(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 109844785 A
(43)申请公布日 2019.06.04

(21)申请号 201780046543.9

(22)申请日 2017.07.14

(30)优先权数据
62/366,444 2016.07.25 US

(85)PCT国际申请进入国家阶段日
2019.01.25

(86)PCT国际申请的申请数据
PCT/US2017/042034 2017.07.14

(87)PCT国际申请的公布数据
WO2018/022315 EN 2018.02.01

(71)申请人 安客诚有限责任公司
地址 美国阿肯色州

(72)发明人 C·鲍威尔 J·廷德尔
B·沃尔什 S·戴维斯

(74)专利代理机构 上海专利商标事务所有限公司 31100
代理人 陈斌

(51)Int.Cl.
G06Q 10/10(2012.01)
G06Q 30/02(2012.01)
G16H 10/60(2018.01)
G06F 16/215(2019.01)

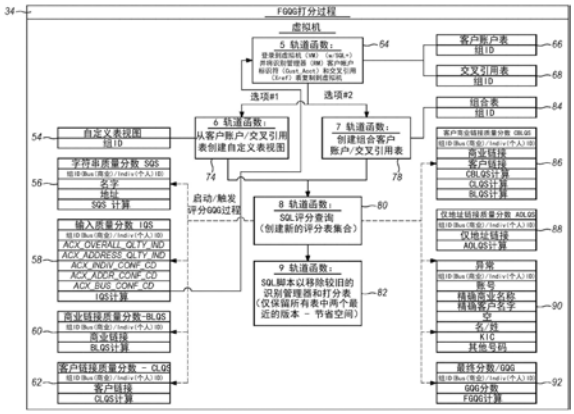
权利要求书4页 说明书17页 附图10页

(54)发明名称

识别质量管理

(57)摘要

识别质量管理系统和方法被用于为包含关于对象的数据结构的数据库确定最终组质量评分(FGQG),其中FGQG是表明在数据库内发生的识别的质量的单个数字分数。使用包含至少三个分量的加权算法来计算FGQG:由字符串距离计算确定的字符串质量分数(SQS);由地址置信码确定的输入质量分数(IQS);以及评估关键字段以确定分组质量的链接质量分数(LQS)。该系统和方法允许跨整个数据库的对识别质量的确定而不是使用采样和外推,并且从而导致更高质量的结果,并且因为该系统和方法是客观的,所以其允许在跨竞争的识别质量解决方案之间进行识别质量的比较。



CN 109844785 A

1.一种用于提高识别质量并由此为数据库改善计算性能的方法,所述数据库包括多个数据结构,每个所述数据结构关于多个对象之一,所述方法包括以下步骤:

a.在关于所述对象之一的多个数据结构的每一个中连接包含字符串的多个字段,并基于连接的字符串的相似度计算字符串质量分数(SGS);

b.基于多个输入总体质量分数和输入置信度分数的加权度量来为所述对象计算输入质量分数(IQS);

c.为所述对象计算链路质量分数(LQS),其中所述LQS计算的输入包括唯一链接的数量,与所述对象相关联的数据结构的数量,最常出现的唯一链接的计数,以及所述唯一链接的数量与所述数据结构的数量的比率;

d.将多个LQS异常应用于针对所述对象的所述LQS分数以确定是否应用所述LQS异常中的任一者,并从而通过用与这样的LQS异常相关联的预定义LQS值替换针对所述对象的先前计算的LQS来覆盖针对所述对象的所述先前计算的LQS;

e.基于所述SGS、IQS和LQS的加权度量来为所述对象计算组质量评分(GQG),其中所述GQG是落入限定范围内的单个数值;

f.将多个GQG异常应用于针对所述对象的所述GQG分数以确定是否应用所述GQG异常中的任一者,并通过用与这样的GQG异常相关联的预定义GQG值替换针对所述对象的先前计算的GQG来覆盖针对所述对象的所述先前计算的GQG,以产生最终的GQG(FGQG);以及

g.对所述数据库中的多个对象重复上述步骤,并将边界阈值FGQG值应用于所述数据库,使得低于所述边界阈值FGQG值的对象在所述数据库中被抑制,其中所述数据库中的所述多个对象在被抑制之后包括比应用所述方法步骤之前更少数量的对象。

2.根据权利要求1所述的方法,其特征在于,为所述对象计算GQG的所述步骤进一步包括为所述对象确定组大小的步骤,其中所述组大小是被分组在所述对象中的单独数据结构的所述数量的度量。

3.根据权利要求2所述的方法,其特征在于,为所述对象计算LQS的所述步骤包括为所述对象计算(a)客户商业链接质量分数(CBLQS)或客户链接质量分数(CLQS)以及(b)仅地址链接质量分数(AOLQS)的步骤。

4.根据权利要求3所述的方法,其特征在于,基于关于所述对象的所述数据结构的数量,所述对象被分类为多个组大小中的一个,并且其中基于所述对象的所述组大小,被应用于计算所述GQG值的权重被确定。

5.根据权利要求4所述的方法,其特征在于,所述多个组大小由三个组大小组成,并且其中被应用于对所述三个组大小中的每个计算所述GQG值的所述权重中的至少一个是不同的。

6.根据权利要求3所述的方法,其特征在于,所述对象包括关于商业组的多个数据结构,并且计算CBLQS或CLQS的所述步骤包括计算CBLQS的步骤。

7.根据权利要求6所述的方法,其特征在于,所述对象包括20个或更少的数据结构,并且在计算所述GQG值时应用的所述权重对于所述SQS值是0.10,对于所述IQS值是0.40,对于所述CGLQS值是0.30,并且对于所述AOLQS值是0.20。

8.根据权利要求6所述的方法,其特征在于,所述对象包括21-50个数据结构,并且应用于所述GQG值的所述权重对于所述SQS值是0.10,对于所述IQS值是0.25,对于所述CBLQS值

是0.40,并且对于所述AOLQS值是0.25。

9.根据权利要求6所述的方法,其特征在于,所述对象包括多于50个的数据结构,并且应用于所述GQG值的所述权重对于所述SQS值是0.10,对于所述IQS值是0.00,对于所述CBLQS值是0.30,并且对于所述AOLQS值是0.20。

10.根据权利要求3所述的方法,其特征在于,所述对象包括关于客户组的多个数据结构,并且计算CBLQS或CLQS的所述步骤包括计算CLQS的步骤。

11.根据权利要求10所述的方法,其特征在于,所述对象包括20个或更少的数据结构,并且在计算所述GQG值时应用的所述权重对于所述SQS值是0.10,对于所述IQS值是0.40,对于所述CLQS值是0.30,并且对于所述AOLQS值是0.20。

12.根据权利要求10所述的方法,其特征在于,所述对象包括21-50个数据结构,并且应用于所述GQG值的所述权重对于所述SQS值是0.10,对于所述IQS值是0.40,对于所述CLQS值是0.30,并且对于所述AOLQS值是0.20。

13.根据权利要求10所述的方法,其特征在于,所述对象包括多于50个的数据结构,并且应用于所述GQG值的所述权重对于所述SQS值是0.10,对于所述IQS值是0.40,对于所述CLQS值是0.30,并且对于所述AOLQS值是0.20。

14.一种通过减少数据库中多个对象的总数来提高识别质量的系统,所述系统包括:

a.多个数据结构,每个所述数据结构关于所述多个对象中的一个;

b.识别质量例程,其中所述识别质量例程包括:

i.字符串质量分数(SGS)子例程,所述字符串质量分数子例程被配置为在关于所述对象的多个数据结构的每一个中连接包含字符串的多个字段,并基于连接的字符串的相似度计算SGS;

ii.输入质量分数(IQS)子例程,所述输入质量分数子例程被配置为基于多个输入总体质量分数和输入置信度分数的加权度量来为所述对象计算IQS;以及

iii.链接质量分数(LQS)子例程,所述链接质量分数子例程被配置为:

1.使用包括唯一链接的数量、与所述对象相关联的数据结构的数量、最常出现的唯一链接的计数以及所述唯一链接的数量与所述数据结构的数量的比率来为所述对象计算LQS;以及

2.确定是否应用多个LQS异常中的任一者,并且从而通过用与这样的LQS异常相关联的预定义LQS值替换针对所述对象的先前计算的LQS来覆盖针对所述对象的所述先前计算的LQS;

iv.组质量评分(GQG)子例程,所述组质量评分子例程被配置为:

1.基于所述SGS、IQS和LQS的加权度量来为所述对象计算GQG,其中所述GQG是落入限定范围内的单个数值;以及

2.将多个GQG异常应用于针对所述对象的所述GQG分数以确定是否应用所述GQG异常中的任一者,并通过用与这样的GQG异常相关联的预定义GQG值替换针对所述对象的先前计算的GQG来覆盖针对所述对象的所述先前计算的GQG,以产生最终的GQG(FGQG);以及

v.边界阈值子例程,该边界阈值子例程被配置为将边界阈值FGQG值应用于所述数据库,使得具有低于所述边界阈值FGQG值的FGQG的对象在所述数据库中被抑制,其中所述数据库中的所述对象的总数在应用边界阈值子例程之后小于应用所述边界阈值子例程之前

所述数据库中的所述对象的总数。

15.根据权利要求14所述的系统，其特征在于，所述识别质量例程被进一步配置为确定针对所述对象的组大小，其中所述组大小是被分组在所述对象中的单独数据结构的所述数量的度量。

16.根据权利要求15所述的系统，其特征在于，LQS子例程被进一步配置成为所述对象计算(a)客户商业链接质量分数(CBLQS)或客户链接质量分数(CLQS)以及(b)仅地址链接质量分数(AOLQS)。

17.根据权利要求16所述的系统，其特征在于，所述识别质量子例程被进一步配置为基于被包含在所述对象中的所述数据结构的数量将所述对象分配给多个组大小中的一个，并且其中所述GQG子例程被进一步配置为基于所述对象的所述组大小确定所述GQG值。

18.根据权利要求17所述的系统，其特征在于，所述多个组大小由三个组大小组成，并且其中被应用于对所述三个组大小中的每个计算所述GQG值的权重中的至少一个是不同的。

19.根据权利要求16所述的系统，其特征在于，所述对象包括关于商业组的多个数据结构，并且其中所述LQS子例程被进一步配置为计算CBLQS。

20.根据权利要求19所述的系统，其特征在于，所述对象包括20个或更少的数据结构，并且其中所述GQG子例程被配置为在计算所述GQG值时应用对于所述SQS值的0.10的权重、对于所述IQS值的0.40的权重、对于所述CGLQS值的0.30的权重以及对于所述AOLQS值的0.20的权重。

21.根据权利要求19所述的系统，其特征在于，所述对象包括21-50个数据结构，并且其中所述GQG子例程被配置为在计算所述GQG值时应用对于所述SQS值的0.10的权重、对于所述IQS值的0.25的权重、对于所述CBLQS值的0.40的权重以及对于所述AOLQS值的0.25的权重。

22.根据权利要求19所述的系统，其特征在于，所述对象包括多于50个的数据结构，并且其中所述GQG子例程被配置为在计算所述GQG值时应用对于所述SQS值的0.10的权重、对于所述IQS值的0.00的权重、对于所述CGLQS值的0.30的权重以及对于所述AOLQS值的0.20的权重。

23.根据权利要求16所述的系统，其特征在于，所述对象包括关于客户组的多个数据结构，并且其中所述LQS子例程被进一步配置为计算CLQS。

24.根据权利要求23所述的系统，其特征在于，所述对象包括20个或更少的数据结构，并且其中所述GQG子例程被配置为在计算所述GQG值时应用对于所述SQS值的0.10的权重、对于所述IQS值的0.40的权重、对于所述CLQS值的0.30的权重以及对于所述AOLQS值的0.20的权重。

25.根据权利要求23所述的系统，其特征在于，所述对象包括21-50个数据结构，并且其中所述GQG子例程被配置为在计算所述GQG值时应用对于所述SQS值的0.10的权重、对于所述IQS值的0.40的权重、对于所述CLQS值的0.30的权重以及对于所述AOLQS值的0.20的权重。

