其中将记录转换成标准化消息格式的步骤包括利用对应于多个标识符中每个的类型指示器的步骤。

10 15、如权利要求 14 所述的方法，其中在使用加密算法的步骤之后，类型指示器是可辨别的。

16、如权利要求 1 所述的方法，其中匹配处理过的记录与确定为反映实体的先前存储的数据的步骤包括步骤：

检索具有多个标识符中至少一个的先前存储的数据；

15 估计是否有不包括在具有多个标识符中至少一个的先前存储数据中的其它标识符包括在处理过的记录中；及

基于该其它标识符，为了与处理过的记录的匹配，分析具有多个标识符中至少一个的先前存储的数据。

20 17、如权利要求 16 所述的方法，其中匹配处理过的记录的步骤还包括检索具有该其它标识符的先前存储的二级数据并将先前存储的二级数据包括在具有多个标识符中至少一个的先前存储的数据中的步骤。

25 18、如权利要求 1 所述的方法，其中匹配处理过的记录与确定为反映实体的先前存储的数据的步骤包括向匹配的处理过的记录指定与确定为反映实体的先前存储的数据的至少一部分关联的持续键的步骤。

19、如权利要求 2 所述的方法，还包括基于用户定义的规则发布消息的步骤。

20 20、如权利要求 19 所述的方法，其中消息包括记录的源和先前存储的数据的至少一个源中的一个。

21、如权利要求 19 所述的方法，其中基于用户定义的规则发布消息的步骤包括以用户定义的时间间隔发布消息的步骤。

22、如权利要求 1 所述的方法，还包括将处理过的记录存储在数据库中的步骤。

5 23、一种处理数据的方法，包括步骤：

接收具有多个标识符的记录，该记录对应于实体；

利用加密算法处理记录中多个标识符中的至少两个，以形成处理过的记录；

比较处理过的记录与先前存储的数据；

10 基于该多个标识符，通过匹配处理过的记录与确定为反映实体的先前存储的数据确定匹配的数据；

基于该多个标识符分析是否有不包括在确定为反映实体的先前存储的数据中的其它标识符包括在处理过的记录中；及

15 基于该其它标识符匹配匹配过的数据与确定为反映实体的先前存储的数据。

24、如权利要求 23 所述的方法，还包括向记录指定或识别源的步骤。

25、如权利要求 24 所述的方法，还包括在匹配匹配的数据的步骤之后解译处理过的记录的至少一部分的步骤。

20 26、如权利要求 25 所述的方法，其中解译后的处理过的记录的至少一部分是源。

27、如权利要求 23 所述的方法，还包括向记录加盐的步骤。

28、如权利要求 23 所述的方法，其中接收的记录包括多个接收的记录。

25 29、如权利要求 23 所述的方法，还包括在利用加密算法处理记录的至少一部分的步骤之前分析记录的步骤。

30、如权利要求 29 所述的方法，其中分析记录的步骤包括比较多个标识符中至少一个与以下中的一个的步骤：

用户定义的准则，及

二级数据库和列表中的一个中的数据集。

31、如权利要求 30 所述的方法，还包括增强记录的步骤。

32、如权利要求 31 所述的方法，其中增强记录的步骤包括根据用户定义的标准格式化多个标识符中的至少一个。

5 33、如权利要求 31 所述的方法，其中增强记录的步骤包括对多个标识符中的至少一个产生变体并将变体包括在记录中。

34、如权利要求 31 所述的方法，其中增强记录的步骤包括：

查询数据集寻找关于记录的二级标识符；及
利用该二级标识符补充记录。

10 35、如权利要求 23 所述的方法，还包括将记录转换成标准化消息格式的步骤。

36、如权利要求 35 所述的方法，其中将记录转换成标准化消息格式的步骤包括利用对应于多个标识符中每个的类型指示器的步骤。

15 37、如权利要求 36 所述的方法，其中在利用加密算法的步骤之后类型指示器是可辨别的。

38、如权利要求 23 所述的方法，其中匹配匹配数据的步骤包括向处理过的记录指定与确定为反映实体的先前存储的数据的至少一部分关联的持续键的步骤。

20 39、如权利要求 24 所述的方法，还包括基于用户定义的规则发布消息的步骤。

40、如权利要求 39 所述的方法，其中消息包括记录的源和先前存储的数据的至少一个源中的一个。

41、如权利要求 40 所述的方法，其中基于用户定义的规则发布消息的步骤包括以用户定义的时间间隔发布消息的步骤。

25 42、如权利要求 23 所述的方法，还包括将处理过的记录存储在数据库中的步骤。

43、一种处理数据的方法，包括步骤：

接收具有多个标识符的记录，该记录对应于实体；

利用加密算法处理记录中多个标识符中的至少两个，以形成处理

过的记录；

比较处理过的记录与先前存储的数据；及

将处理过的记录与确定为反映与实体关系的先前存储的数据关联。

5 44、如权利要求 43 所述的方法，还包括向记录指定或识别源的步骤。

45、如权利要求 44 所述的方法，还包括在将处理过的记录与确定为反映与实体关系的先前存储的数据关联的步骤之后解译处理过的记录的至少一部分的步骤。

10 46、如权利要求 45 所述的方法，其中解译后的处理过的记录的至少一部分是源。

47、如权利要求 43 所述的方法，还包括向记录加盐的步骤。

48、如权利要求 43 所述的方法，其中接收的记录包括多个接收的记录。

15 49、如权利要求 43 所述的方法，其中确定为反映与实体关系的先前存储的数据包括两个先前接收的记录的至少一部分。

50、如权利要求 43 所述的方法，还包括在利用加密算法处理记录中多个标识符的至少两个的步骤之前分析记录的步骤。

20 51、如权利要求 50 所述的方法，其中分析记录的步骤包括比较多个标识符中至少一个与以下中的一个的步骤：

用户定义的准则，及

二级数据库和列表中的一个中的数据集。

52、如权利要求 51 所述的方法，还包括增强记录的步骤。

25 53、如权利要求 52 所述的方法，其中增强记录的步骤包括根据用户定义的标准格式化多个标识符中的至少一个。

54、如权利要求 52 所述的方法，其中增强记录的步骤包括对多个标识符中的至少一个产生变体并将变体包括在记录中。

55、如权利要求 52 所述的方法，其中增强记录的步骤包括：

查询数据集寻找关于记录的二级标识符；及

利用该二级标识符补充记录。

56、如权利要求 43 所述的方法，还包括将记录转换成标准化消息格式的步骤。

5 57、如权利要求 56 所述的方法，其中将记录转换成标准化消息格式的步骤包括利用对应于多个标识符中每个的类型指示器的步骤。

58、如权利要求 57 所述的方法，其中在利用加密算法的步骤之后类型指示器是可辨别的。

59、如权利要求 44 所述的方法，还包括基于用户定义的规则发布消息的步骤。

10 60、如权利要求 59 所述的方法，其中消息包括记录的源和先前存储的数据的至少一个源中的一个。

61、如权利要求 59 所述的方法，其中基于用户定义的规则发布消息的步骤包括以用户定义的时间间隔发布消息的步骤。

15 62、如权利要求 43 所述的方法，还包括将处理过的记录存储在数据库中的步骤。

63、如权利要求 43 所述的方法，还包括将与实体的关系存储在数据库中的步骤。

64、一种处理数据的方法，包括步骤：

20 接收具有多个标识符的记录，该记录对应于实体，且该多个标识符中的至少一个已经先前利用加密算法进行了处理；

比较该记录与先前存储的数据，先前存储的数据的至少一部分已经先前利用加密算法进行了处理；

匹配记录与确定为反映实体的先前存储的数据；及

发布多条消息，其中该多条消息中的至少一条是噪声。

25 65、如权利要求 64 所述的方法，其中发布多条消息的步骤以用户定义的时间间隔出现。

66、如权利要求 64 所述的方法，其中多条消息中的至少一条包括记录的源。

67、如权利要求 64 所述的方法，还包括将记录存储在数据库中

的步骤。

68、一种处理数据的方法，包括步骤：

接收具有多个标识符的记录，该记录对应于实体，且该多个标识符中的至少一个已经先前利用加密算法进行了处理；

5 比较记录与先前存储的数据，先前存储的数据的至少一部分已经先前利用加密算法进行了处理；

关联记录与被确定为反映与实体的关系的先前存储的数据；及发布多条消息，其中多条消息中的至少一条是噪声。

10 69、如权利要求 68 所述的方法，其中发布多条消息的步骤以用户定义的时间间隔出现。

70、如权利要求 68 所述的方法，其中多条消息中的至少一条包括记录的源。

71、如权利要求 68 所述的方法，还包括将记录存储在数据库中的步骤。

