(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 101978348 B
(45)授权公告日 2015.11.25

(21)申请号 200880128089.2
(22)申请日 2008.12.30
(30)优先权数据
12/015,085 2008.01.16 US
(85)PCT国际申请进入国家阶段日
2010.09.16
(86)PCT国际申请的申请数据
PCT/US2008/088530 2008.12.30
(87)PCT国际申请的公布数据
WO2009/091494 EN 2009.07.23
(73)专利权人 起元技术有限责任公司
地址 美国马萨诸塞州
(72)发明人 阿伦·安德森
(74)专利代理机构 北京林达刘知识产权代理事务所(普通合伙) 11277
代理人 刘新宇
(51)Int.Cl.
G06F 7/00(2006.01)
(56)对比文件
WO 2006102227 A2,2006.09.28,
US 20070239741 A1,2007.10.11,
审查员 李昕宇

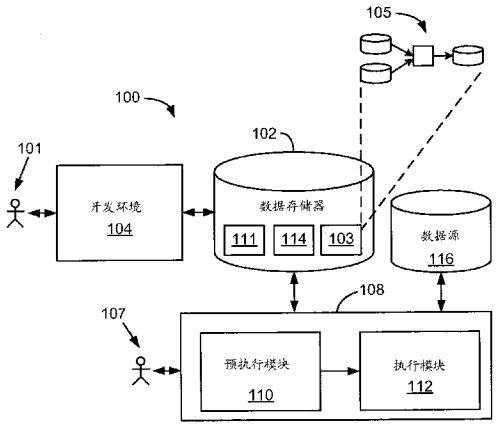
权利要求书3页 说明书18页 附图2页

(54)发明名称

管理关于近似串匹配的档案

(57)摘要

在一个方面,一般,描述了一种用于管理档案的方法。所述档案用来确定与在记录中出现的串相关联的近似匹配。该方法包括处理记录以确定一组串代表的步骤,所述串代表对应于在记录中出现的串。该方法也包括为该组中的至少一些串代表的每个产生多个接近代表的步骤,所述多个接近代表的每个根据该串中至少一些相同字符而产生。该方法也包括存储条目在档案中的步骤。每个所存储的条目表示基于它们各自的接近代表的在至少两个串之间的潜在近似匹配。



CN 101978348 B

1. 一种用于管理用来确定与在记录中出现的串相关联的近似匹配的档案的方法，该方法包括步骤：

处理记录以确定一组串代表，所述串代表对应于在记录中出现的串；

为该组中的至少一些串代表的每个产生多个删除变型，所述多个删除变型的每个通过从对应串删除一个或多个字符而产生；

存储条目在档案中，每个条目表示基于它们各自的删除变型的在至少两个串之间的潜在近似匹配；

对于该组中的至少一些串代表的每个，确定记录中对应串的出现频率；以及

对于该组中的至少一些串代表的每个，产生表示所述档案中相应串的重要性的重要性值，

其中，所述重要性值基于包括该串的出现频率和在档案中作为该串的潜在近似匹配而代表的至少一些串的出现频率的总和；以及

基于存储的条目，匹配来自一数据集的记录与来自另一数据集的另一记录。

2. 根据权利要求 1 所述的方法，其中，每个串代表包括一个串。

3. 根据权利要求 2 所述的方法，其中，每个接近代表包括该串中的至少一些相同字符。

4. 根据权利要求 3 所述的方法，其中，为该组中的给定串产生多个删除变型的步骤包括产生其每个使得从给定串中删除不同的字符的删除变型。

5. 根据权利要求 4 所述的方法，其中，为该组中的给定串产生多个删除变型的步骤包括产生其每个使得从给定串中删除单个字符的删除变型。

6. 根据权利要求 5 所述的方法，其中，为该组中的给定串产生多个删除变型的步骤包括产生其中的至少一些使得从给定串中删除多个字符的删除变型。

7. 根据权利要求 4 所述的方法，其中，产生其每个是从给定串中删除不同的字符的删除变型的步骤包括：如果给定串比预定长度短则产生其每个使得从给定串中删除单个字符的删除变型，以及如果给定串比预定长度长则产生其中的至少一些使得从给定串中删除多个字符的删除变型。

8. 根据权利要求 1 所述的方法，其中，该重要性值是基于除以该总和的值来产生的。

9. 根据权利要求 1 所述的方法，还包括步骤：通过确定短语中的串是否对应于近似匹配来确定包括多个串的不同短语是否对应于近似匹配，其中基于它们相应的重要性值来选择所述短语中的串。