26.根据权利要求23所述的系统，其特征在于，所述对象包括多于50个的数据结构，并且其中所述GQG子例程被配置为在计算所述GQG值时应用对于所述SQS值的0.10的权重、对

于所述IQS值的0.40的权重、对于所述CLQS值的0.30的权重以及对于所述AOLQS值的0.20的权重。

27.一种用于提高识别质量的方法,所述方法包括以下步骤:

a.在数据库中的各自关于多个对象之一的多个数据结构的每一个中连接包含字符串的多个字段,并基于连接的字符串的相似度计算字符串质量分数(SGS);

b.基于多个输入总体质量分数和输入置信度分数的加权度量来为所述对象计算输入质量分数(IQS);

c.为所述对象计算链路质量分数(LQS),其中所述LQS计算的输入包括唯一链接的数量,与所述对象相关联的数据结构的数量,最常出现的所述唯一链接的计数,以及所述唯一链接的数量与所述数据结构的数量的比率;

d.基于所述SGS、IQS和LQS的加权度量来为所述对象计算组质量评分(GQG),其中所述GQG是落入限定范围内的单个数值;以及

e.对所述数据库中的多个对象重复上述步骤,并将边界阈值FGQG值应用于所述数据库,使得低于所述边界阈值FGQG值的对象在所述数据库中被抑制,其中所述数据库中的所述多个对象在应用所述边界阈值FGQG值的所述步骤之后包括比应用所述方法步骤之前少至少百分之三(3%)的对象。

28.根据权利要求27所述的方法,其特征在于,进一步包括在应用所述边界阈值FGQG值的步骤之后对数据库应用刷新的步骤,并且其中所述刷新步骤以比应用所述方法步骤之前快至少7%地被执行。

29.根据权利要求27所述的方法,其特征在于,在应用所述边界阈值FGQG值步骤之后,所述数据库占据小至少3%的存储空间。

识别质量管理

技术领域

[0001] 本发明的领域是确定应用于对象数据库(诸如包含各自属于人、家庭、商业等的数据结构的数据库)的识别解决方案的质量。

背景技术

[0002] 数据“识别”可以被定义为一种过程,通过该过程,数据库中的数据结构的集合(诸如记录)被识别为关于同一对象,并且因此那些数据结构被放置在同一组中。例如,给定包含数亿个记录的营销数据库,其中每条记录都关于个人客户,识别可被应用以确定这些记录中的某些记录实际上关于同一客户,即使那些记录中的数据可能不相同。在一个简单的例子中,数据识别可被用于确定都具有相同的地址的名为Jimmy Smith的人和名为James R.Smith的人实际上是同一个人,即使对于这个人在数据库中存在两条单独的记录。其他具体的例子包括识别记录关于同一个人,即使那些记录包含不同的姓氏(诸如当一个女人在结婚后更改她的姓氏时)、具有不同地址但相同或相似名字的人(诸如体现在近期的搬家)以及由于个人收入或兴趣的变化而具有不同“倾向”(例如,对特定产品或服务或产品或服务的类别喜爱)的人。识别不仅限于客户,并且可被应用于家庭、商业或任何其他对象类,其中数据库包含可能关于同一对象的多条记录或其他数据结构。

[0003] 继续营销数据库的例子,高质量组可包括特征,诸如足够相似的商业名称和联系信息;足够相似的地址信息;以及其他信息相似性(诸如电话号码、帐号和电子邮件帐户)。相比之下,低质量组以此类数据的差异为特征,其中可行的现存信息不足以确认记录实际上一起属于同一组。低质量组的特征可包括完全不相似的商业名称或联系信息(诸如完全不同的商业名称)、完全相似的地址信息(诸如具有不同套房号、公寓号、街道号、街道名、邮政编码或ZIP+4码的记录)或不足的或不同的其他现存信息(诸如不同的或缺失的电话号码、帐号或电子邮件帐号)。

[0004] 数据识别在许多行业中具有重要价值。例如,在营销数据库关于客户的情况下,数据识别提供有效的客户标识以创建客户的“单一视图”,从而增加零售商与该客户的互动将为该客户带来积极体验的机会,并且从而推动零售商的销售增长。使用这种单一视图,零售商更好地理解了其与客户的关系,可以能够通过更多的营销渠道与客户接触,并且可以更准确地标识将被客户视为合乎需要的交叉销售机会。营销人员通常不确切地知道他们的客户是谁,因此他们往往无法有效地接触他们的目标受众。随着营销技术的改进,营销人员逐渐认识到,线上营销渠道(电子邮件、社交媒体、网站等)和线下营销渠道(诸如店内营销)实际上是一个单一的营销生态系统;数据识别提高了营销人员利用这些新技术的能力。

[0005] 准确的数据识别带来了许多挑战。举例来说,那些挑战包括如下:信息(诸如关于客户的信息)可能处于不断变化的状态;

[0006] 实现识别解决方案可能成本高昂;用于处理诸如客户之类的数据对象的商业策略(即,商业“规则”)可能随时间不断地修改,因此在每次变更之后需要对数据识别解决方案进行变更;创建包含所有必要信息以执行合理准确的识别的初始数据库可能成本高昂,因

为该初始数据库可能在许多情况下包含数亿条记录或甚至更多;并且事实是对识别解决方案的每次变更必须被全面地测试和审计,以确保变更实际上不会对识别结果产生负面影响。

[0007] 用于创建识别解决方案的现有技术方法,并且特别是对那些识别解决方案的测试和审核,主要是手动执行的。传统方法整体上包括数据样本的手动审查、该样本的审计以及接下来该样本的结果到数据库的外推。使用样本是因为手动测试和审计非常耗时,因此,由于数据库的大小,对整个数据库的审查通常是不切实际的。这种方法导致整体上对识别解决方案的有效性的不完整了解,并且即使仅使用数据样本也需要很多工时才能成功执行,因为有效性在某种程度上与样本大小成比例。由于该方法中的手动元素,结果本质上是主观的,并且因此执行审计和测试的不同人可能获得关于识别解决方案的有效性的不同结果。因此,有效地比较不同的识别方法变得困难或不可能,并且同样地,将一个提供商的识别产品与竞争者的识别产品进行准确地比较是困难的或不可能的。手动过程所需的时间也是一个很大的限制,因为随着数据不断变化,手动结果可能需要很长时间才能生成,因此它不再准确地反映对自审计和外推过程开始以来已更改的数据库的识别解决方案的有效性。这驱使手动解决方案趋向越来越小的样本大小,从而由于过度依赖外推而降低了识别测试和审计的准确性。

[0008] 因此期望的是用于确定客观的识别质量的自动化系统和方法,该自动化系统和方法可以利用数据结构的更大部分或甚至整个数据库进行分析,而不是利用随后外推的小样本,该自动化系统和方法为商业规则变更或与受识别系统影响的数据结构的数据库相关的其他变更提供更快的周转,并且提供改进的方法以跨产品、跨数据库和跨竞争解决方案产品比较识别解决方案的性能度量。

发明内容

[0009] 本发明涉及一种系统和方法,该系统和方法使用数据识别来评估对于任一解决方案的组级别上的总体质量。组质量是指组中存在的数据结构的相关性,并且还包括两个或更多个数据结构(例如,记录)已作为关于同一对象被正确地组合在一起有多可靠。在本发明的某些实施例中,自动识别方法的第一步包含对组内字符串的相似性的分析,并且可能还包含组内任何其他定义信息的重现。字符串分析产生字符串质量分数(SQS)。在识别计算中使用的其他分数是输入质量分数(IQS)和一个或多个链接质量分数(LQS)。这些分数根据基于应用的公式被组合成作为单个数值(诸如但不限于1到100之间的值,其中100是最高质量)的最终组质量评分(FGQG),其被分配给每个组以反映识别该组中的数据结构确实实际上关于同一对象的质量。以这种方式,数值被用作直接涉及总体组质量的评分机制。分数可被解释为百分比;例如,如果发现95的FGQG,这可被描述为95%的结果,表示高质量和高度可靠的数据。阈值边界值可被应用以便抑制被认为是低质量的组(即,低于阈值边界值的数值,从而表示对分组准确性的低置信度)。

[0010] 本发明降低了对于生产数据库的存储要求,因为被视为不可接受的分组数据(即,低于边界阈值值)被抑制,并且所得到的生产环境将仅由高得分组组成。由于组的数量减少,因此随后在生产环境中对于使用该数据的任何应用的处理时间减少,因为更少的组需要被处理。

[0011] 本发明为用户提供数据结构所关于的对象(诸如零售客户)的实时快照,从而消除滞后时间并提高用户及时做出明智的商业决策的能力。这在某些应用中尤其重要,诸如响应于用户访问特定网站或社交媒体出口的在线营销消息的服务,该服务必须被实时执行,以便用户的体验不会因为该服务而降低。在实时在线环境中,使用较老的手动方法进行识别质量测量是不可能的。

[0012] 本发明提供了识别质量的客观测量,从而考虑到变更数据库的某些方面(例如,商业规则)的结果的客观分析,以确定这种变更对识别质量的影响。同样,数字度量可被用于客观地比较竞争的识别产品,以便客户可以做出关于替换识别解决方案的明智决策。

[0013] 尽管简单地使现有手动技术自动化可以潜在地改善样本大小,但是可以看出这种方法没有考虑到如上所述的客观分析和从中产生的优点。本发明表现出与先前分析识别质量的手动方法中使用的主观方法的完全背离。

[0014] 本发明的这些和其他特征、目标及优点将通过结合如以下描述的附图考虑以下对优选实施例和所附权利要求书的详细描述而变得更好理解。

附图说明

[0015] 图1-A是在提供识别解决方案的数据服务公司的数据存储库上实现该系统和方法的一个实施例时观察到的商业记录减少的概览图。

[0016] 图1-B是在提供识别解决方案的数据服务公司的数据存储库上实现该系统和方法的一个实施例时观察到的客户记录减少的概览图。

[0017] 图1-C是在提供识别解决方案的数据服务公司的数据存储库上实现该系统和方法的一个实施例时观察到的记录减少的概览图。

[0018] 图2-A是根据本发明的一个实施例的最终组质量评分(FGQG)的准确性审计的概览图。

[0019] 图2-B是按分数范围的根据本发明的一个实施例的FGQG的准确性审计的概览图。

[0020] 图3-A是根据本发明的一个实施例的在实现每周刷新过程时所处理的记录的数量以及观察到的减少的概览图。

[0021] 图3-B是在每周刷新过程中的总处理时间以及实现本发明一个实施例时观察到的处理时间减少的概览图。

[0022] 图3-C是在实现本发明的一个实施例时观察到的每月大小减少的概览图。

[0023] 图3-D是每月刷新文件处理时间以及在实现本发明的一个实施例时观察到的减少的概览图。

[0024] 图4-A是根据本发明的一个实施例的数据识别过程的高级概览图。

[0025] 图4-B是将FGQG逻辑示出为被并入到图4-A中所示的数据识别解决方案中的示意图。

[0026] 图5-A是在身份识别数据存储库上实现本发明的一个实施例时观察到的用于处理的硬盘存储要求方面的减少的概览图。

[0027] 图5-B是在身份识别数据存储库上实现本发明的一个实施例时观察到的用于处理的硬盘存储要求方面的减少的概览图。

[0028] 图5-C是在身份识别数据存储库上实现本发明的一个实施例时观察到的用于处理

的硬盘存储要求方面的减少的概览图。

具体实施方式

[0029] 在更详细地描述本发明之前，应当理解，本发明不限于所描述的特定实施例和实现，并且在描述特定实施例和实现时所使用的术语仅用于描述那些特定实施例和实现的目的，而并不旨在进行限制，因为本发明的范围将仅通过权利要求书限定。