15 72、如权利要求 68 所述的方法，还包括将与实体的关系存储在数据库中的步骤。

73、在处理数据的系统中，一种包含用于由计算机执行以执行包括以下步骤的方法的程序指令的计算机可读介质：

接收具有多个标识符的记录，该记录对应于实体；

20 利用加密算法处理记录中多个标识符中的至少两个，以形成处理过的记录；

比较处理过的记录与先前存储的数据；及

匹配处理过的记录与确定为反映实体的先前存储的数据，确定为反映实体的先前存储的数据包括至少两个先前接收记录的至少一部分。

25 74、如权利要求 73 所述用于执行方法的计算机可读介质，方法还包括向记录指定或识别源的步骤。

75、如权利要求 74 所述用于执行方法的计算机可读介质，方法还包括在匹配处理过的记录与确定为反映实体的先前存储的数据的步

骤之后解译处理过的记录的至少一部分的步骤。

76、如权利要求 75 所述用于执行方法的计算机可读介质，其中解译后的处理过的记录的至少一部分是源。

5 77、如权利要求 73 所述用于执行方法的计算机可读介质，方法还包括向记录加盐的步骤。

78、如权利要求 73 所述用于执行方法的计算机可读介质，方法还包括接收多个接收的记录。

10 79、如权利要求 73 所述用于执行方法的计算机可读介质，方法还包括在利用加密算法处理记录的至少一部分的步骤之前分析记录的步骤。

80、如权利要求 79 所述用于执行方法的计算机可读介质，其中分析记录的步骤包括比较多个标识符中的至少一个与以下中的一个：

用户定义的准则；及

二级数据库和列表中的一个中的数据集。

15 81、如权利要求 80 所述用于执行方法的计算机可读介质，方法还包括增强记录的步骤。

82、如权利要求 81 所述用于执行方法的计算机可读介质，其中增强记录的步骤包括根据用户定义的标准格式化多个标识符中的至少一个。

20 83、如权利要求 81 所述用于执行方法的计算机可读介质，其中增强记录的步骤包括对多个标识符中的至少一个产生变体并将该变体包括在记录中。

84、如权利要求 81 所述用于执行方法的计算机可读介质，其中增强记录的步骤包括：

25 查询数据集寻找关于记录的二级标识符；及

用二级标识符补充记录。

85、如权利要求 73 所述用于执行方法的计算机可读介质，方法还包括将记录转换成标准化消息格式的步骤。

86、如权利要求 85 所述用于执行方法的计算机可读介质，其中

将记录转换成标准化消息格式的步骤包括利用对应于多个标识符中每个的类型指示器的步骤。

87、如权利要求 86 所述用于执行方法的计算机可读介质，其中在利用加密算法的步骤之后类型指示器是可辨别的。

5 88、如权利要求 73 所述用于执行方法的计算机可读介质，其中匹配处理过的记录与确定为反映实体的先前存储的数据的步骤包括步骤：

检索具有多个标识符中至少一个的先前存储的数据；

10 估计是否有不包括在具有多个标识符中至少一个的先前存储的数据中的其它标识符包括在处理过的记录中；及

15 基于该其它标识符，为了与处理过的记录匹配，分析具有多个标识符中至少一个的先前存储的数据。

89、如权利要求 88 所述用于执行方法的计算机可读介质，其中匹配处理过的记录的步骤还包括检索具有该其它标识符的先前存储的二级数据并将先前存储的二级数据包括在具有多个标识符中至少一个的先前存储的数据中的步骤。

90、如权利要求 73 所述用于执行方法的计算机可读介质，其中匹配处理过的记录与确定为反映实体的先前存储的数据的步骤包括向匹配的处理过的记录指定与确定为反映实体的先前存储的数据的至少一部分关联的持续键的步骤。

20 91、如权利要求 74 所述用于执行方法的计算机可读介质，方法还包括基于用户定义的规则发布消息的步骤。

92、如权利要求 91 所述用于执行方法的计算机可读介质，其中消息包括记录的源和先前存储的数据的至少一个源中的一个。

25 93、如权利要求 91 所述用于执行方法的计算机可读介质，其中基于用户定义的规则发布消息的步骤包括以用户定义的时间间隔发布消息的步骤。

94、如权利要求 73 所述用于执行方法的计算机可读介质，方法还包括将处理过的记录存储在数据库中的步骤。

95、在处理数据的系统中，一种包含用于由计算机执行以执行包括以下步骤的方法的程序指令的计算机可读介质：

5 接收具有多个标识符的记录，该记录对应于实体；
利用加密算法处理记录中多个标识符中的至少两个，以形成处理
过的记录；

10 比较处理过的记录与先前存储的数据；
基于该多个标识符，通过匹配处理过的记录与确定为反映实体的
先前存储的数据确定匹配的数据；
基于该多个标识符分析是否有不包括在确定为反映实体的先前
存储的数据中的其它标识符包括在处理过的记录中；及
15 基于该其它标识符匹配匹配的数据与确定为反映实体的先前存
储的数据。

96、如权利要求 95 所述用于执行方法的计算机可读介质，方法
还包括向记录指定或识别源的步骤。

15 97、如权利要求 96 所述用于执行方法的计算机可读介质，方法
还包括在匹配匹配的数据的步骤之后解译处理过的记录的至少一部分
的步骤。

98、如权利要求 97 所述用于执行方法的计算机可读介质，其中
解译后的处理过的记录的至少一部分是源。

20 99、如权利要求 95 所述用于执行方法的计算机可读介质，还包
括向记录加盐的步骤。

100、如权利要求 95 所述用于执行方法的计算机可读介质，其中
接收的记录包括多个接收的记录。

25 101、如权利要求 95 所述用于执行方法的计算机可读介质，还包
括在利用加密算法处理记录的至少一部分的步骤之前分析记录的步
骤。

102、如权利要求 101 所述用于执行方法的计算机可读介质，其
中分析记录的步骤包括比较多个标识符中至少一个与以下中的一个的
步骤：

用户定义的准则，及
二级数据库和列表中的一个中的数据集。

103、如权利要求 102 所述用于执行方法的计算机可读介质，方法还包括增强记录的步骤。

5 104、如权利要求 103 所述用于执行方法的计算机可读介质，其中增强记录的步骤包括根据用户定义的标准格式化多个标识符中的至少一个。

10 105、如权利要求 103 所述用于执行方法的计算机可读介质，其中增强记录的步骤包括对多个标识符中的至少一个产生变体并将变体包括在记录中。

106、如权利要求 103 所述用于执行方法的计算机可读介质，其中增强记录的步骤包括：

查询数据集寻找关于记录的二级标识符；及
利用该二级标识符补充记录。

15 107、如权利要求 95 所述用于执行方法的计算机可读介质，还包括将记录转换成标准化消息格式的步骤。

108、如权利要求 107 所述用于执行方法的计算机可读介质，其中将记录转换成标准化消息格式的步骤包括利用对应于多个标识符中每个的类型指示器的步骤。

20 109、如权利要求 108 所述用于执行方法的计算机可读介质，其中在利用加密算法的步骤之后类型指示器是可辨别的。

110、如权利要求 95 所述用于执行方法的计算机可读介质，其中匹配匹配数据的步骤包括向处理过的记录指定与确定为反映实体的先前存储的数据的至少一部分关联的持续键的步骤。

25 111、如权利要求 96 所述用于执行方法的计算机可读介质，还包括基于用户定义的规则发布消息的步骤。

112、如权利要求 111 所述用于执行方法的计算机可读介质，其中消息包括记录的源和先前存储的数据的至少一个源中的一个。

113、如权利要求 112 所述用于执行方法的计算机可读介质，其

中基于用户定义的规则发布消息的步骤包括以用户定义的时间间隔发布消息的步骤。

114、如权利要求 95 所述用于执行方法的计算机可读介质，还包括将处理过的记录存储在数据库中的步骤。

5 115、在处理数据的系统中，一种包含用于由计算机执行以执行包括以下步骤的方法的程序指令的计算机可读介质：

接收具有多个标识符的记录，该记录对应于实体；

利用加密算法处理记录中多个标识符中的至少两个，以形成处理过的记录；

10 比较处理过的记录与先前存储的数据；及

关联处理过的记录与确定为反映与实体关系的先前存储的数据。

116、如权利要求 115 所述用于执行方法的计算机可读介质，方法还包括向记录指定或识别源的步骤。

15 117、如权利要求 116 所述用于执行方法的计算机可读介质，方法还包括在将处理过的记录与确定为反映与实体关系的先前存储的数据关联的步骤之后解译处理过的记录的至少一部分的步骤。