10. 根据权利要求 9 所述的方法，其中短语中串的重要性值是基于该总和的，并且基于该串的长度、短语中串的位置、其中出现串的记录的字段、和其中该字段出现的记录的源中的至少一个。

11. 根据权利要求 1 所述的方法，还包括步骤：对于档案中至少一些条目的每个产生与条目相关联的分值，其量化至少两个串之间的潜在近似匹配的质量。

12. 根据权利要求 11 所述的方法，还包括步骤：通过将与条目相关联的分值与阈值比较来确定与条目相关联的串是否对应于近似匹配。

13. 根据权利要求 11 所述的方法，其中，该分值基于用来确定至少两个串之间的潜在近似匹配的各自删除变型之间的对应性。

14. 根据权利要求 1 所述的方法，其中，处理记录以确定对应于在记录中出现的串的一

组串代表包括：修改在至少一个记录中出现的串以产生修改串，以便包括在该组串代表中。

15. 根据权利要求 14 所述的方法，其中，修改串包括去除或替换标点。

16. 根据权利要求 14 所述的方法，其中，修改串包括将串编码为不同的代表。

17. 根据权利要求 16 所述的方法，其中，修改串包括将串编码为数字代表。

18. 根据权利要求 17 所述的方法，其中，将串编码为数字代表包括：将串中的每个字符映射为素数，并且将串表示为映射到串中的字符的素数的乘积。

19. 根据权利要求 1 所述的方法，其中，该档案包括基于来自用户的输入而表示至少两个串之间的潜在近似匹配的至少一些条目。

20. 根据权利要求 1 所述的方法，其中，每个条目包括串和分值，其中在串之间存在潜在近似匹配，而所述分值量化串之间的潜在近似匹配的质量。

21. 根据权利要求 1 所述的方法，还包括步骤：对于该组中的至少一些串代表的每个，基于相应串的出现频率而产生代表相应串的重要性的重要性值。

22. 根据权利要求 11 所述的方法，还包括步骤：使用档案中的条目来识别作为可能的误判的潜在近似匹配。

23. 根据权利要求 22 所述的方法，其中，在第一串和第二串之间的可能的误判潜在近似匹配是基于记录中第一串的出现频率和记录中第二串的出现频率来识别的。

24. 根据权利要求 22 所述的方法，其中，基于档案中存储的 n-gram 频率来识别可能的误判潜在近似匹配，n-gram 是 n 字母单词片段。

25. 根据权利要求 22 所述的方法，还包括步骤：响应于作为可能的误判的潜在近似匹配的识别，调整与代表潜在近似匹配的条目相关联的分值。

26. 根据权利要求 1 所述的方法，还包括步骤：对于在与记录中的串对应的、该组中的串代表产生重要性值，包含计算如下内容的除以总和的值：

记录中对应串的出现频率，和

在档案中作为记录中对应串的潜在近似匹配而代表的串的出现频率。

27. 一种用于管理用来确定与在记录中出现的串相关联的近似匹配的档案的系统，该系统包括：

用于处理记录以确定一组串代表的部件，所述串代表对应于在记录中出现的串；

用于对于该组中的至少一些串代表的每个产生多个删除变型的部件，所述多个删除变型的每个通过从对应串删除一个或多个字符而产生；

用于存储条目在档案中的部件，每个条目表示基于它们各自的删除变型的在至少两个串之间的潜在近似匹配；

用于对于该组中的至少一些串代表的每个确定记录中对应串的出现频率的部件；

用于对于该组中的至少一些串代表的每个，产生表示所述档案中相应串的重要性的重要性值的部件，

其中，所述重要性值基于包括该串的出现频率和在档案中作为该串的潜在近似匹配而代表的至少一些串的出现频率的总和；以及

用于基于存储的条目匹配来自一数据集的记录与来自另一数据集的另一记录的部件。

28. 一种用于管理用来确定与在记录中出现的串相关联的近似匹配的档案的系统，该系统包括：

数据源，用于存储记录；

计算机系统，包括

用于处理该数据源中的记录以确定一组串代表的部件，所述串代表对应于在记录中出现的串；

用于对于该组中的至少一些串代表的每个，产生多个删除变型的部件，所述多个删除变型的每个通过从对应串删除一个或多个字符而产生；

用于对于该组中的至少一些串代表的每个，确定记录中对应串的出现频率的部件；

用于对于该组中的至少一些串代表的每个，产生表示所述档案中相应串的重要性的重要性值的部件；以及

与所述计算机系统耦接的数据存储器，存储包括条目的档案，每个条目表示基于它们各自的删除变型的在至少两个串之间的潜在近似匹配，