[0030] 以下定义被用于描述下文阐述的实施例：

[0031] 商业记录：包含专属于单个商业的标识元素的行。例如，商业名称、商业地址、商业电话等。

[0032] 客户记录：包含专属于个体客户的标识元素的行。例如，客户姓名、客户地址、客户电话等。

[0033] 商业组：关于单个商业的商业记录的集群；高质量组包含具有足够相似的特征的记录。

[0034] 客户组：关于单个客户的客户记录集群；高质量组包含具有足够相似的特征的记录。

[0035] 置信码：一种自定义规则集，被设计为帮助用户创建自动决策规则并被用于匹配结果的评估。

[0036] 质量指标：由邮政处理产品生成的值，以指示地址的总体可送达性。

[0037] 链接字段：通常指CLQS、AOLQS和/或BLQS(各自如下所定义)。

[0038] 过度链接组：一个低质量的组，包含可能关于多个客户或商业的记录的集群。

[0039] 帐号与商业和/或客户记录相关联的客户帐号或第三方帐号，一个帐号属于一个个体。

[0040] 保持链接(M)：一串字符，指示数据提供商的身份解析工具先前已标识或摄取了特定记录。

[0041] 导出链接(M)：一串字符，指示特定记录是新的，并且尚未被数据提供商的身份解析工具标识或摄取。

[0042] 组大小分类可被使用以企图更准确地分析数据结构的整个存储库并防止有偏差的结果。表1中示出的以下组大小分类被应用于本发明的一个实施例中，如下所述。

[0043] 组大小分类记录计数范围

极小	2-3 条记录
加小	4-8 条记录
小	9-20 条记录
中	21-50 条记录
大-A	51-100 条记录
大-B	101-250 条记录
大-C	251-500 条记录
加大 A	501-1,000 条记录
加大 B	1,001-3,000 条记录
加大 C	3001-10,000 条记录
加加大 A	10,001-20,000 条记录
加加大 B	20,001-50,000 条记录
加加大 C	50,001+条记录

[0044]

[0045] 表1

[0046] 为了在针对涉及营销的数据对象的实施例中说明商业和客户记录两者，用于评估这两种类型的对象的整体组质量的两种方法已被开发。所描述的实施例中的所有记录将遵循所概述的方法，但是本发明不限于此。每个得分分量的加权与组的大小分类相关。根据组的大小分类，得分分量的加权被分为以下三个部分。注意，CBLQS是客户商业链接质量分数，AOLQS是仅地址链接质量分数，且CLQS是客户链接质量分数，其中所有值在下文更详细地被描述。

[0047] a.极小、加小和小：

[0048] i.FGQG(商业组)＝SQS(.10)+IQS(.40)+CBLQS(.30)+AOLQS(.20)

[0049] ii.FGQG(客户组)＝SQS(.10)+IQS(.40)+CLQS(.30)+AOLQS(.20)

[0050] b.中

[0051] i.FGQG(商业组)＝SQS(.10)+IQS(.25)+CBLQS(.40)+AOLQS(.25)

[0052] ii.FGQG(客户组)＝SQS(.10)+IQS(.40)+CLQS(.30)+AOLQS(.20)

[0053] c.大(A、B、C)、加大(A、B、C)和加加大(A、B、C)

[0054] i.FGQG(商业组)＝SQS(.10)+IQS(0.00)+CBLQS(0.65)+AOLQS(.25)

[0055] ii.FGQG(客户组)＝SQS(.10)+IQS(.40)+CLQS(.30)+AOLQS(.20)

[0056] 1.字符串质量分数-SQS

[0057] 该SQS是通过分析连接字段中存在的字符串的相似性而形成的值在1-100范围内的分数。例如,在Oracle数据库环境中使用Oracle MIN/MAX函数为每组找到(COMPANY_NAME,STREET_NUMBER,PRE_DIRECTIONAL,STREET_SUFFIX,POST_DIRECTIONAL,SECONDARY_UNIT_DESIGNATOR,SECONDARY_NUMBER)和MAX&MIN。它分析数据的物理输入。

[0058] 2.输入质量分数-IQS

[0059] IQS分析表2中示出的以下五个置信码(CDs)和指标(INDs)字段:

[0060]

字段	分数范围
OVERALL_QLTY_IND	1-6
ADDRESS_QLTY_IND	1-9, X
NDIV_CONF_CD	1-6, 0
ADDR_CONF_CD	1-5
BUS_CONF_CD	1-6, 0

[0061] 表2

[0062] 表2的置信码和指标具有三步等式。针对每个码的最高分数被实现在以下等式中:

$$置信码或质量指标计算 = 100 * \left(\frac{(Max+1)-(实际分数)}{Max}\right)$$

$$(Confidence\ or\ Quality\ Indicator\ Calculation = 100 * (\frac{(Max+1)-(Actual\ Score)}{Max}))$$

[0063] 每个分数的比率取决于最高分数。来自置信码和指标的所有分数被加在一起,然后乘以置信码的总权重。所有五个置信码和质量指标的权重均等于20%。每个指标具有一个最大值。诸如X和0(零)之类的值被替换为(Max+1),因此对于其IQS部分该值得到0%的评分。

[0064] 3.链接质量分数-LQS

[0065] 链接质量分数(LQS)是临时组质量评分(GQG)的一般组成部分,其确定不同链接字段的质量。该链接质量分数能被应用于任何链接字段。在优选实施例中,这在通用质量标准(GQG)公式中显示为商业链接质量分数(BLQS),仅地址链接质量分数(AOLQS),客户链接质量分数(CLQS)等。以下表示的表达式是原始表达式的简化版本(更详细的版本在下文进一步被示出)。主要的LQS公式如下:

[0066]

$$LQS = (\frac{100*M*T*(R-U+(1.5-(\frac{U}{R})))}{R^2})$$

$$CLQS = (\frac{100*M*T*(R-U+(1.5-(\frac{U}{R})))}{R^2})$$

$$BLQS = (\frac{100*M*T*(R-U+(1.5-(\frac{U}{R})))}{R^2})$$

$$CBLQS = BLQS + ((1 - \frac{BLQS}{100}) * CLQS)$$

[0067] 其中U是组的链接字段内唯一链接的数量;R是组内的记录数量;U(max)或M是最常

出现的唯一链接的计数，并且T是T等式。

[0068] CBLQS是LQS的变体，其首先计算BLQS，并且接下来基于剩余分数的百分比和CLQS调整分数，以便说明商业组内的客户链接的值。这被用于对包含商业记录和客户记录两者的组进行适当评分。其仅在计算商业组的GQG时被使用。

[0069] 为了计算M，如果M的字符串包含“OOMSUSO”，则使用以下等式：

[0070] $$M = ((1 - \frac{M}{R}) * M + M)$$

[0071] 该等式如异常一样起作用并出现在最后。例如，具有R＝4，U＝3，M＝2的组将得到为0的LQS分数。(如果该等式被应用于该示例组，则M变为3并且该组获得实际的LQS分数，而非0；这不是预期的结果)。

[0072] 如果M的字符串包含“OOMSUS1”，则M＝最常出现的链接组的计数。如果碰到MDP异常(下文的异常5)，则M计算逻辑的这部分将被忽略。

[0073] 为了绘制T等式的图形，y轴是T常数，并且x轴是唯一值与记录的比率，即U/R。具体而言，以下值被应用到T：

[0074] i.如果组大小＝“极小”，“加小”且(U/R)≤0.364：则T＝-0.5(U/R)+1

[0075] ii.如果组大小＝“极小”，“加小”且(U/R)＞0.364：则T＝-1.2(U/R)+1.2

[0076] iii.如果组大小!＝“极小”，“加小”：则T＝-0.5(U/R)+1

[0077] 4.异常

[0078] 以下五个异常应用到LQS公式(共5个)：

[0079] 异常1：该异常使具有值为1的M的组的LQS保持为零。该异常减少了处理时间。应用的逻辑是M必须≥2。如果M＜2，则LQS＝0。请注意，在进行保持计算之后，上述比较中引用的M是M的实例。

[0080] 异常2：该异常绕过了LQS计算的其余部分，并给与具有完全链接匹配的组值为100的LQS。这也减少了处理时间。应用的逻辑是，如果M＝R，则LQS＝100。

[0081] 异常3：该异常允许在针对商业组的计算中为U计数为空值。该异常仅影响商业组。应用的逻辑是，如果Record_Type(记录类型)＝′B(商业)′且Business_Link(商业链接)为Null(空)，则针对U，但不针对M，计数空值。异常4：该异常允许在针对客户组的计算中为U计数为空值。该异常仅影响客户组。应用的逻辑是，如果Record_Type(记录类型)＝′C(客户)′且Consumer_Link(客户链接)为Null(空)，则针对U，但不针对M，计数为空值。

[0082] 异常5(M减少百分比或MDP)该异常允许两个最常出现的唯一链接的比较，以便更好地识别过度链接的组。在计算M时，我们需要创建一个循环，以便找到对应于两个最常出现的唯一链接的前两个M(M1和M2)。记录总数(R)也需要为该计算被导出。该计算和循环将在保持M计算之前发生。这些值将被用于创建两个比率来比较，以便确定该组是否真的过度链接并且M值需要被降低。如果M字符串包含“OOMSUSO”，则由此产生的M将是被用于计算调整后的M的值。如果字符串不包含“OOMSUSO”，则保持M的计算被跳过，并且该计算的结果被赋值给M。该逻辑取代上面列出的关于为M并且包含“OOMSUSO”的字符串的逻辑。受影响的组大小是加大A以及更大的那些组大小。所有组类型都会受到影响。变量是M1(从通过组的第一循环导出的值，以找到最常出现的非空M)；M2(从通过组的第二循环导出的值，以找到下一个最常出现的非空M值)；以及R(组中记录总数)。应用的公式是：

[0083] MDP＝M2/M1

[0084] MDPCV＝M1/R

[0085] MDPCV是组中最常出现的链接与组中记录总数的比率,根据等式：

[0086] Let MDPT＝(M1+M2)/R

[0087] If MDPT＞＝.9and string that occurs most often contains

[0088] ‘00MSUS0’：

[0089] M＝M1

[0090] Run M Calculation

[0091] Else If MDPT＞＝.9and string that occurs most often

[0092] contains‘00MSUS1’

[0093] M＝M1

[0094] Else If MDPT＜.9

[0095] If MDP＞MDPCV and string that occurs most

[0096] often contains‘00MSUS0’：

[0097] M＝M1(1-MDP)

[0098] Run M Calculation

[0099] Else If MDP＞MDPCV and string that occurs most

[0100] often contains‘00MSUS1’:M＝M1(1-MDP)

[0101] Else If MDP＜＝MDPCV andstring that occurs most often

[0102] contains‘00MSUS0’:M＝M1

[0103] Run M Calculation

[0104] Else If MDP＜＝MDPCV and string that occurs most often

[0105] ccontains‘00MSUS1’:M＝M1

[0106] (让MDPT＝(M1+M2)/R

[0107] 如果MDPT＞＝.9且最常出现的字符串包含“00MSUS0”：

[0108] M＝M1

[0109] 运行M计算

[0110] 否则如果MDPT＞＝.9且最常出现的字符串包含“00MSUS1”：

[0111] M＝M1

[0112] 否则如果MDPT＜.9

[0113] 如果MDP＞MDPCv且最常出现的字符串包含“00MSUS0”：

[0114] M＝M1(1-MDP)

[0115] 运行M计算

[0116] 否则如果MDP＞MDPCV且最常出现的字符串包含“00MSUS1”：

[0117] M＝M1(1-MDP)

[0118] 否则如果MDP＜＝MDPCV且最常出现的字符串包含“00MSUS0”：

[0119] M＝M1

[0120] 运行M计算

[0121] 否则如果MDP＜＝MDPCV且最常出现的字符串包含“00MSUS1”:M＝M1)