118、如权利要求 117 所述用于执行方法的计算机可读介质，其中解译后的处理过的记录的至少一部分是源。

20 119、如权利要求 115 所述用于执行方法的计算机可读介质，方法还包括向记录加盐的步骤。

120、如权利要求 115 所述用于执行方法的计算机可读介质，其中接收的记录包括多个接收的记录。

25 121、如权利要求 115 所述用于执行方法的计算机可读介质，其中确定为反映与实体关系的先前存储的数据包括两个先前接收的记录的至少一部分。

122、如权利要求 115 所述用于执行方法的计算机可读介质，方法还包括在利用加密算法处理记录中多个标识符的至少两个的步骤之前分析记录的步骤。

123、如权利要求 122 所述用于执行方法的计算机可读介质，其

中分析记录的步骤包括比较多个标识符中至少一个与以下中的一个的步骤：

 用户定义的准则，及
 二级数据库和列表中的一个中的数据集。

5 124、如权利要求 123 所述用于执行方法的计算机可读介质，方法还包括增强记录的步骤。

125、如权利要求 124 所述用于执行方法的计算机可读介质，其中增强记录的步骤包括根据用户定义的标准格式化多个标识符中的至少一个。

10 126、如权利要求 124 所述用于执行方法的计算机可读介质，其中增强记录的步骤包括对多个标识符中的至少一个产生变体并将变体包括在记录中。

127、如权利要求 124 所述用于执行方法的计算机可读介质，其中增强记录的步骤包括：

15 查询数据集寻找关于记录的二级标识符；及
 利用该二级标识符补充记录。

128、如权利要求 115 所述用于执行方法的计算机可读介质，方法还包括将记录转换成标准化消息格式的步骤。

20 129、如权利要求 128 所述用于执行方法的计算机可读介质，其中将记录转换成标准化消息格式的步骤包括利用对应于多个标识符中每个的类型指示器的步骤。

130、如权利要求 129 所述用于执行方法的计算机可读介质，其中在利用加密算法的步骤之后类型指示器是可辨别的。

25 131、如权利要求 116 所述用于执行方法的计算机可读介质，方法还包括基于用户定义的规则发布消息的步骤。

132、如权利要求 131 所述用于执行方法的计算机可读介质，其中消息包括记录的源和先前存储的数据的至少一个源中的一个。

133、如权利要求 131 所述用于执行方法的计算机可读介质，其中基于用户定义的规则发布消息的步骤包括以用户定义的时间间隔发

布消息的步骤。

134、如权利要求 115 所述用于执行方法的计算机可读介质，方法还包括将处理过的记录存储在数据库中的步骤。

135、如权利要求 115 所述用于执行方法的计算机可读介质，方法还包括将与实体的关系存储在数据库中的步骤。

136、在处理数据的系统中，一种包含用于由计算机执行以执行包括以下步骤的方法的程序指令的计算机可读介质：

接收具有多个标识符的记录，该记录对应于实体，且该多个标识符中的至少一个已经先前利用加密算法进行了处理；

10 比较该记录与先前存储的数据，先前存储的数据的至少一部分已经先前利用加密算法进行了处理；

匹配记录与确定为反映实体的先前存储的数据；及
发布多条消息，其中多条消息中的至少一条是噪声。

15 137、如权利要求 136 所述用于执行方法的计算机可读介质，其中发布多条消息的步骤以用户定义的时间间隔出现。

138、如权利要求 136 所述用于执行方法的计算机可读介质，其中多条消息中的至少一条包括记录的源。

139、如权利要求 136 所述用于执行方法的计算机可读介质，方法还包括将记录存储在数据库中的步骤。

20 140、在处理数据的系统中，一种包含用于由计算机执行以执行包括以下步骤的方法的程序指令的计算机可读介质：

接收具有多个标识符的记录，该记录对应于实体，且该多个标识符中的至少一个已经先前利用加密算法进行了处理；

25 比较记录与先前存储的数据，先前存储的数据的至少一部分已经先前利用加密算法进行了处理；

关联记录与确定为反映与实体的关系的先前存储的数据；及
发布多条消息，其中多条消息中的至少一条是噪声。

141、如权利要求 140 所述用于执行方法的计算机可读介质，其中发布多条消息的步骤以用户定义的时间间隔出现。

142、如权利要求 140 所述用于执行方法的计算机可读介质，其中多条消息中的至少一条包括记录的源。

143、如权利要求 140 所述用于执行方法的计算机可读介质，还包括将记录存储在数据库中的步骤。

5 144、如权利要求 140 所述用于执行方法的计算机可读介质，还包括将与实体的关系存储在数据库中的步骤。

机密数据共享和匿名实体解析度

5 相关申请的交叉引用

本申请要求于 2002 年 11 月 6 日在美国专利局提交的编号为 60/424,240 的临时申请的权益。

10 政府资助的研究或开发
不适用。

技术领域

本发明涉及处理和检索数据库中的数据，更具体而言，涉及以机密和匿名方式对数据的提交、比较和匹配/关联。

15 背景技术

鉴于 2001 年的 911 事件，各种团体（例如，公司、政府代理机构或自然人）都面临一个共同的两难问题：团体如何共享帮助相同或独立团体检测可能的恐怖分子或其它问题团体的存在的特定信息（例如，恐怖分子观察名单，黑名单或者实际或可能的问题实体名单），同时保持这种信息的安全性和机密性并隔离任何与事件不相关的信息？

25 提供或公开特定信息的犹豫及管理特定信息的使用与公开的法律是在信息以侵犯团体隐私或对其造成伤害的方式使用的考虑下声明的。这种伤害包括身份盗用，非法的直接市场活动，非法或入侵式政府活动，受保护阶级（例如，种族，宗教，性别，民族）迫害和歧视，反竞争实践，诽谤和/或信誉或经济损害。

响应这种两难问题，或任何需要共享机密数据的情况，有这样一种系统是有利的，其中各种团体可以如下方式向内部或外部处理或仓

库提供数据:(a)不公开任何识别实体的数据(例如, 姓名或社会安全号), 充分识别数据中的每条记录(例如, 源和记录号); (b)准备数据, 从而:(i)不管源是什么, 从相同的数据都得到相同的唯一值, 及(ii)这种数据可以标准但机密的格式发送, 以便保护数据的机密性和安全性,
5 (c)将数据与先前提供的数据进行比较, 同时数据仍然是机密格式, (d)通过匹配比较后的数据构造可识别实体(例如, 通过利用可能的别名, 地址, 编号和/或其它识别信息的机密表示, 使用持续密钥及分析和增强记录), (e) 通过关联比较后的数据构造相关实体, 和/或(f)产生用于合适团体的消息(例如, 利用相关记录识别元素—例如, 源和记录号),
10 这种信息有时候是以机密方式发送的, 如:(i)以时间间隔为基础和/或其中至少一条消息是噪声(例如, 不与匹配或关联对应但发布以便最小化与业务量模式分析对应的特定脆弱性的消息)和(ii)在这种消息通过可逆加密算法(例如, 编码、加密或其它用于产生机密等级但可以例如通过利用解码或解密反转的算法)处理以后。

15 当前的系统利用各种方式以相对机密的方式在团体内部或之间传输数据。例如, 有些当前系统利用可逆加密算法, 通过了解接收者将使用可比较解码或解密方法(即, 将编码/加密的数据反转、返回或修改回原始数据的格式表示)来解译并理解数据, 在发送数据之前修改数据以便产生一些机密等级并降低在发送过程中丢失数据的危险。
20 但是, 一旦数据被解译了, 这种数据就要以造成恰恰是编码/加密处理要保护的伤害的方式被分析和使用。

其它当前系统利用常常作为文档签名的不可逆加密算法(例如, 单向函数, 如 MD-5 或其它用于产生机密等级但不可逆的算法), 从而当文档在团体之间共享时使非法的文档改变可以被检测到。事实上, 有些现有的不可逆加密算法使数据:(a)不管源是什么, 对相同的数据都导致完全相同的唯一值, 及(b)不可解译和不可逆, 从而保护数据的机密性和安全性。即使数据是相同的, 但是与没有小改变的数据相比, 在使用不可逆加密算法以后, 数据中任何小的改变(如额外的空间)都导致不同的值。有些当前系统利用不可逆加密算法处理一部分数据,

然后基于完全相同的处理过的数据一一对应地匹配与合并记录。例如，目前医院中的系统可以通过单向函数处理电子病人记录中的社会安全号，然后基于处理过的社会安全号一一对应地匹配与合并数据库中的记录。

但是，还没有现有系统能最低限度：(a) 在至少一部分这种接收到的数据通过加密算法（例如，可逆加密算法，如编码或解密，或者不可逆加密算法，如单向函数）处理以后，以一对多或多对多为基础（即，接收到的数据包括一条或多条与先前存储在数据库中的数据匹配的记录，先前存储在数据库中的匹配数据包括不止一条先前接收到的记录）匹配接收到的数据与先前存储在数据库中的数据，这限制了当前系统中建立可识别实体的能力，同时数据还是机密格式，(b) 移动到初始匹配过程以外，来分析是否有任何附加信息在初始匹配中获得，然后基于该附加信息匹配先前存储在数据库中的其它数据，这进一步限制了当前系统构造可识别实体的能力，(c) 利用这一段(a)和(b)中识别的那些功能的全部或部分来不仅匹配相同的实体，还以某种方式（例如，航线预订名单上的乘客是航线观察名单上一个自然人的室友）关联确定为相关的各种实体，和/或(d)发布多条消息，其中多条消息中至少有一条仅仅是噪声。