其中，所述重要性值基于包括该串的出现频率和在档案中作为该串的潜在近似匹配而代表的至少一些串的出现频率的总和；以及

用于基于存储的条目匹配来自一数据集的记录与来自另一数据集的另一记录的部件。

管理关于近似串匹配的档案

技术领域

[0001] 本发明涉及管理关于近似串匹配的档案 (archive)。

背景技术

[0002] 关于近似串匹配（也被称做“模糊”或“不精确”串匹配或搜索）的各种技术被用来根据串度量（也叫做“相似函数”）寻找在一定偏差内匹配给定样式串的串。被搜索的串可以是被称为“文本”的较大串的子串、或者可以是包含在例如数据库的记录中的串。串度量的一种类别是“编辑距离”。编辑距离的一个示例是 Levenshtein 距离，其对需要将一个串转换为另一串的编辑操作（字符的插入、删除或替换）的最小数目进行计数。近似串匹配包括在线匹配和离线匹配，在在线匹配中在匹配开始之前无法处理（或“编索引”）要被搜索的文本，在离线匹配中在匹配开始之前能够处理文本。

发明内容

[0003] 在一个方面中，一般，描述一种用于管理确定与记录中出现的串关联的近似匹配的档案的方法。该方法包括：处理记录以确定对应于在记录中出现的串的一组串代表；为该组中的至少一些串代表的每个产生多个接近代表，所述多个接近代表的每个是根据该串中的至少一些相同字符而产生的；以及在档案中存储条目，所述条目的每个表示基于它们各自的接近代表的在至少两个串之间的潜在近似匹配。

[0004] 各个方面能够包括以下特征的一个或多个。

[0005] 每个串代表包括一个串。

[0006] 每个接近代表包括该串中的至少一些相同字符。

[0007] 为该组中的给定串产生多个接近串包括产生其每个使得从给定串中删除不同的字符的接近串。

[0008] 为该组中的给定串产生多个接近串包括产生其每个使得从给定串中删除单个字符的接近串。

[0009] 为该组中的给定串产生多个接近串的步骤包括产生其中的至少一些使得从给定串中删除多个字符的接近串。

[0010] 产生其每个是从给定串中删除不同的字符的接近串包括：如果给定串比预定长度短则产生其每个使得从给定串中删除单个字符的接近串，以及如果给定串比预定长度长则产生其中的至少一些使得从给定串中删除多个字符的接近串。

[0011] 该方法还包括对于该组中的至少一些串代表的每个，确定记录中对应串的出现频率。

[0012] 该方法还包括：对于该组中的至少一些串代表的每个，基于包括该串的出现频率和在档案中作为该串的潜在近似匹配而代表的至少一些串的出现频率的总和产生表示相应串的重要性的重要性值。

[0013] 该重要性值是基于该总和的反来产生的。

[0014] 该方法还包括：通过确定短语中的串是否对应于近似匹配来确定包括多个串的不同短语是否对应于近似匹配，其中基于它们相应的重要性值来选择所述短语中的串。

[0015] 短语中串的重要性值是基于该总和的，并且基于该串的长度、短语中串的位置、其中出现串的记录的字段、和其中该字段出现的记录的源中的至少一个。

[0016] 该方法还包括：对于档案中至少一些条目的每个产生与条目相关联的分值，其量化至少两个串之间的潜在近似匹配的质量。

[0017] 该方法还包括：通过将与条目相关联的分值与阈值比较来确定与条目相关联的串是否对应于近似匹配。

[0018] 该分值基于用来确定至少两个串之间的潜在近似匹配的各自接近代表之间的对应性。

[0019] 处理记录以确定对应于在记录中出现的串的一组串代表包括：修改在至少一个记录中出现的串以产生修改串，以便包括在该组串代表中。

[0020] 修改串包括去除或替换标点。

[0021] 修改串包括将串编码为不同的代表。

[0022] 修改串包括将串编码为数字代表。

[0023] 将串编码为数字代表包括：将串中的每个字符映射为素数，并且将串表示为映射到串中的字符的素数的乘积。

[0024] 该档案包括基于来自用户的输入而表示至少两个串之间的潜在近似匹配的至少一些条目。

[0025] 在另一方面，一般，描述了一种存储在计算机可读介质中的计算机程序，用于管理用来确定与在记录中出现的串相关联的近似匹配的档案。该计算机程序包括指令，用于促使计算机来：处理记录以确定一组串代表，所述串代表对应于在记录中出现的串；为该组中的至少一些串代表的每个，产生多个接近代表，所述多个接近代表的每个根据该串中的至少一些相同字符而产生；以及存储条目在档案中，每个条目表示基于它们各自的接近代表的在至少两个串之间的潜在近似匹配。