[0122] 本发明描述的实现还包括影响总体FGQG的许多其他异常。帐号异常存在以允许对

组中帐号的分析。该异常影响所有组大小以及商业组和客户组。在初始GQG被计算之后，等式(如下所述)将被应用于每个组。空字段被包括在该计算中。出于该异常的目的，任何和所有空值都将被视为一个“唯一”。出于任何其他目的，空字段将没有值。空或空白将不被视为帐号UMAX(ANU)。例如，如果在具有10条记录的组中不存在帐号，则ANU＝0，UAN＝1，R＝10。在异常计算中，这些将被首先计算。在这里，ANU是针对帐号字段的组中的M;UAN是唯一帐号的总数;并且R是组中记录的总数。适用的等式是：

[0123] $$FGQG = ((\frac{100-GQG}{3}))((\frac{ANU}{R}))((1-\frac{UAN-1}{R})) + GQG$$

[0124] 商业匹配异常存在，以确保在某些字段上具有完全匹配的组获得为100的其链接分数。该异常影响所有组大小，但仅影响商业组类型。该分数将在帐户计算后被计算，并且将只有在以下字段值完全匹配时才会被计算：COMPANY_NAME、SUBURB、STATE以及POSTCODE。仅当前面列出的所有四个字段都完全匹配时，才应用以下等式：

[0125] BLQS＝SQS

[0126] 客户完全匹配异常确保在某些字段上具有完全匹配的小的组获得为100的其链接分数。这仅适用于具有50分或更低的AOLQS和/或CLQS的客户极小或加小组。应用的逻辑是，如果整个组中STREET_NUMBER、STREET_NAME、SUBURB、STATE、POSTCODE相等，则AOLQS＝SQS。

[0127] 名字/姓氏异常被实现，以便解决其中分数由记录中的名字和姓氏的颠倒被显著降低的问题。它适用于所有组大小，但仅适用于客户类型的组。应用的逻辑如下：

[0128] String＝FIRSTNAME+LASTNAME of the first record

[0129] String 1＝FIRSTNAME+LASTNAME

[0130] String 2＝LASTNAME+FIRSTNAME

[0131] IfString 1＝＝String

[0132] Then Match Count++;

[0133] Else if

[0134] String 2＝＝String

[0135] Then Match Count++;

[0136] Else if

[0137] Do nothing

[0138] IfmatchRate＞＝.5&&CLQS＜50

[0139] then CLQS＝matchRate*100;

[0140] (字符串＝第一条记录的FIRSTNAME+LASTNAME

[0141] 字符串1＝FIRSTNAME+LASTNAME

[0142] 字符串2＝LASTNAME+FIRSTNAME

[0143] 如果字符串1＝＝字符串

[0144] 则匹配计数++;

[0145] 否则如果

[0146] 字符串2＝＝字符串

[0147] 则匹配计数++;

[0148] 否则如果

[0149] 不做任何事

[0150] 如果匹配率＞＝.5并且CLQS＜50

[0151] 则CLQS＝匹配率*100;)

[0152] 关键指标计算(KIC)异常比较组中的电话号码。空值不被计入该异常中。要利用针对该异常的逻辑,P必须大于0且C必须大于或等于2。该异常影响所有组大小和组类型。适用的公式是：

[0153] $$FGQG = FGQG + \left(\left(\frac{100-FGQG}{\left(2-\frac{FGQG}{100}\right)}\right)(KIC)\right)$$

[0154] KIC＝AO

[0155] $$O = \frac{P}{R}$$

[0156] 其中A＝协定;O＝发生;P＝存在的电话号码;C＝最常出现的唯一电话号码的计数;R＝组内的记录数;以及U＝该字段中唯一电话号码的数量。应用的逻辑是：

[0157] IfP＞0AND C＞＝2(如果P＞0且C＞＝2)

[0158] 则

[0159] $$FGQG = FGQG + \left(\left(\frac{100-FGQG}{\left(2-\frac{FGQG}{100}\right)}\right)(KIC)\right)$$

[0160] 否则：

[0161] FGQG＝FGQG

[0162] 次要号码异常比较组内的次要号码(字段id＝Secondary_Number(次要号码))。空值不被计入该异常中。要利用针对该异常的逻辑,P必须大于0且P/R必须大于或等于基于组大小定义的常量。列出的变量与KIC异常逻辑中使用的变量不同。这将在每个其他异常之后被计算。该异常影响所有组大小和组类型。适用的公式是：

[0163] $$FGQG = FGQG - \left(FGQG * \left(\frac{1-\frac{C}{P}}{1.3}\right)\right)$$

[0164] P/R

[0165] C/P

[0166] 其中P＝存在的次要号码(不包括空值);C＝最常出现的唯一次要号码的计数(不包括空值);以及R＝组内的记录数。应用的逻辑是：

[0167] 1.如果P＞0且SIZE_BAND＝＝‘TINY(极小)’且P/R＞＝.79

[0168] 则

[0169] $$FGQG = FGQG - \left(FGQG * \left(\frac{1-\frac{C}{P}}{1.3}\right)\right)$$

[0170] 2.否则如果P＞0且SIZE_BAND＝＝‘XSMALL(加小)’且P/R＞＝.79

[0171] 则

[0172] $$FGQG = FGQG - \left(FGQG * \left(\frac{1-\frac{C}{P}}{1.3}\right)\right)$$

[0173] 3.否则如果P＞0且SIZE_BAND＝＝‘SMALL(小)’且P/R＞＝.7

[0174] 则

[0175] $$FGQG = FGQG - \left(FGQG * \left(\frac{1-\frac{C}{P}}{1.3}\right)\right)$$

[0176] 4.否则如果P＞0且SIZE_BAND＝＝‘MEDIUM(中)’且P/R＞＝.77

[0177] 则

[0178] $$FGQG = FGQG - (FGQG * (\frac{1-C}{1.3}))$$

[0179] 5.否则如果P＞0且SIZE_BAND＝＝‘LARGE_A(大A)’且P/R＞＝.59

[0180] 则

[0181] $$FGQG = FGQG - (FGQG * (\frac{1-C}{1.3}))$$

[0182] 6.否则如果P＞0且SIZE_BAND＝＝‘LARGE_B(大B)’且P/R＞＝.59

[0183] 则

[0184] $$FGQG = FGQG - (FGQG * (\frac{1-C}{1.3}))$$

[0185] 7.否则如果P＞0且SIZE_BAND＝＝‘LARGE_C(大C)’且P/R＞＝.53

[0186] 则

[0187] $$FGQG = FGQG - (FGQG * (\frac{1-C}{1.3}))$$

[0188] 8.否则如果P＞0且SIZE_BAND＝＝‘XLARGE_A(加大A)’且P/R＞＝.53则

[0189] $$FGQG = FGQG - (FGQG * (\frac{1-C}{1.3}))$$

[0190] 9.否则如果P＞0且SIZE_BAND＝＝‘XLARGE_B(加大B)’且P/R＞＝.50则

[0191] $$FGQG = FGQG - (FGQG * (\frac{1-C}{1.3}))$$

[0192] 10.否则如果P＞0且SIZE_BAND＝＝‘XLARGE_C(加大C)’且P/R＞＝.50则

[0193] $$FGQG = FGQG - (FGQG * (\frac{1-C}{1.3}))$$

[0194] 11.否则如果P＞0且SIZE_BAND＝＝‘XXLARGE_A(加加大A)’且P/R＞＝0.47

[0195] 则

[0196] $$FGQG = FGQG - (FGQG * (\frac{1-C}{1.3}))$$

[0197] 12.否则如果P＞0且SIZE_BAND＝＝‘XXLARGE_B(加加大B)’且P/R＞＝.46

[0198] 则

[0199] $$FGQG = FGQG - (FGQG * (\frac{1-C}{1.3}))$$

[0200] 13.否则如果P＞0且SIZE_BAND＝＝‘XXLARGE_C(加加大A)’且P/R＞＝.42

[0201] 则

[0202] $$FGQG = FGQG - (FGQG * (\frac{1-C}{1.3}))$$

[0203] 14.否则

[0204] FGQG＝FGQG

[0205] LQS公式之前已被介绍,以下是对原始LQS公式的更详细的解释。

[0206] $$LQS, AOLQS, CLQS, BLQS = (1 - (\frac{(U-(1.5-\frac{U}{R}))}{R}))((\frac{M}{R})(T)) * 100$$

[0207] 1.变量

[0208] a.R

[0209] i.仅组内的记录的计数

[0210] b.U

[0211] i.组内的唯一链接的计数

[0212] c.M

[0213] i.通过对组内两个最常出现的链接计数,比较公式,并且基于公式比较输入逻辑来计算。

[0214] ii.最常出现的链接的计数被定义为变量M1。具有第二最大计数的链接被定义为变量M2。

[0215] iii.公式如下:MDP＝M2/M1,MDPCV＝M1/R,MDPT＝(M1+M2)/R。

[0216] iv.链接可被保持或导出。保持链接比导出链接更有价值

[0217] v.为了说明这个,实现了附加的逻辑:

[0218] IfMDPT＞＝.9&&M1is Maintained(如果MDPT＞＝.9并且M1被保持):

[0219] M＝M1

[0220] $M = ((1 - \frac{M}{R}) * M + M)$

[0221] Else If MDPT＞＝.9&&M1is Derived(否则如果MDPT＞＝.9并且M1被导出):

[0222] M＝M1

[0223] Else If MDPT＜.9(否则如果MDPT＜.9)

[0224] IfMDP＞MDPCV&&M1is Maintained(如果MDP＞MDPCV并且M1被保持):

[0225] M＝M1(1-MDP)

[0226] $M = ((1 - \frac{M}{R}) * M + M)$

[0227] Else IfMDP＞MDPCV&&M1is Derived(否则如果MDP＞MDPCV并且M1被导出):

[0228] M＝M1(1-MDP)

[0229] Else IfMDP＜＝MDPCV&&M1is Maintained(否则如果MDP＜＝MDPCV并且M1被保持):

[0230] M＝M1

[0231] $M = ((1 - \frac{M}{R}) * M + M)$

[0232] Else IfMDP＜＝MDPCV M1is Derived(否则如果MDP＜＝MDPCV M1被导出):

[0233] M＝M1

[0234] vi. $M = ((1 - \frac{M}{R}) * M + M)$

[0235] 上面的公式简单地将M的百分比加回到自身,以便恰当地说明保持链接比导出链接更有价值的事实。

[0236] 2.分量

[0237] a.当最常出现的链接的计数(M)等于记录(R)时,LQS等于100。当计算M时,空链接字段被完全忽略。符合这些准则的组不使用该表达式。

[0238] i.其背后的逻辑是:当M在组中与R一样频繁出现时,链接上存在100％的匹配,因此该组需要得到为100的LQS分数。这也减少了处理时间。

[0239] b.这部分$1 - (\frac{(U-(1.5-\frac{U}{R}))}{R})$是唯一链接的数量和记录数之间的关系。

[0240] i.表达式的这一部分背后的逻辑是这样的:随着组内U与R的比率增加,分数将降

低，并且反之亦然；随着U与R的比率减小，分数将提高。唯一链接的数量由唯一链接数(U)与记录(R)的关系被减少。例如，具有值为5的R的组中的3个唯一(U)将获得低于具有值为3的U且值为100的R的组的分数。

[0241] ii.常数1.5基于表达式的优化图形化表示被选择。

[0242] iii.最常出现的链接计数(M)小于2的组不使用该表达式。

[0243] c.该部分$\left(\left(\frac{M}{R}\right)(T)\right)$是M和R之间的关系。

[0244] i.然后将该关系乘以相应的T常数(T)。T由U和R之间的线性关系计算。精确的线性关系基于组大小被选择(下文更多详细信息)。

[0245] ii.表达式的这一部分背后的逻辑是这样的：随着组内M与R的比率增加，分数将提高，并且反之亦然；随着M与R的比率减小，分数将降低。

[0246] d.该部分*100将小数转换为整数。

[0247] e.AOLQS、BLQS和CLQS遵循相同的公式和逻辑。我们选择遵循相同的逻辑，因为所有链接字段的表现都相同(除了它们为不同字段集显示链接信息的事实)，以便于使用以及避免混淆。