就象这样，此外，没有现有系统可以使用加密算法共享和比较机密数据（包括通过让个人可识别信息保持加密格式，但不作为限制），构造可识别或相关实体并以保持原始数据安全性和机密性的方式通知合适的实体。

提供本发明是为了解决这些和其它问题。

25 发明内容

本发明的一个目的是提供用于处理数据库中数据的方法、程序和系统。该方法、程序和系统优选地包括步骤：(a)接收一条或多条记录，每条记录都具有多个标识符（例如，多个已知类型的数据值，如对于已知“名”类型的“John”和对于已知“姓”类型的“Smith”的两

个(2)数据值，这两个值有时候是来源于被解析成对应于独立已知类型的独立值的一个数据值，如当对应于已知“姓名”类型的原始数据值 John Smith 被解析成对应于已知“名”类型的 John 和对应于已知“姓”类型的 Smith 的两个(2)数据值时)，每条记录对应于至少一个实体，

5 (b) 利用加密算法处理记录中多个标识符中的至少两个，(c) 有时候，在将处理过的记录发送到独立的系统或数据库以后，比较处理过的记录与先前存储的数据，及 (d) 匹配处理过的记录与确定为反映实体的先前存储的数据，确定为反映实体的先前存储的数据包括至少两条先前接收的记录的一部分。

10 但是，还期望该方法、程序和系统还包括步骤：(a) 为记录指定或识别源（例如，提供该记录的组织，该组织内的特定系统，及代表该特定系统内记录的唯一标识符），(b) 在使用加密算法之前向记录加盐（即，用于拼凑、修改、扭曲或覆盖处理过的数据的附加数据），及 (c) 有时候，在匹配处理过的数据与确定为反映实体的先前存储的数据的步骤之后，解译至少一部分处理过的记录（如源）。

15 还期望该方法、程序和系统还包括在使用加密算法的步骤之前分析记录的步骤，该步骤可以包括：(a) 对照用户定义的准则（如用户定义的标准）或二级数据库中的一个或多个数据集（如对于二级标识符查询二级数据库）或列表比较标识符，及 (b) 增强记录，如通过：
20 (i) 对至少一个标识符产生至少一个变体并在记录中包括该变体，及
(ii) 利用二级标识符补充记录。

25 还期望该方法、程序和系统还包括将记录转换成标准消息格式的步骤。例如，该方法、程序和系统可以通过对每个标识符使用类型指示器（例如，标号、变量、标志或其它对应于类型的指示器，如对应于类型的 XML 标志，如姓名或电话号码）将记录转换成标准消息格式。作为进一步的例子，其中记录包含对应于一个(1)姓类型和二个(2)电话号码类型的三个(3)标识符，利用<2>作为姓类型的类型标识符和<3>作为电话号码类型的类型标识符，一条标准化后的记录可以导致以下：
<2>Smith</2><3>111-222-3333</3><3>222-111-

3131</3>。还期望在使用加密算法的步骤之后类型指示器可以是可辨别的。例如，在加密算法中处理以后，在这段中上面阐述的标准化后的记录可以导致以下：

<2>23ff0ad398g13ef82kcks83cke821apw</2><3>bcke39sck30cvk
5 1002ckwlAeMn301L3b</3><3>23kaPek309cwf319oc3f921ldks8773q</
3>。

10 还期望匹配处理过的记录的步骤包括步骤：(a) 检索具有完全相同标识符的先前存储的数据，(b) 估计是否有不存在于先前存储的数据中的其它标识符包括在处理过的记录中，(c) 基于该其它标识符为了与处理过的记录进行匹配而分析先前存储的数据，(d) 重复这些步骤，直到基于该其它标识符先前存储的数据为了与处理过的记录进行匹配都分析了，及(e) 向匹配的处理过的记录指定与至少一部分先前存储的数据关联的持续键(即，唯一的数字或字母数字标识符，最低限度可以用于区分对应于特定实体的一条或多条记录和对应于不同实体的其它记录)。就先前存储的数据的持续键作为任何匹配的结果而改变而言，参照改变的持续键，系统可以保存任何先前的持续键。
15

20 还期望该方法、程序和系统包括步骤：(a) 基于用户定义的规则发布一条或多条消息，规则如：(i) 其中消息包括记录的源和/或先前存储的数据的源，该源可以用于识别其它源中的相关信息，(ii) 其中至少一条消息是噪声，和/或(iii) 以用户定义的时间间隔，和/或(b) 将处理过的记录存储在数据库中。

25 还期望该方法、处理和系统包括步骤：(a) 接收具有多个标识符的记录，该记录对应于一个实体，(b) 利用加密算法处理记录中多个标识符中的至少两个(有时候在使用加密算法之前分析记录)，(c) 比较处理过的记录与先前存储的数据，(d) 基于多个标识符中的至少一个匹配处理过的记录与确定为反映实体的先前存储的数据，(e) 分析是否有不包括在匹配的先前存储的数据中的其它标识符包括在处理过的记录中，及(f) 基于该其它标识符匹配匹配数据与确定为反映实体的先前存储的二级数据(有时候将处理过的记录存储在数据

库中)。

还期望该方法、系统和程序还包括步骤：(a)接收具有多个标识符的记录，该记录对应于一个实体，(b)利用加密算法处理记录的至少一部分(有时候在使用加密算法之前分析并增强记录)，(c)5 比较处理过的记录与先前存储的数据，(d)关联处理过的记录与确定为反映与实体关系的先前存储的数据，及(e)将与实体的关系存储在数据库中。

还期望该方法、系统和程序还包括向处理过的记录指定持续键的步骤。

10 还期望该方法、系统和程序还包括步骤：(a)接收具有多个标识符的记录，该记录对应于一个实体，而且多个标识符中的至少一个已经利用加密算法进行了处理；(b)比较记录与先前存储的数据，先前存储的数据的至少一部分已经利用加密算法进行了处理；(c)匹配记录与确定为反映与实体关系的先前存储的数据和/或关联记录与确定为反映与实体关系的先前存储的数据；及(d)发布多条消息，其中多条消息中的至少一条是噪声。

本发明的这些和其它方面与属性将参考以下附图和相应说明进行讨论。

20 附图说明

图1是根据本发明的系统的功能性方框图。

图2是接收和处理记录步骤的流程图。

图3是图2中分析记录块的流程图。

图4是图2中比较连接数据块的流程图。

25 图5-7是进一步说明图2中比较连接数据块的流程图。

1 具体实施方式

尽管本发明可以有许多不同形式的实施方式，但在理解本公开内容应当看作本发明原理的范例，而不是要将本发明限定到所说明的特

定实施方式的前提下将在附图中示出，而且在此详细描述其特定实施方式。

用于在存入数据库之前处理数据的数据处理系统 10 在图 1 至 7 中说明。系统 10 包括至少一台具有处理器 14 和存储器 16 的传统计算机 12。存储器 16 既用于操作系统 10 的可执行软件的存储，又用于数据库和随机存取存储器中数据的存储。依赖于相关的机密性和安全需求，全部或部分软件可以在各种不同的应用和设备中实现。例如，软件可以利用以下任何一个最低限度在任何计算机可读介质上实现、存储或提供：(a) 安装在源系统上的软件应用，(b) 在有任何篡改时自销毁单元的盒子单元，和/或 (c) CD、DVD 或软磁盘。计算机 12 可以从一个或多个源 18₁ - 18_n 接收输入。

数据包括一条或多条具有多个标识符的记录。每条记录对应于一个或多个实体。该一个或多个实体可以是自然人、组织、动产、不动产、蛋白质、化学或器官混合物、生物或原子结构，或其它可以通过识别数据来表示的项。例如，包含姓名、雇主姓名、家庭地址、工作地址、工作电话号码、家庭电话号码、汽车牌照号码、汽车类型和社会安全号标识符的记录最低限度可以对应以下实体：(a) 自然人，(b) 组织（例如，雇主或航线），和/或 (c) 财产（例如，汽车）。

系统 10 从一个或多个源 18₁ - 18_n 接收数据，并如图 2 所示处理每个接收到的记录。软件存储在存储器 16 中，并由处理器 14 处理或实现。

如图 2 所说明的，系统 10 在步骤 20 接收接收到的记录并以包括以下步骤的方式处理接收到的记录：(a) 如果接收到的记录不是标准格式（例如，XML），则在步骤 22 将接收到的记录转换成标准格式，和 (b) 在步骤 24 将记录识别、指定或归于一个源（例如，一个或多个识别源 18₁ - 18_n 中记录的标识符 - 如记录的一个或多个主键，如组织 ID、系统 ID 和记录 ID）（归于所接收记录的源的标识符是“属性数据”）。可选地，有时候使用编码的交叉引用表，由此属性数据由属性键表示，而该属性键用于在必要时定位属性值。

通过：(a) 比较所接收记录 20 的标识符中的至少一部分与用户定义的准则和/或规则来执行几种功能，如：(i) 步骤 28 中的姓名标准化（例如，同根姓名列表比较），(ii) 步骤 30 中的地址净化（例如，同邮递标准比较），(iii) 步骤 32 中的地球编码（例如，确定地理位置，如维度和经度坐标），(iv) 步骤 34 中的域测试或转化（例如，比较性别域以验证 M/F 或将 Male 转化成 M），(v) 步骤 36 中的用户定义的格式化（例如，将所有社会安全号格式化成 999-99-9999 格式）和/或 (vi) 步骤 38 中的变体产生和包括（例如，公共值的代替值或拼写错误），(b) 通过在步骤 42 中使系统 10 访问一个或多个数据库（这可以包括处理先前识别的，从而使系统以级联方式访问附加数据库）来搜索可以在步骤 44 中添加到所接收记录的附加数据（这可以作为用于接收的新记录进行提交并在步骤 20 中进行处理），在步骤 40 中补充所接收的记录，及(c) 在步骤 46 中建立并包括哈希键（例如，所接收记录中特定数据的组合，如根名的前三个字母、姓的前四个字母和社会安全号的后五个数字），系统 10 还在步骤 26 通过分析和增强所接收记录的多个标识符中的一个或多个处理在步骤 20 接收到的记录（记录的分析和增强后的标识符是“增强数据”）。任何新的、修改过的或增强的数据都可以存储在新创建的域中，以保持原始数据的完整性。通过分析和增强每条记录中的标识符，对于相同实体的标识符更有可能匹配（通过原始标识符，或通过新的、修改过的或增强的数据）。

然后，在步骤 48 中通过加密算法处理增强数据的全部或部分，而属性数据的全部或部分有时候是通过加密算法处理的，为了保护和机密性，这可以包括向增强数据或属性数据加盐，如：(a) 利用不可逆加密算法（例如，单向函数）处理可识别实体的增强数据（不可逆的处理数据是“匿名数据”）和 (b) 有时候利用可逆加密算法（例如，加密或编码）处理属性数据(可逆的处理数据是“受保护的属性数据”）。

然后，匿名数据和受保护的属性数据在步骤 50 中连接（有时候还通过可逆加密算法进一步处理）（连接的匿名数据和受保护的属性

数据是“连接数据”），并在步骤 52 中转移、处理并保存在仓库（“仓库”）中。

步骤 52 中仓库的位置并不太关键，因为步骤 52 中仓库中的连接数据是机密格式。此外，在这个例子中，只有属性数据易于从连接数据反转（例如，解密或解码）。如此，即使非法团体能够反转属性数据，这种团体也不能够访问、读取或估计增强数据。但是，整个连接数据可以用于比较和身份识别或相关目的，同时还保持机密性。

步骤 52 中仓库中的系统在步骤 54 中比较连接数据与先前存储的数据（如来自其它源和可能是所存储数据的仓库）并匹配任何反映相同或相关实体的数据。如图 4 所说明的，这类似地在以本发明人为发明人、以“实时数据仓库”为标题、申请号为 10/331,068 并且公开号为 2003/0154194A1、在 2003 年 8 月 14 日公开的发明中示例，仓库中的系统在步骤 56 分析连接数据的第一标识符并在步骤 58 中确定这种标识符是否是候选列表建立者标识符（例如，可以帮助区分实体的标识符）。例如，表示自然人类型实体的社会安全号的标识符有助于区分实体并将用于建立用于匹配或关系建立的可能候选列表。如果标识符是候选列表建立者标识符，则系统将在步骤 60 中确定该标识符是否通常在实体之间是独特的，如通过比较该标识符与公共标识符列表并确定这种标识符是否在该列表上。如果标识符：(a) 不是候选列表建立者标识符或 (b) 是候选列表建立者标识符但通常在实体之间不是独特的，则系统将在步骤 62 中确定连接数据中是否有任何附加的未比较的标识符。

除非基于用户定义的准则，在步骤 66 中系统确定标识符不应当被认为是通常独特的标识符，否则如果标识符是候选列表建立者和通常独特的标识符，则系统将在步骤 64 中检索先前存储的数据中完全相同的标识符的全部出现。例如，如果：(a) 标识符表示社会安全号且处理后的值对应于值 999-99-9999（例如，在社会安全号未知的情况下作为缺省值用于源系统中的值）和 (b) 对应于社会安全号标识符的用户定义准则是如果完全相同的社会安全号个数达到五十（50）（或

某个其它设定的数量)则停止检索出现, 在相同的社会安全号个数达到 51 的时候, 系统将确定该社会安全号不是通常独特的标识符, 并将停止检索出现。

如果在步骤 66 中标识符仍然被认为是通常独特的标识符, 则检索到的出现在步骤 68 中更新或添加到候选列表和关系记录。但是, 如果在步骤 66 中系统确定标识符不被认为是通常独特的标识符, 则在步骤 70 中系统停止基于公共标识符的匹配(如通过将标识符添加到公共标识符列表)并在步骤 72 中不匹配先前基于公共标识符匹配的记录。最后, 系统在步骤 62 确定连接数据中是否有任何附加的未比较的标识符。

一旦系统在步骤 62 确定不再有未比较的标识符, 则系统在步骤 74 检索对应于候选列表的所有用于信用和/或身份识别的标识符(这种标识符是否是候选建立标识符)并在步骤 76 中比较连接数据与候选列表, 从而使系统能够在步骤 78 创建信用指示器(如相似指示器和相关指示器)并利用信用指示器更新候选列表和关系记录。然后, 系统在步骤 80 确定是否有任何基于相似指示器的匹配, 如果有匹配被识别出来, 则在步骤 82 中估计匹配的记录是否包含任何新的或先前未知的标识符, 该标识符可以是添加或更新候选列表/关系记录的候选列表建立者标识符。这种处理在步骤 84 重复, 直到没有更多的匹配可以辨别。然后, 系统将在步骤 86 中为所有匹配的记录指定相同的持续键。如果对于任何记录都没有找到匹配, 则在步骤 88 中为连接数据指定新的持续键。贯穿整个处理, 系统保留连接数据的全部属性, 而且通过合并、净化或删除功能, 匹配处理不丢失任何数据。

比较连接数据与先前存储的数据以便找出相同实体匹配的步骤的其它例子在图 5-7 中说明(尽管没有明确指出, 图 5-7 识别出了关于识别相关实体关联性所使用的几种功能)。从空数据库 90 开始, 连接数据记录一 92(源于公司 AAA、预订系统, 预订系统中的记录 21, 或 AAA-Res-0021, 在标准化、分析和增强, 匿名化和连接以后)在仓库中接收。假定数据库 90 中没有记录, 则系统在所存储的数据中将

找不出完全相同的候选列表和通常独特的标识符的出现（这可以基于变体和/或哈希键来启用类“模糊逻辑”能力），导致步骤 94 中的空候选。假定空候选列表，在步骤 96 中将检索不到对应于候选列表的标识符，将不创建信用指示器，因此将不出现类似匹配。如此，系统将在 5 步骤 98 中向连接数据记录一指定新的持续键并将该记录添加到数据库。

如在图 6 中进一步说明的，当在步骤 100 中连接数据记录二在仓库中接收时，发生相同的处理，如源于公司 BBB、进餐系统，进餐系统中的记录 0486（或 BBB-Din-0486）。系统将在步骤 102 检索所存储数据中完全相同的候选列表和通常独特的标识符的全部出现，（与持续键一起）识别出要添加到候选列表的相同的电话号码。系统在步骤 104 中检索对应于候选列表的所有标识符，得到对应于被检索的第一记录的数据。只基于匹配的电话号码，系统将创建关系信用指示器（如指示两个人是室友），但电话号码匹配将不指示高度信用相似匹配缺少验证相似性的附加信息，如相同的姓名或社会安全号。如此，10 没有相似信用指示器指示匹配，系统将在步骤 106 中向连接数据记录 15 二指定新的持续键并将记录添加到数据库。

如在图 7 中进一步说明的，在步骤 108 中一接收到连接数据记录，如源于公司 CCC、汽车租赁系统，汽车租赁系统中的记录 0356（或 CCC-Car-0356），系统将在步骤 110 中检索所存储数据中完全相同的候选列表和通常独特的标识符的全部出现，（与连接数据记录一的持续键一起）从连接数据记录一得到识别社会安全号的候选列表并（与连接数据记录二的持续键一起）从连接数据记录二得到识别司机驾驶执照号码的候选列表。然后，系统将在步骤 112 中从先前存储的记录 20 检索对应于显示全部标识符的候选列表的所有标识符。然后，系统可以基于连接数据记录三和检索到的数据创建信用指示器，使要创建的相似信用指示器指示匹配。基于相似匹配，系统将在步骤 114 中向所有记录指定相同的持续键并将连接数据记录三添加到数据库。有时候，25 为了系统能够在稍后认识到持续键 PK: 00000002 已经改变成持续键