[0248] f.T等式

[0249] i.T等式基于组大小(R)加速或减速U与R的关系。它增加了第三维以改变比率，以便基于组大小(R)提高或抑制分数降低的速率。

[0250] ii.这背后的逻辑是：随着组大小的变化，U与R的比率和M与R的比率的准确性变得更少地表明组质量。T增加了第三个因素来解决前面提到的观察所引起的担忧并提高组内比率的准确性。

[0251] (1)例如，具有10比200或10比2000的U与R的比率的组与具有值为10的U和值为20的R的组相比可以具有相似或更低的质量，即使较大的组具有较低的U与R的比率。这是由于可能存在20条记录的10个组或200条记录的10个组的事实。

[0252] iii.考虑到它们的线性表示来选择T等式，以便分别基于U与R的关系和基于R来修改M与T的关系。这背后的逻辑是，唯一链接的计数越大，组内的“链接不一致”就越大。在T等式的图上，y轴是T常数，并且x轴是唯一与记录的比率，即U/R。

[0253] iv.如果组大小＝＝“极小”或“加小”且$\left(\frac{U}{R}\right) \leqslant 0.364$，使用：

[0254] $$T = -0.5\left(\frac{U}{R}\right) + 1$$

[0255] (1)我们选择这些限制以确保该等式会真正表明组质量。

[0256] (2)-0.5存在以便使其更能表明组的U与R的比率。

[0257] (3)+1存在以便将等式保持在正确的图形界限中。

[0258] (4)我们选择将组范围限制为极小和加小，因为该等式仅在这些大小的组内保持为真。

[0259] (5)我们选择$\left(\frac{U}{R}\right) \leqslant 0.364$以确保该等式会真正表明组质量。

[0260] v.如果组大小＝＝“极小”或“加小”且$\left(\frac{U}{R}\right) > 0.364$，使用：

[0261] $$T = -1.2\left(\frac{U}{R}\right) + 1.2$$

[0262] (1)-1.2存在以便以更快的速率降低分数。

[0263] (2)+1.2存在以设置最小值,以使结果始终为正。

[0264] (3)我们选择将组范围限制为极小和加小,因为该等式仅在这些大小的组内真正具有指示性。

[0265] (4)我们选择(U/R)＞0.364以确保该等式会真正表明组质量。

[0266] vi.如果组大小!＝极小或加小,使用:

[0267] $$T = -0.5(\frac{U}{R}) + 1$$

[0268] (1)-0.5存在以便使其更能表明组的U与R的比率。

[0269] (2)+1存在以便将等式保持在正确的图形界限中。

[0270] (3)*CLQS然后将其与CLQS相乘以获得CLQS分数的百分比。我们选择将修改后的BLQS与CLQS相乘,以考虑组中的客户记录。

[0271] (4)BLQS+该数字被加回BLQS。我们选择将修改后的分数加回BLQS,以具有对组中的商业对客户记录的准确描绘。我们选择加到BLQS而不是CLQS,因为该公式将仅被应用到商业组,因此我们希望BLQS和商业记录在链接分数中具有更重的权重。这也允许CLQS弥补商业链接分数中的一些缺点或差异。

[0272] 以下是原始LQS公式的简化的细分,以便人们理解之前列出的简化版本。

[0273]
$$\begin{aligned}
LQS &= (1 - (\frac{(U-(1.5-\frac{U}{R}))}{R}))((\frac{M}{R})(T)) * 100 \\
&= (\frac{M*T*100}{R}) - (\frac{(U-(1.5-\frac{U}{R}))}{R}) * (\frac{M*T*100}{R}) \\
&= (\frac{M*T*100}{R}) - (\frac{U*M*T*100-(1.5-\frac{U}{R})*M*T*100}{R^2}) \\
&= (\frac{R}{R})(\frac{M*T*100}{R}) - (\frac{U*M*T*100-(1.5-\frac{U}{R})*M*T*100}{R^2}) \\
&= (\frac{R*M*T*100}{R^2}) - (\frac{U*M*T*100-(1.5-\frac{U}{R})*M*T*100}{R^2}) \\
&= (\frac{R*M*T*100-U*M*T*100+(1.5-\frac{U}{R})*M*T*100}{R^2}) \\
&= (\frac{100*M*T*(R-U+(1.5-\frac{U}{R}))}{R^2})
\end{aligned}$$

[0274] 现在参考图1-A至3-D,应用上述本发明实施例的测试运行的结果可被说明。用于第一次测试的数据包括792,250,327条商业记录和678,160,909条客户记录。用于该测试的边界阈值是值为40的FGQG分数,这意味着对于得到小于40分的FGQG的每个组,该组在最终结果中被抑制。对该测试数据使用本文描述的实施例,被标识为商业的记录数减少了3.77%(或29,883,069条记录),并且被标识为客户的记录数减少了0.33%(或2,234,017条记录)。这些结果在图1-A至图1-C中被图形化地示出。可以理解,通过将可应用数据库中的记录数减少数百万,显著的过程改善在识别技术平台中被认可。减少的记录数降低了存储需求,同时也减少了处理时间。后者是以下事实的结果:应用于数据库的每个商业过程都需要对每条记录进行迭代,并且因此记录计数的减少会导致针对每个商业过程的处理时间缩短。由于单个数据库可以在任何给定时间段(诸如一周或一个月)中被用于许多商业过程,因此对于营销提供商或其数据服务提供商的总处理时间的减少可以是非常显著的。

[0275] 使用与刚刚描述的测试中相同的识别系统,进一步的测试被执行以验证系统和方法的总体准确性。第n个(随机)样本测试床被创建,并且被从商业记录集中手动审计。为了防止偏差,共有1,022个组的此样本测试床,涵盖了所有组大小分类并反映了所有FGQG范围(评分范围从1至100)。据观察,针对1,022个采样的组中的511个组,值为60或更高的FGQG被发现。换句话说,总的采样的组的50%得到值为60(即60%)或更高的FGQG。换句话说,总的

采样的组的16.6％得到值为90％或更高的FGQG。图2-A以饼图示出了总体商业准确性，并且图2-B按FGQG大小范围分解了商业准确性结果。

[0276] 再次使用具有同一营销数据库的同一营销技术平台，如本文所述的本发明的实现被应用于双月刷新过程数据文件。该过程采用最旧的存储库标识符(总记录包括完整数据存储库的1/13)并通过数据识别对其进行重新处理，以应用由内部产品增强引起的任何必要变更。接下来，所有记录接着被分配新链接，并且新的存储库标识符在必要时被分配。依次的，此过程可确保跨数据库和数据存储库之间的一致性，以形成单一且准确的客户视图。被选择用于测试的双月刷新数据文件由148,396,077条记录组成，并且该文件被处理达约249.6小时。在应用如本文所述的本发明的实现之后，双周刷新数据的大小减少了10,899,000条记录(记录大小减少了多于7％)，并且总处理时间减少了19.2小时。由此实现的显著改善由图3-A和图3-B的饼图来图形化示出。被选择用于测试的每月刷新数据文件由296,792,154条记录组成，并且该文件被处理达约499.2小时。在应用如本文所述的本发明的实现之后，每月刷新数据的大小减少了21,798,000条记录，并且总处理时间减少了38.5小时，从而表示多于7％的处理时间改善。由此在每月的基础上实现的显著改善由图3-C和图3-D的饼图来图形化示出。

[0277] 再次使用具有同一营销数据库的同一营销技术平台，如本文所述的本发明的实现被应用于另一测试环境。该测试环境由626,260,073条记录组成，并且静态数据消耗550GB的总空间。在应用该系统和方法之前处理该记录集需要3.1TB的分配。以下三个测试用例利用了如上定义的环境。

[0278] 测试用例1在数据调用方面使用了非常保守的，值为30的阈值分数，并且将被发送用于处理的记录数减少了17,264,573条记录或2.8％。该记录的减少相当于17GB或3％的存储空间的节省。该减少还最小化了用于处理的分配空间。由于设定的阈值分数为30，需求减少了96GB或3％。存储需求的改善由图5-A的饼图图形化示出。

[0279] 测试用例2在数据调用方面使用了稍微不那么保守的，值为40的阈值分数，并且将被发送用于处理的记录数减少了29,883,069条记录或4.8％。该记录的减少相当于27.5GB或5％的存储空间的节省。该减少还最小化了用于处理的分配空间。由于设定的阈值分数为40，需求减少了155GB或5％。存储需求的改善由图5-B的饼图图形化示出。

[0280] 测试用例3在数据调用方面使用了更不那么保守的，值为50的阈值分数，并且将被发送用于处理的记录数减少了47,053,346条记录或7.5％。该记录的减少相当于41.5GB或5％的存储空间的节省。该减少还最小化了用于处理的分配空间。由于设定的阈值分数为50，需求减少了233GB或7.5％。存储需求的改善由图5-C的饼图图形化示出。

[0281] 测试用例1-3被用于展示通过实现该系统和方法创建的数据传输时间和网络资源分配。在使用ftp、sftp或使用connect direct(直接连接)的sftp连接进行数据传输期间，这些节省可在下游被测量。假设传输速度为30Mbps，且原始文件大小为550GB，则传输时间约为1天19小时45分钟。该系统和方法可以将传输时间减少到1天16小时27分钟(超过6％的改善)到1天18小时24分钟之间的任何时间。如果这些资源是另一个网络事务所需的，这也可以减少所需的带宽分配并且仍然产生与原始文件相同的传输时间。假设原始文件大小为550GB，传输时间大约为1天19小时45分钟，则需要30Mbps的传输速度。随着系统和方法的实现，文件大小将降至508.5GB至533GB之间的任何大小，这将仅需要28Mbps至29Mbps的传输

速度以在相似时间内完成传输。

[0282] 因此可以理解，上述本发明的实现为用户提供了包含关于对象的数据结构的数据库(诸如包含记录的营销数据库)，以确定在该数据库中利用数据识别的解决方案的整体组质量。这允许用户基于期望的边界阈值来抑制被认为是低质量的组，该边界阈值可以基于期望的风险等级和特定应用来修改。由于这种方法是客观的，并且其参数易于修改，因此该方法允许商业规则逻辑被更改，以满足新的和不断变化的行业需求。该方法进一步允许更快的处理和更小的数据库占用空间，从而导致与被用于实现本发明的该实现的技术有关的显著成本节省。

[0283] 如本文描述的本发明旨在用作总体识别系统的一部分，其一个例子由图4-A示出。输入文件10可以表示例如包含客户记录或商业记录的来自商家的数据文件。在块12处，这些文件被传输到营销服务提供商。基于客户账户标识符(Cust_Acct_ID)对数据的第一次搜索完成，并且匹配被标识并被发送以匹配覆盖块48。在块16处，源数据操作被执行，以便标准化用于进一步处理的剩余数据。在步骤18处，数据接着根据每个记录表示的数据的类型被划分：在该示例中，数据被划分成全球数据组20、美国数据组22、部分数据组24，并且其他(即，除此之外不能被分类的被拒绝的数据)被划分到组26。来自组26的数据立即进入覆盖块48，而其他数据进行进一步处理。

[0284] 部分数据组24被摄取在数据附加服务32处，目的是附加数据以使来自部分数据组24的数据完整并因此可用。这可包括例如为数据记录确定最佳地址。拒绝(即，仍然找不到匹配)接着被直接发送到覆盖块48。匹配被发送到美国卫生/数据质量块30，其也是美国数据组22被发送到的地方。在美国卫生/数据质量块30处，数据被标准化、清理以及更新。这可包括例如标识名称异常(诸如缩写、拼写错误、格式问题，不完整内容和其他各种内容问题)。它可包括对地址信息的改善(诸如更正以验证交货地点、基于居住的改善、新的物流以及其他各种内容问题)。它可包括评估电子邮件地址、语法验证/更正以及电子邮件地址的更改(COA)。它还可包括电话确认或更正、NPA/NXX内容验证、区域代码的重新调整、固定电话与移动电话的标识以及标记私人电话和禁止的能力。在全球卫生/数据质量块28处对全球数据组20执行类似的过程。