PK: 00000001, 对应于连接数据记录二的前一持续键, PK: 00000002, 将参照新的持续键, PK: 00000001, 进行保存。

此外, 如图 2 所说明的, 连接数据和任何结果关系和/或匹配将在步骤 116 中存储, 一连串消息将在步骤 118 中基于用户定义的规则发送, 规则如: (i) 以设置的时间间隔为基础, (ii) 有时候包括仅仅是噪声的消息, 以最小化业务量分析攻击, 和/或 (iii) 关于真正的消息, 从匹配的数据中识别受保护的属性数据。例如, 尽管不是对应于属性数据的原始数据 (这是仓库未知的), 但相关团体可以给予可以解译的受保护的属性数据 (例如, AAA-Res-0021, BBB-Din-0486 和/或 CCC-Car-0356), 从而使这种组织可以从其它团体请求或与其共享对应于属性数据的特定记录。作为进一步的例子, 在匹配或关系没有在时间间隔内确定的情况下, 将发送噪声消息。适当的过程将建立, 以保持基本属性数据的机密性和安全性。

如前所述, 可以看到在不背离本发明主旨与范围的前提下可以实现许多变体和修改。应当理解, 不是要打算或推断出要将本发明限制于在此所说明的特定装置。当然, 正如落在权利要求的范围的所有这种修改要由所附权利要求覆盖。

图1

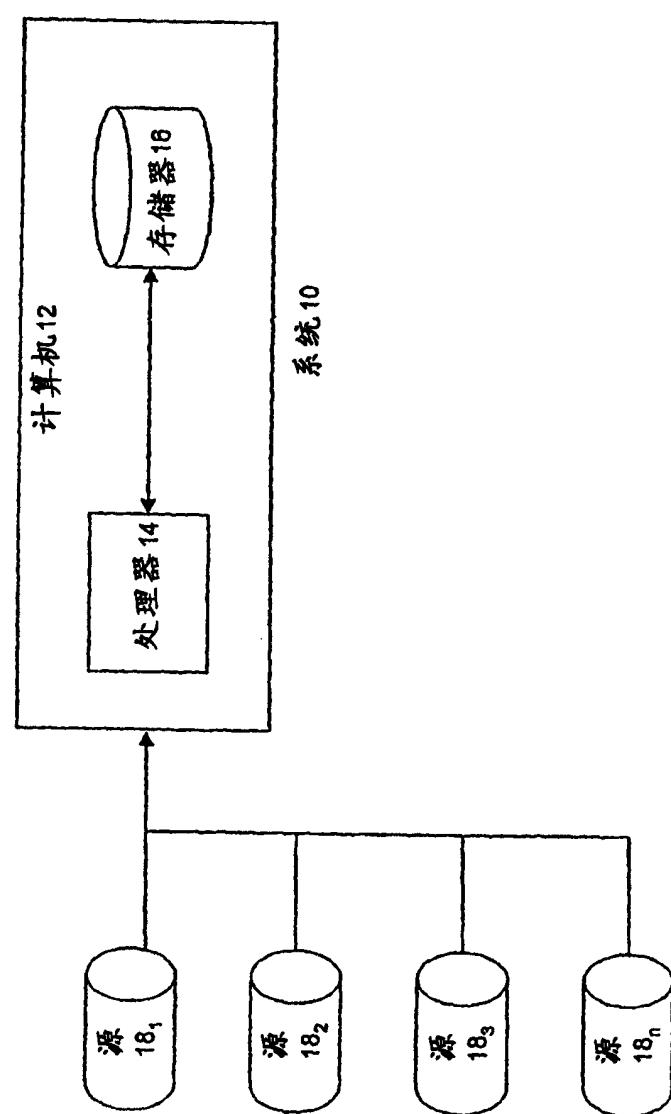


图 2
接收和处理接收的记录

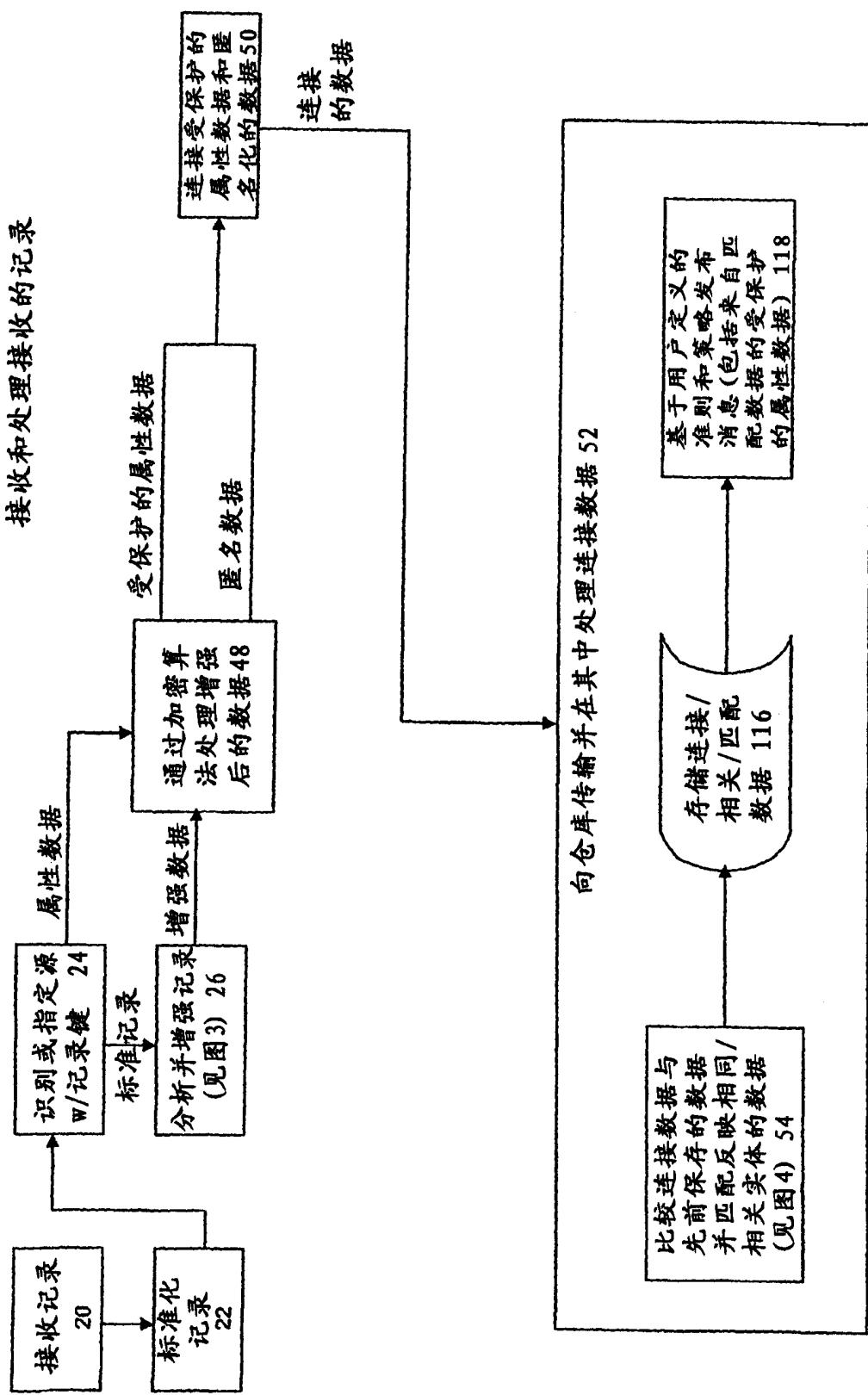
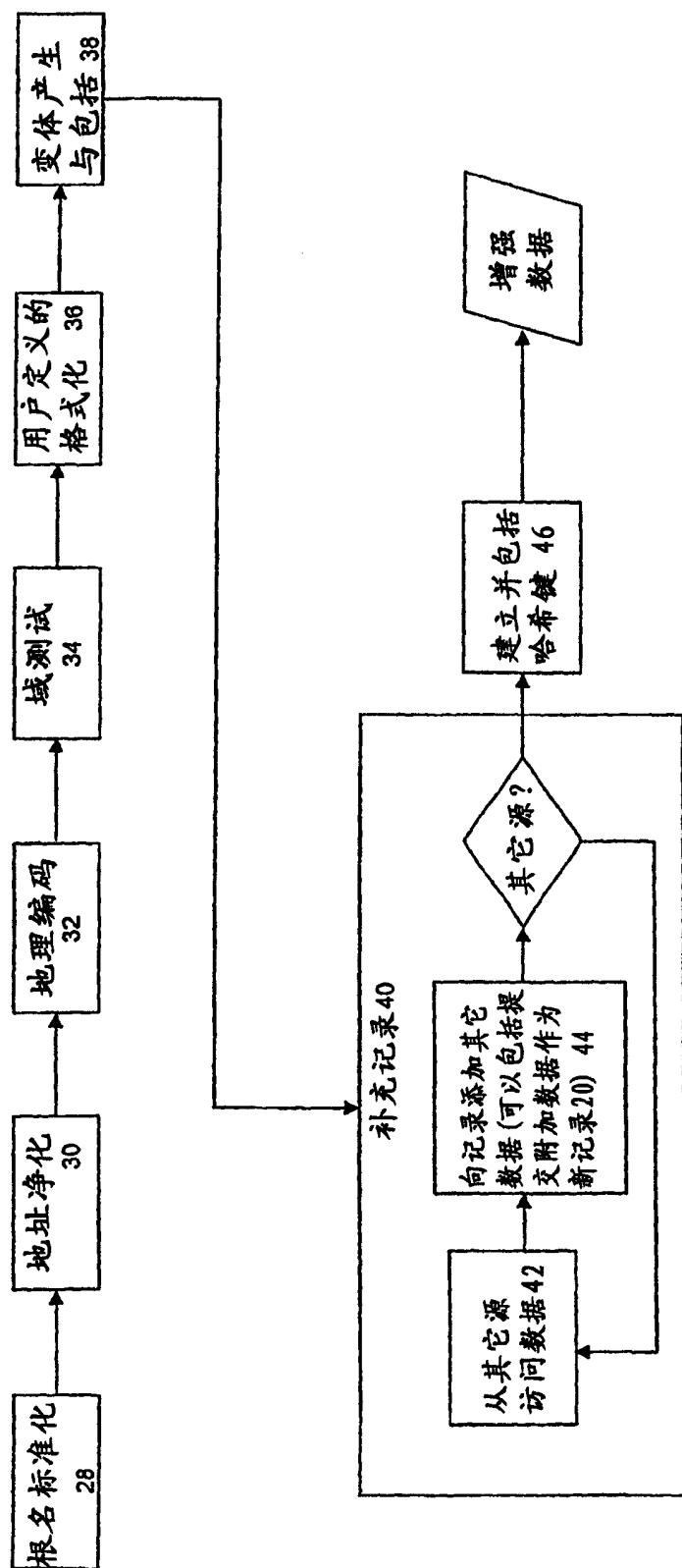


图3
分析和增强记录26



4

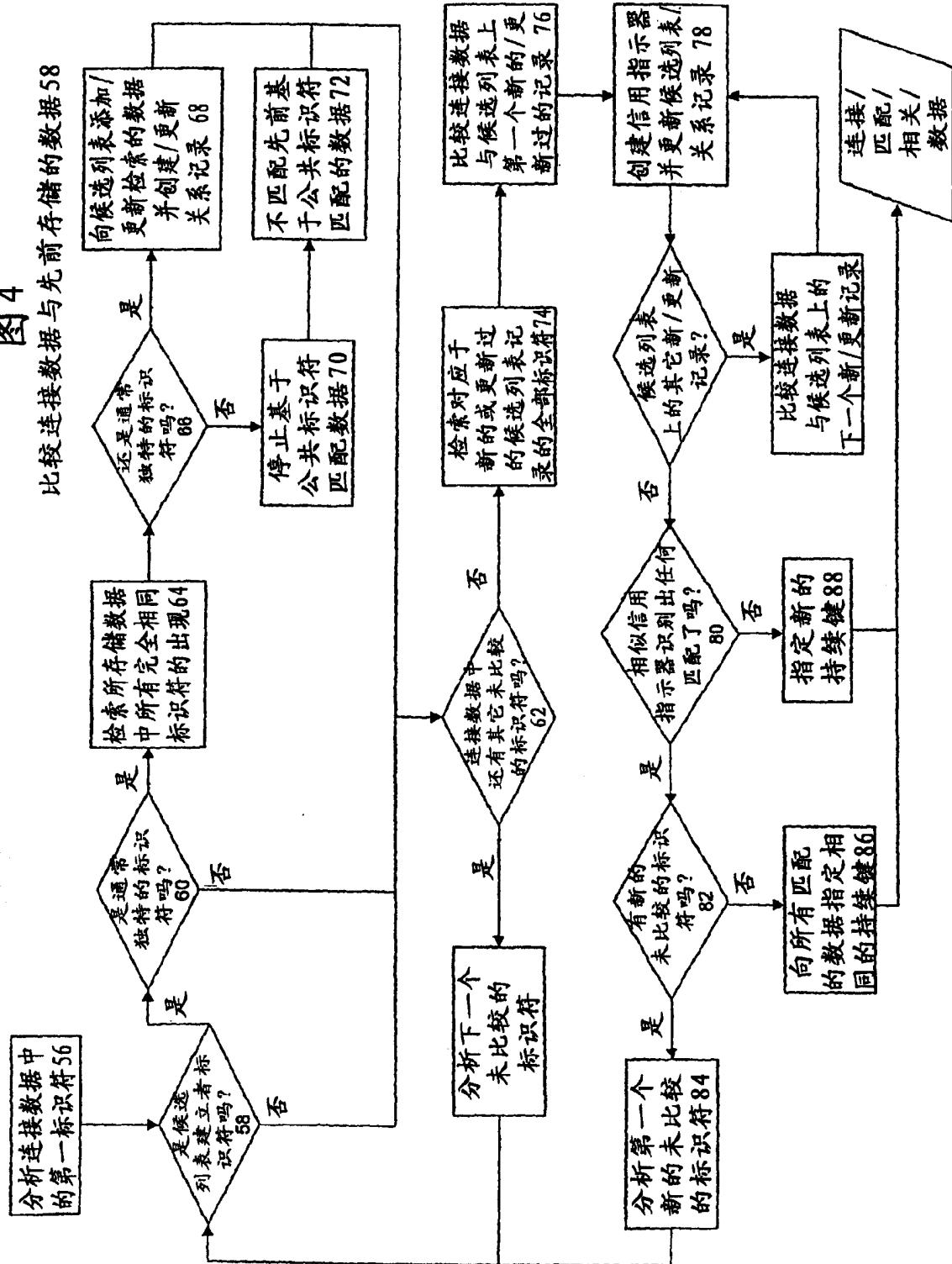
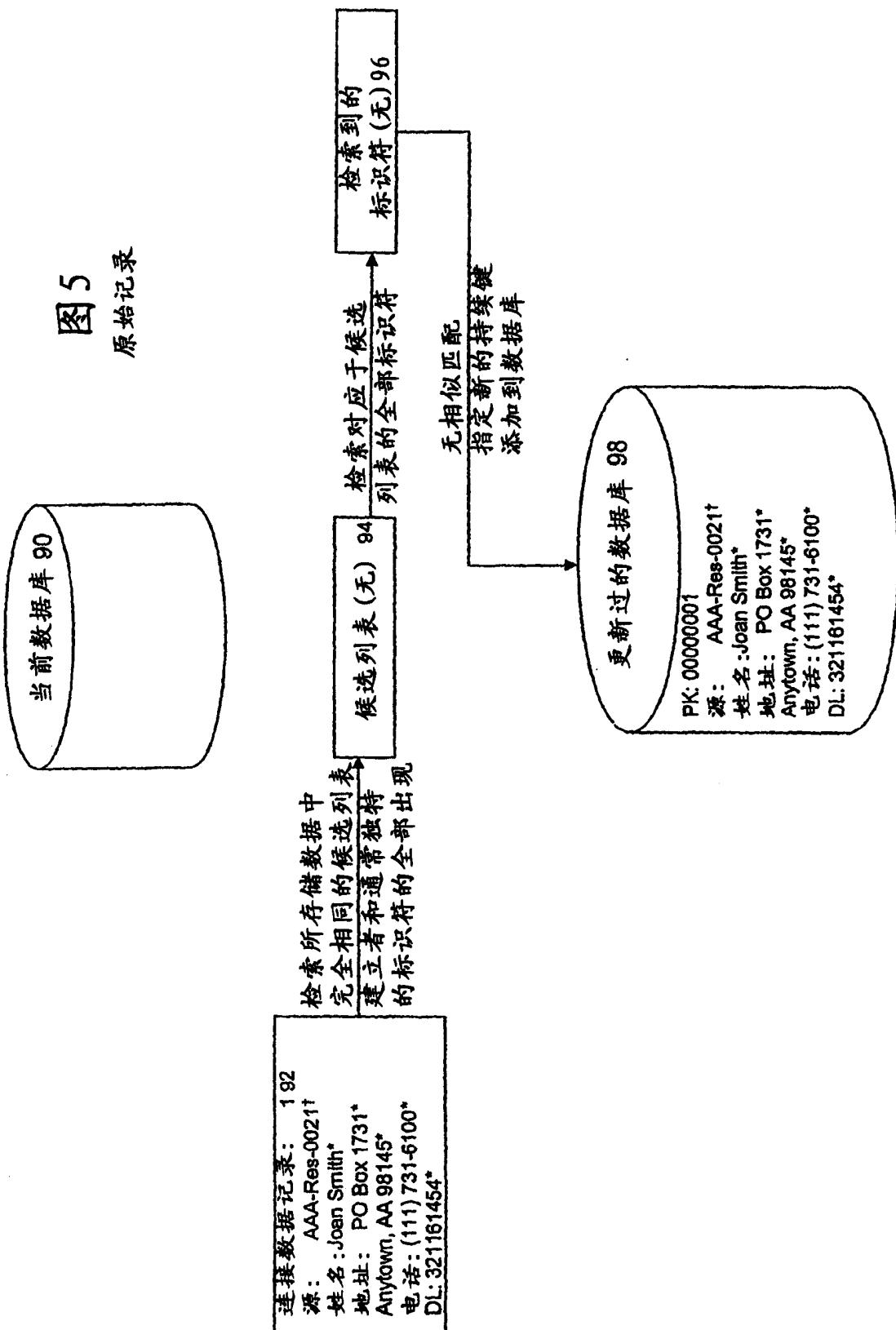
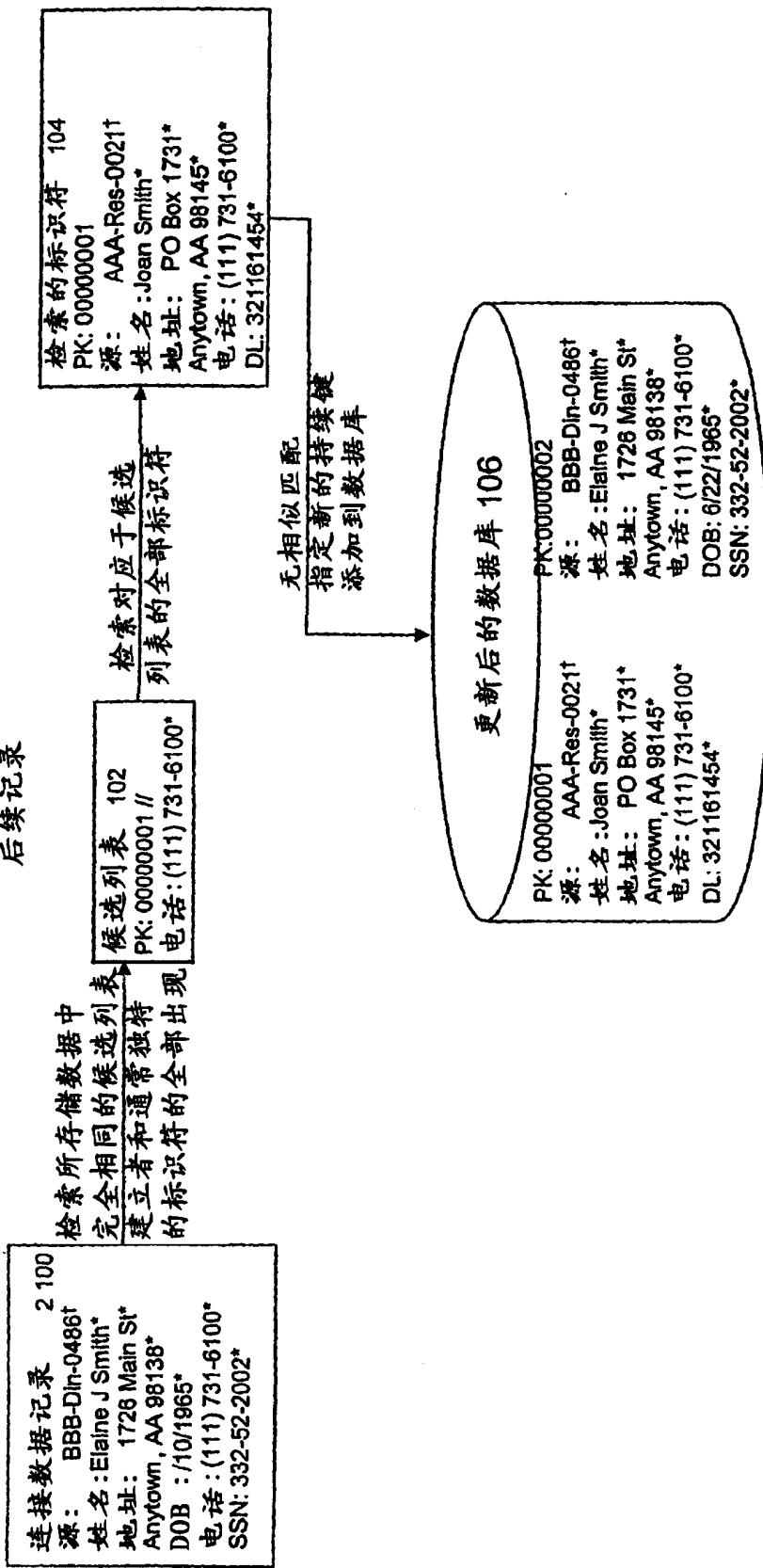