[0285] 在数据从美国卫生/数据质量块30移动之后FGQG打分块34被执行。在该示例中，FGQG打分仅对美国数据执行，但是本发明不限于此。该块将参考下面的图4-B被详细描述。额外的重新卫生/数据质量过程在块36处被执行，并且块36的输出与块28的输出结合，作为对最终数据质量操作块38的输入。然后处理进行到基于客户链接块40的搜索，其中如果匹配发生，则数据被移动到覆盖块48。如果未找到匹配，则在匹配块42处对所有级别，而不仅仅是客户链接，执行匹配，，并且再次，如果找到匹配，则数据被移动到覆盖块48。未找到匹配的剩余数据被发送到块44以被添加到数据库，之后对于新被添加的记录在块46处执行事务完成，并且处理的这最后一个线程进行到覆盖块48。

[0286] 在覆盖块48处，已由营销服务提供商接收并且现在已通过识别系统被处理的新数据被覆盖到主存储库数据库上。然后事务输出文件52在块50处被转换并被传输回营销商。

[0287] 现在转到图4-B，FGQG打分块34内的处理可被更详细地描述。该处理被示出为由过程和数据输入和输出组成的“虚拟机”，尽管它可在专用计算硬件或专门编程的通用计算硬件上被实现。处理在子例程64处开始，其中用户可登录到虚拟机(VM)并将识别管理器(RM)

客户帐户标识符(Cust_Acct)和交叉引用表复制到虚拟机。这里的输入是客户账户表66和交叉引用表68。处理进行到用于自定义表视图的选项#1,块70,或用于组合的客户账户/交叉引用表的创建的选项#2,块72。这些过程分别在块74和78处被执行。块74的输出是自定义表视图54,块78的输出是组合表84。块76启动FGQG打分过程,该过程在块80处由SQL中的一系列打分查询执行。这里的输出匹配在上文提供的计算中被描述的各种分数,并且包括块56处的字符串质量分数SQS、块58处的输入质量分数IQS、块60处的商业链接质量分数BLQS、块62处的客户链接质量分数CLQS、块86处的客户/商业链接质量分数CBLQS、块88处的仅地址质量链路分数AOLQS、块90处的异常以及块92处的最终组质量评分分数FGQG。最后,在块92处执行清理操作,如较旧的识别管理和排序表被移除,因为仅所有表的最新版本被保留以便节省存储空间。

[0288] 除非以其他方式说明,否则本文中所使用的所有技术和科学术语具有如本发明所属的本领域的普通技术人员共同理解的相同含义。虽然类似于或等同于本文所描述的方法或材料的任何方法和材料可在实践或测试本发明时使用,本文中描述了有限数目的示例性的方法和/或材料。本领域的那些技术人员将领会,更多的修改是可能的,而不背离本文中的发明概念。

[0289] 本文中使用的所有术语应当以与上下文一致的尽可能最宽的方式来解释。当本文中使用分组时,该组中的所有个体成员以及该组中所有可能的组合和子组合均旨在被个体地包括。当在此说明范围时,该范围旨在包括该范围内的所有子区域和单个点。本文中引用的所有参考都被通过援引纳入在此到不存在与本说明书的公开不一致的程度。

[0290] 本发明已参考某些优选和替换实施例来描述,这些实施例旨在仅为示例性的而非旨在限制如所附权利要求书中阐述的本发明的整个范围。

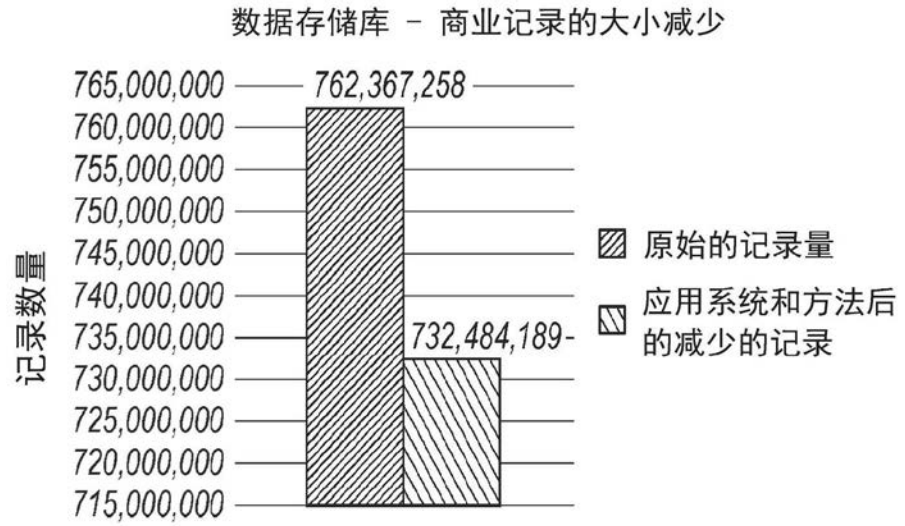

数据存储库 - 商业记录的大小减少
765,000,000
760,000,000
755,000,000
750,000,000
745,000,000
740,000,000
735,000,000
730,000,000
725,000,000
720,000,000
715,000,000
记录数量
762,367,258
732,484,189
原始的记录量
应用系统和方法后的减少的记录

图1A

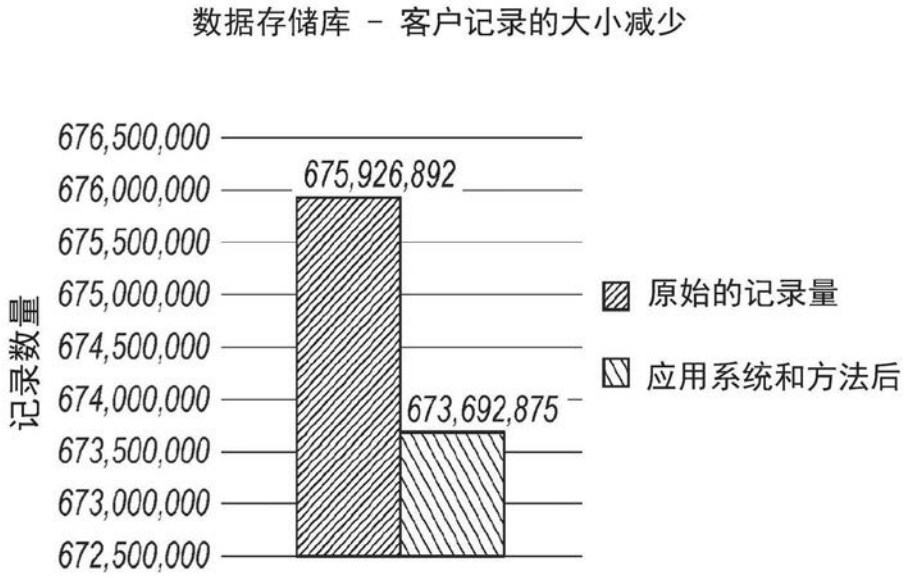

数据存储库 - 客户记录的大小减少
676,500,000
676,000,000
675,500,000
675,000,000
674,500,000
674,000,000
673,500,000
673,000,000
672,500,000
记录数量
675,926,892
673,692,875
原始的记录量
应用系统和方法后

图1B

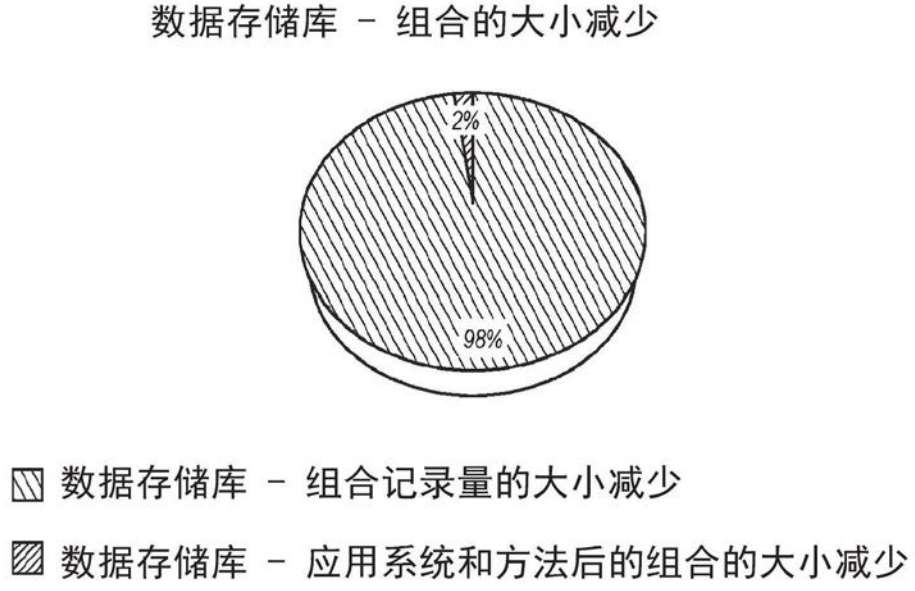

数据存储库 - 组合的大小减少
2%
98%
数据存储库 - 组合记录量的大小减少
数据存储库 - 应用系统和方法后的组合的大小减少

图1C

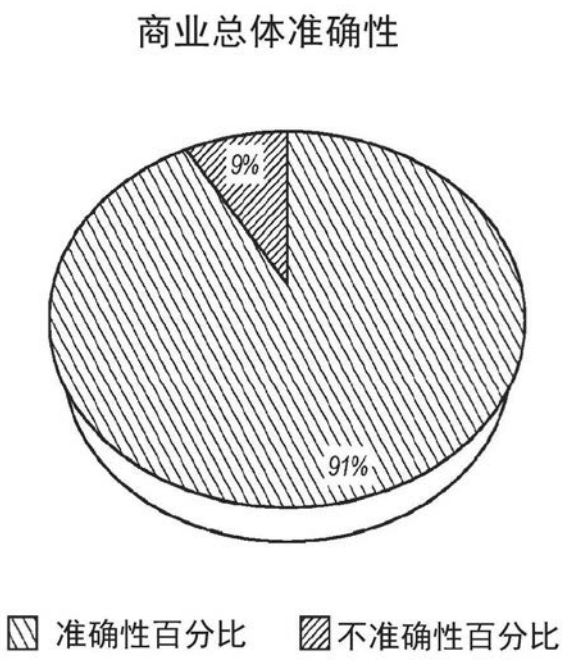

商业总体准确性
9%
91%
准确性百分比
不准确性百分比

图2A

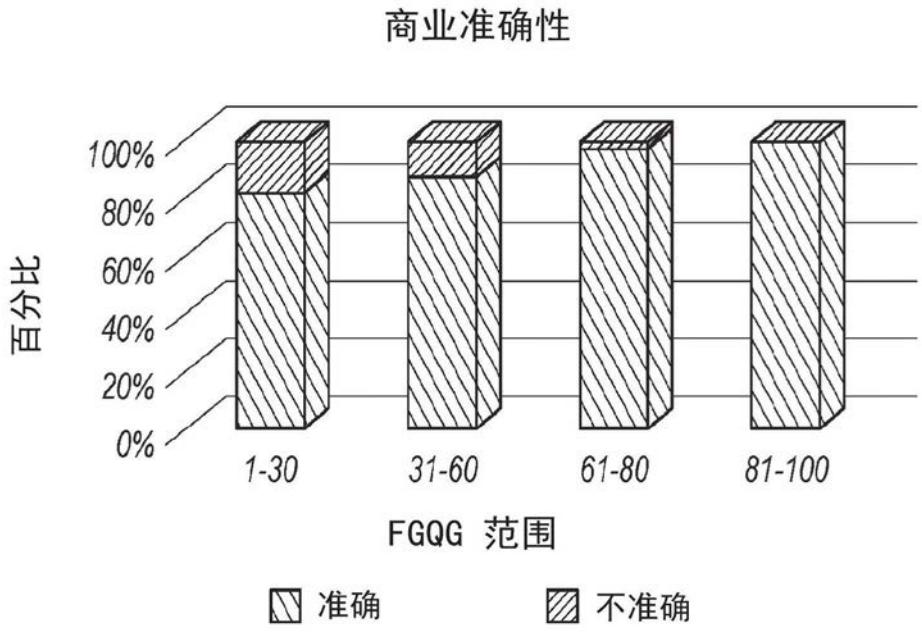

商业准确性
100%
80%
60%
40%
20%
0%
百分比
1-30
31-60
61-80
81-100
FGQG 范围
准确
不准确

图2B

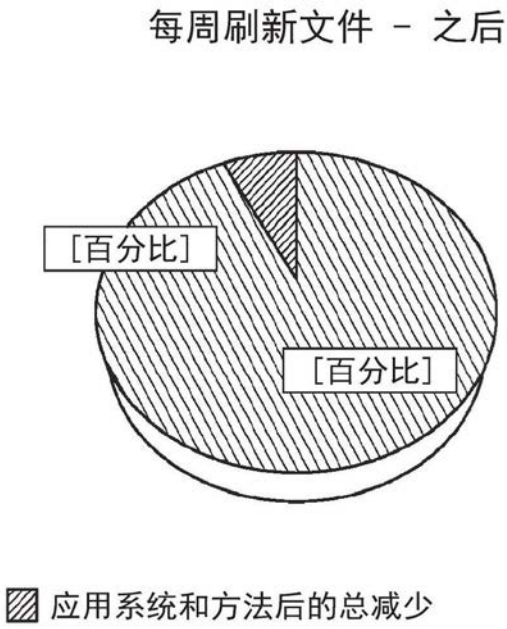

每周刷新文件 - 之后
[百分比]
[百分比]
应用系统和方法后的总减少

图3A

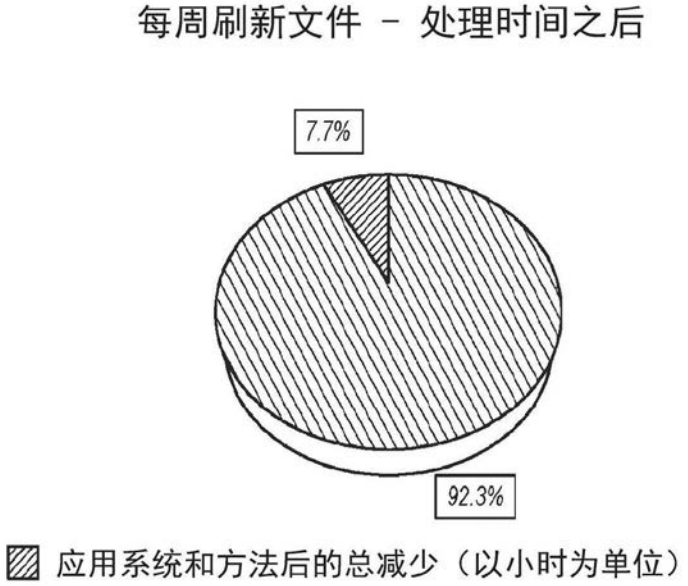

每周刷新文件 - 处理时间之后
7.7%
92.3%
应用系统和方法后的总减少（以小时为单位）

图3B

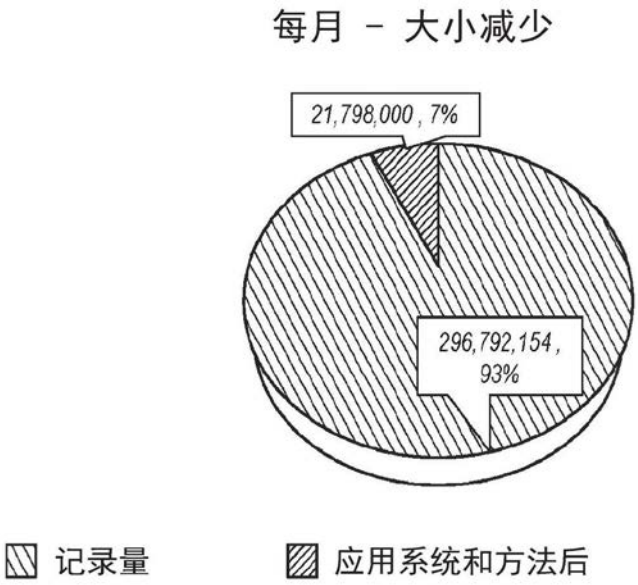

每月 - 大小减少
21,798,000 , 7%
296,792,154 , 93%
记录量
应用系统和方法后

图3C

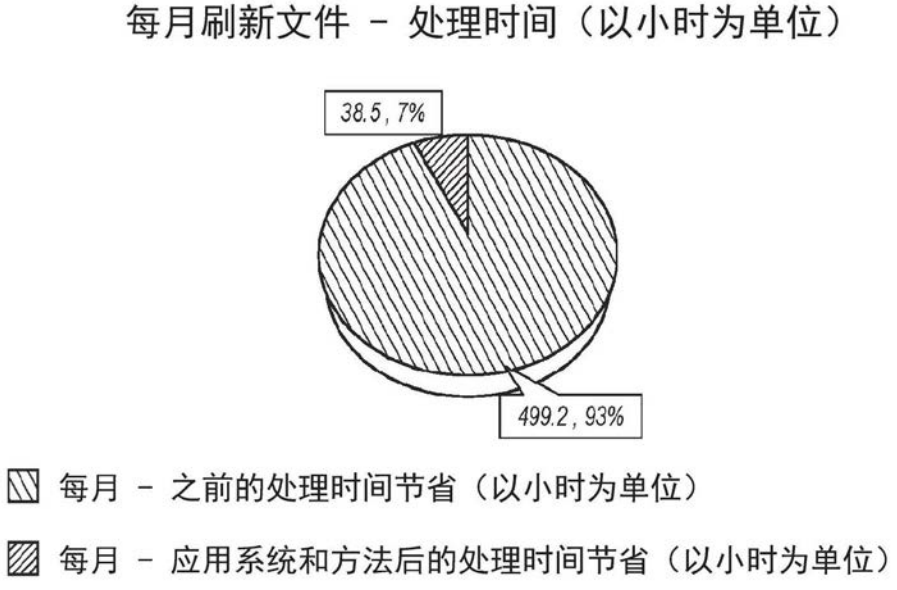

每月刷新文件 - 处理时间（以小时为单位）
38.5, 7%
499.2, 93%
每月 - 之前的处理时间节省（以小时为单位）
每月 - 应用系统和方法后的处理时间节省（以小时为单位）

图3D

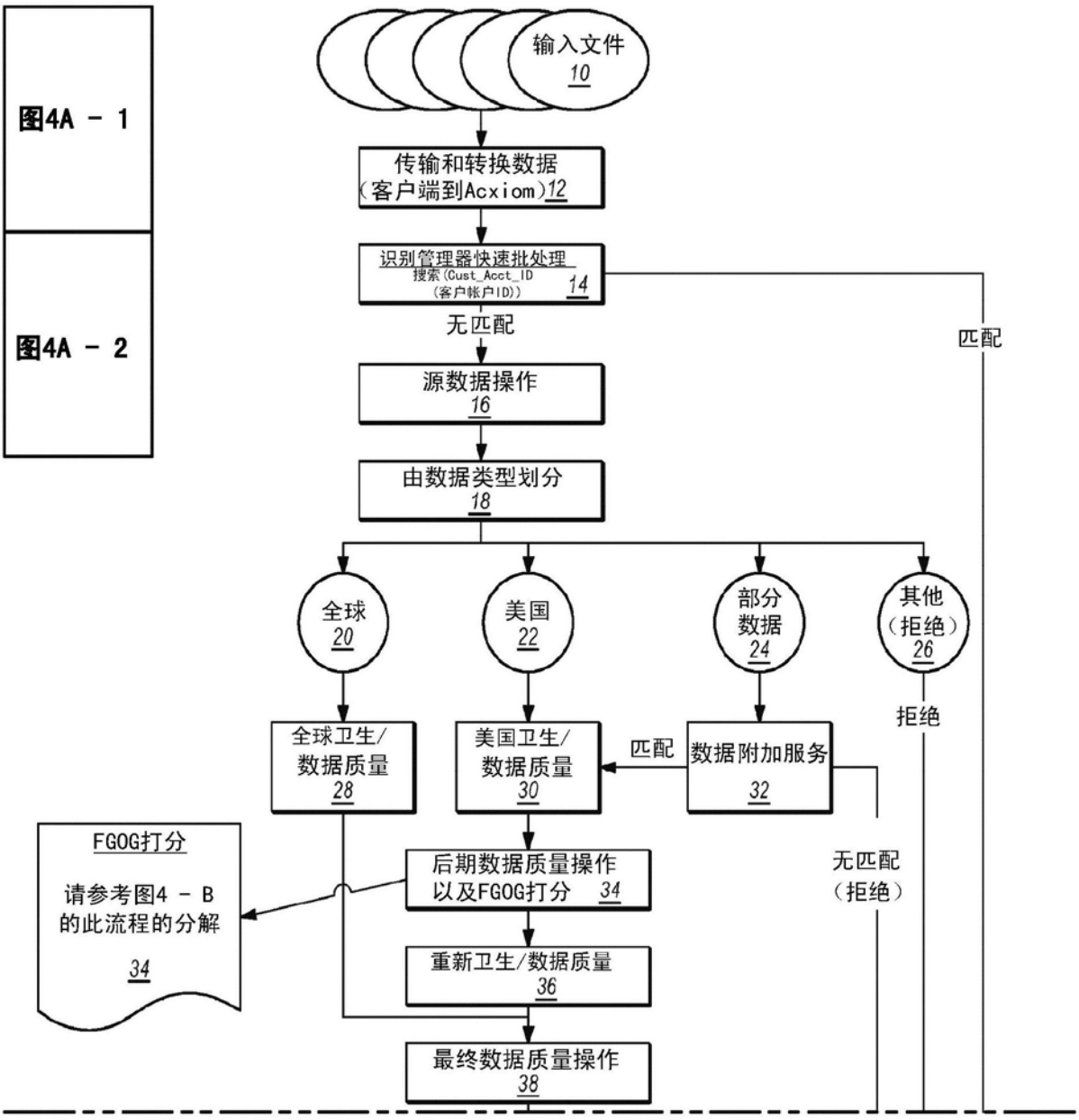


图4A - 1
图4A - 2
输入文件
10
传输和转换数据
(客户端到Acxiom)12
识别管理器快速批处理
搜索(Cust_Acct_ID
(客户帐户ID))
14
匹配
无匹配
源数据操作
16
由数据类型划分
18
全球
20
美国
22
部分
数据
24
其他
(拒绝)
26
拒绝
全球卫生/
数据质量
28
美国卫生/
数据质量
30
匹配
数据附加服务
32
FGOG打分
请参考图4 - B
的此流程的分解
34
后期数据质量操作
以及FGOG打分 34
无匹配
(拒绝)
重新卫生/数据质量
36
最终数据质量操作
38