图 5
原始记录



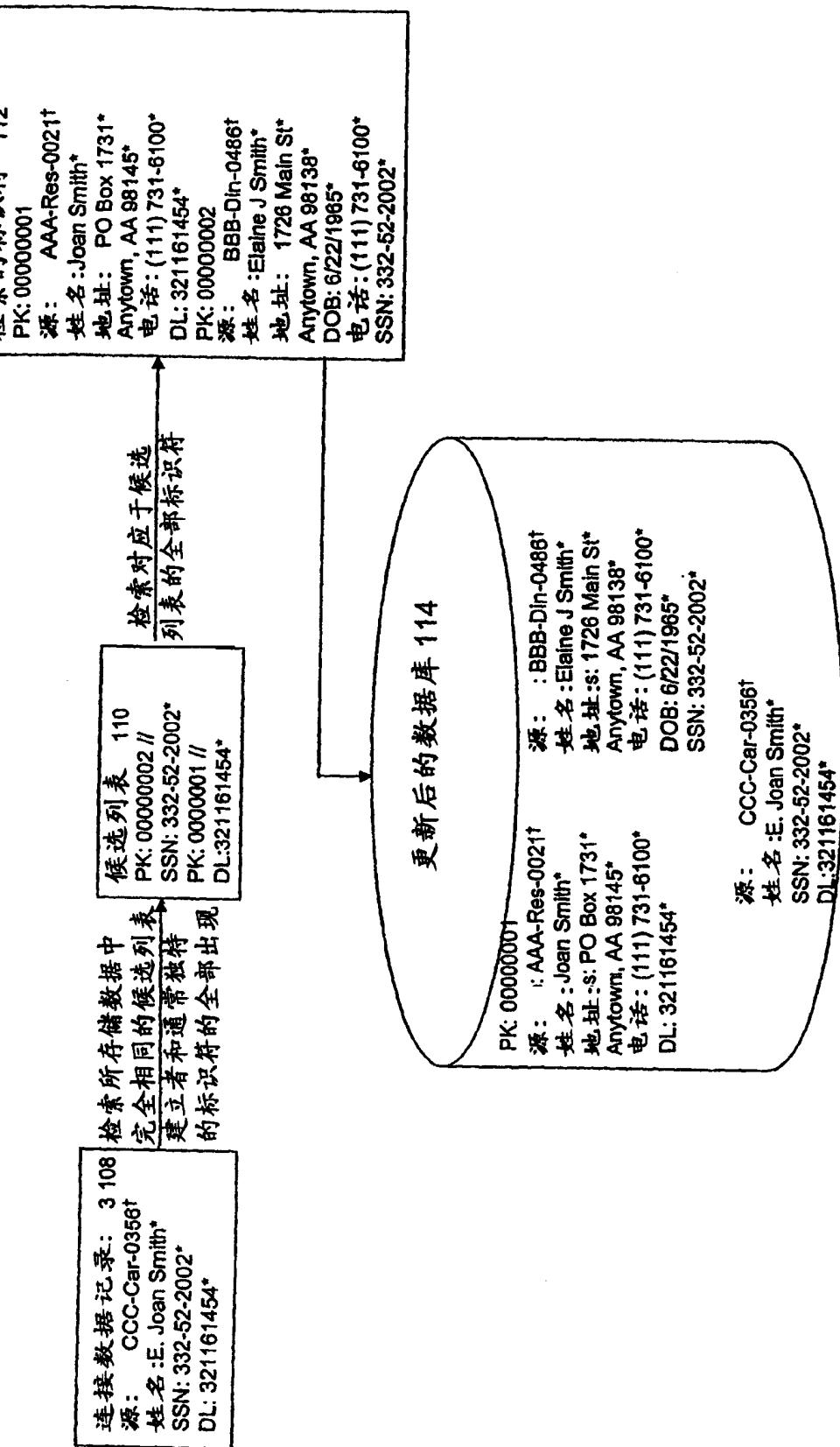
† 通过可逆加密算法(例如, 加密)处理的
• 通过不可逆加密算法(例如, 单向函数)处理的

图 6
后续记录



* 通过不可逆加密算法(例如, 加密)处理的
† 通过可逆加密算法(例如, 单向函数)处理的

图 7
第三记录



* 通过不可逆加密算法(例如, 加密)处理的
† 通过可逆加密算法(例如, 单向函数)处理的