图4A-1

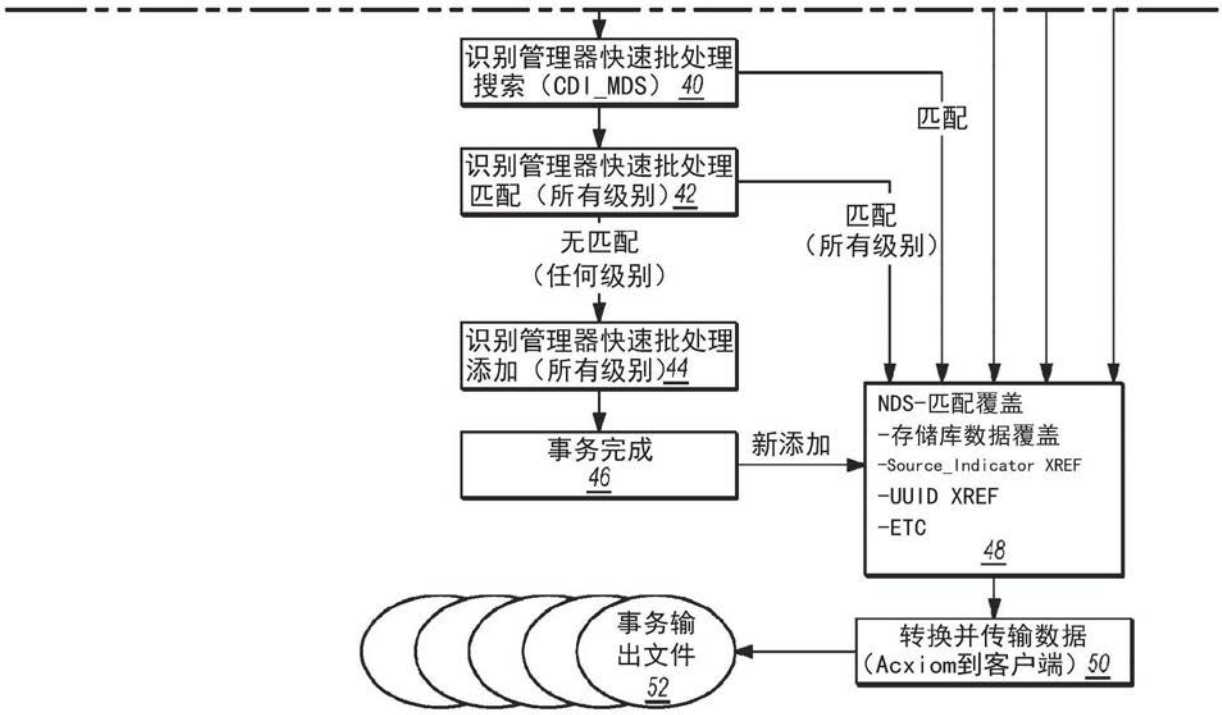

识别管理器快速批处理
搜索（CDI_MDS） 40
匹配
识别管理器快速批处理
匹配（所有级别）42
匹配
（所有级别）
无匹配
（任何级别）
识别管理器快速批处理
添加（所有级别）44
事务完成
46
新添加
NDS-匹配覆盖
-存储库数据覆盖
-Source_Indicator XREF
-UUID XREF
-ETC
48
事务输
出文件
52
转换并传输数据
（Acxiom到客户端）50

图4A-2

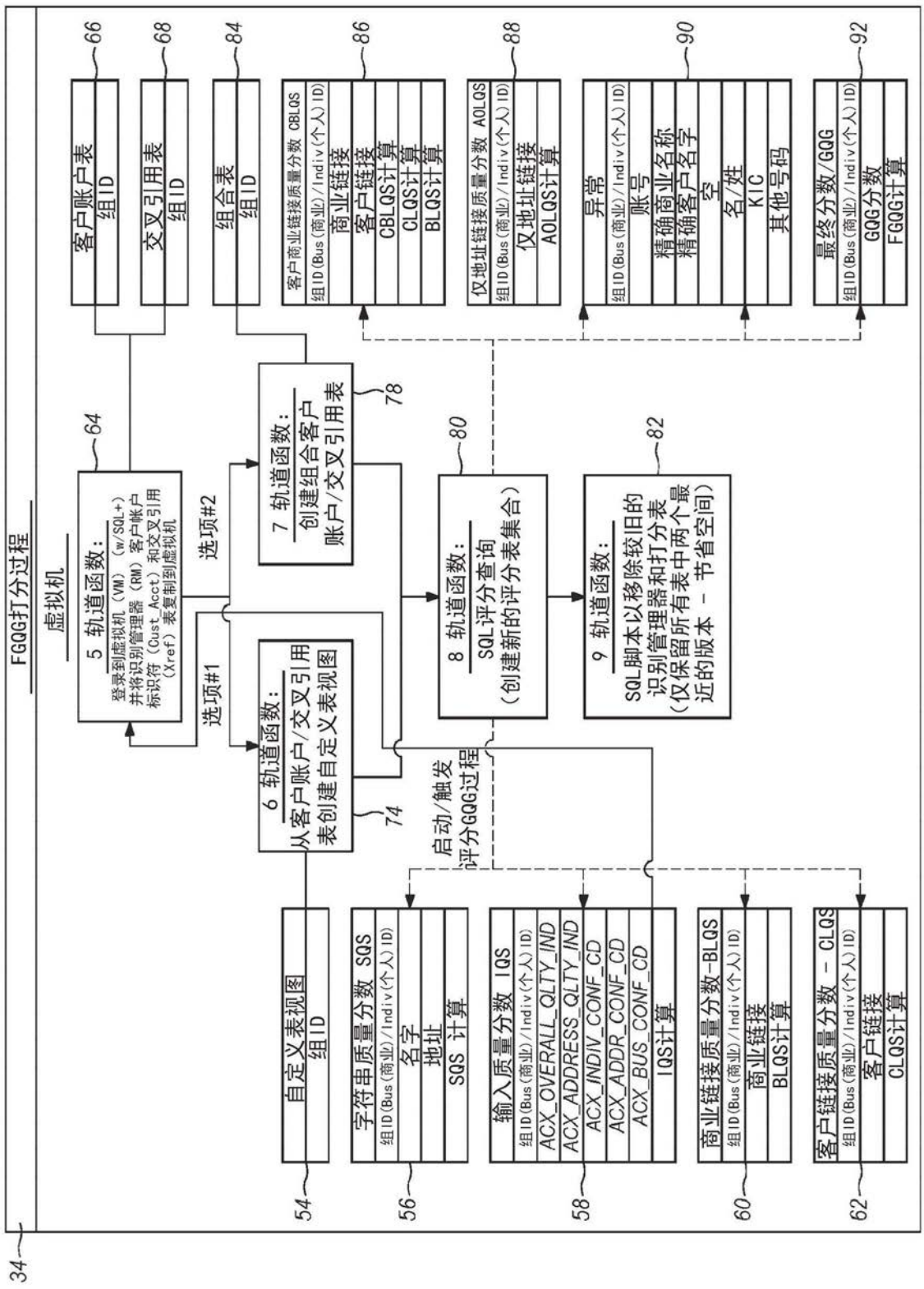

FGQG打分过程
虚拟机
34
5 轨道函数：
登录到虚拟机（VM）（w/SQL+）并将识别管理器（RM）客户帐户标识符（Cust Acct）和交叉引用（Xref）表复制到虚拟机
64
选项#1
选项#2
6 轨道函数：
从客户账户/交叉引用表创建自定义表视图
74
7 轨道函数：
创建组合客户账户/交叉引用表
78
8 轨道函数：
SQL评分查询
（创建新的评分表集合）
80
9 轨道函数：
SQL脚本以移除较旧的识别管理器和打分表（仅保留所有表中两个最近的版本 - 节省空间）
82
启动/触发评分GQG过程
客户账户表
组ID
66
交叉引用表
组ID
68
组合表
组ID
84
客户商业链接质量分数 CBLQS
组ID(Bus(商业)/Indiv(个人)ID)
商业链接
客户链接
CBLQS计算
CLQS计算
BLQS计算
86
仅地址链接质量分数 AOLQS
组ID(Bus(商业)/Indiv(个人)ID)
仅地址链接
AOLQS计算
88
异常
组ID(Bus(商业)/Indiv(个人)ID)
账号
精确商业名称
精确客户名字
空
名/姓
KIC
其他号码
90
最终分数/GQG
组ID(Bus(商业)/Indiv(个人)ID)
GQG分数
FGQG计算
92
自定义表视图
组ID
54
字符串质量分数 SQS
组ID(Bus(商业)/Indiv(个人)ID)
名字
地址
SQS 计算
56
输入质量分数 IQS
组ID(Bus(商业)/Indiv(个人)ID)
ACX_OVERALL_QLTY_IND
ACX_ADDRESS_QLTY_IND
ACX_INDIV_CONF_CD
ACX_ADDR_CONF_CD
ACX_BUS_CONF_CD
IQS计算
58
商业链接质量分数-BLQS
组ID(Bus(商业)/Indiv(个人)ID)
商业链接
BLQS计算
60
客户链接质量分数 - CLQS
组ID(Bus(商业)/Indiv(个人)ID)
客户链接
CLQS计算
62

图4B

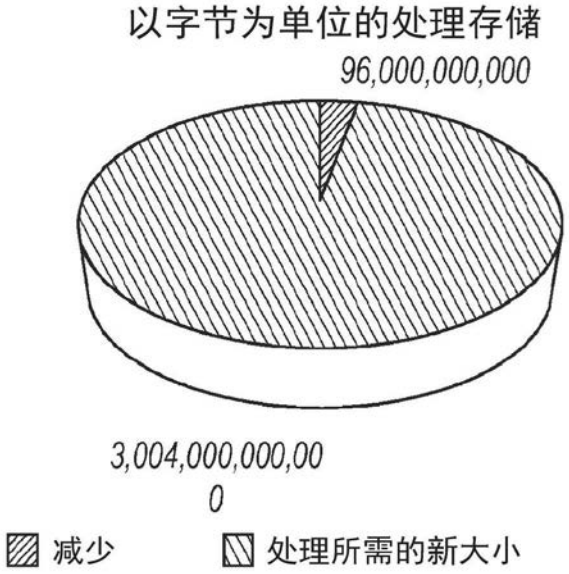

以字节为单位的处理存储
96,000,000,000
3,004,000,000,00
0
减少
处理所需的新大小

图5A

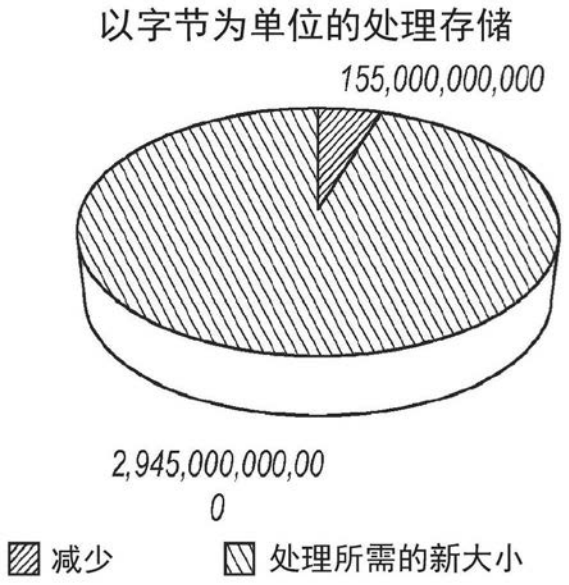

以字节为单位的处理存储
155,000,000,000
2,945,000,000,00
0
减少
处理所需的新大小

图5B

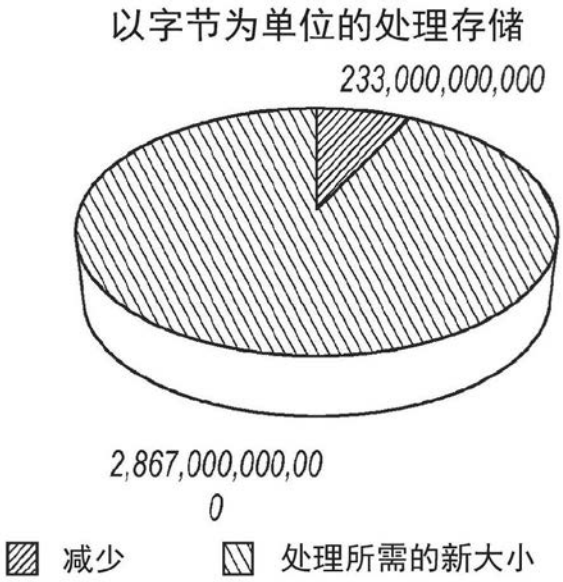


图5C