[0026] 在另一方面，一般，描述了一种用于管理用来确定与在记录中出现的串相关联的近似匹配的档案的系统。该系统包括：用于处理记录以确定一组串代表的部件，所述串代表对应于在记录中出现的串；用于对于该组中的至少一些串代表的每个产生多个接近代表的部件，所述多个接近代表的每个根据该串中的至少一些相同字符而产生；以及用于存储条目在档案中的部件，每个条目表示基于它们各自的接近代表的在至少两个串之间的潜在近似匹配。

[0027] 在另一方面，一般，描述了用于管理用来确定与在记录中出现的串相关联的近似匹配的档案的系统。该系统包括：数据源，存储记录；计算机系统，被配置来处理该数据源中的记录以确定一组串代表，所述串代表对应于在记录中出现的串，为该组中的至少一些串代表的每个，产生多个接近代表，所述多个接近代表的每个根据该串中的至少一些相同字符而产生；以及与所述计算机系统耦接的数据存储器，存储包括条目的档案，每个条目表示基于它们各自的接近代表的在至少两个串之间的潜在近似匹配。

[0028] 各个方面能够包括以下优点的一个或多个。

[0029] 在典型的数据库应用中，不同记录的给定字段在它们的内容一致时匹配。如联合

和聚集之类的操作一般基于在特定字段中出现的匹配关键字将记录分到组中。可是，在一些应用中，能够使用近似串匹配来执行联合或聚集以比较关键字是有用的。如果在预定准则下两个记录对应的关键字字段足够接近，则它们被说成近似匹配。例如，当使用多于一个的数据源来执行该操作时，对于包括单词或短语的关键词，在每个源中的单词的准确拼写可能不同或者一个短语可能包含在另一中不存在的单词。

[0030] 维持档案以存储在一个或多个数据源的记录中出现的接近的串对。这些对和诸如由档案提供的分值的相关信息提高了联合、聚集和使用近似串匹配的其他操作的效率。在一些实施方式中，可以从计算图形的组件访问该档案，该计算图形对来自该数据源的数据执行操作，如下更详细描述的。

[0031] 通过以下描述以及权利要求书，本发明的其它特征和优点将变得明了。

附图说明

[0032] 图 1 是用于执行基于图形的计算的系统的框图。

[0033] 图 2 是计算图形的图。

[0034] 图 3 是预处理过程的流程图。

具体实施方式

[0035] 1 系统概述

[0036] 关于近似串匹配（或“模糊匹配”）的技术能够适用于各种类型的系统，包括存储数据集 (dataset) 的不同形式的数据库系统。如这里使用的，数据集包括数据的任何集合，其允许部分值据被组织为具有用于各个字段（也叫做“属性”或“列”）的值的记录。数据库系统和所存储的数据集能够采用多种形式中的任意一个，此类精密复杂的数据库管理系统或文件系统存储简单平实的文件。各种数据库系统的一个方面是记录结构的类型，其用于数据集（这能够包括用于每个记录中的字段的字段结构）中的记录。在一些系统中，数据集的记录结构可以简单地将各个文本文件定义为记录，并且文件的内容表示一个或多个字段的值。在一些系统中，不需要单个数据集中的全部记录具有相同的结构（如，字段结构）。

[0037] 利用与图形的顶点关联的计算组件以及在与图形的链接（弧、边缘）对应的组件之间的数据流，复杂的计算通常能够经过直接图形而被表示为数据流。在美国专利 5,966,072,EXECUTING COMPUTATIONS EXPRESSED ASGRAPHS 中描述了实现这样的基于图形的计算的系统，通过引用在此并入。用于执行基于图形的计算的一种办法是执行多个过程，每个过程与图形的不同顶点关联，并且根据图形的链接在各过程之间建立通信路径。例如，通信路径能够使用 TCP/IP 或 UNIX 域套接字，或使用共享存储器在各过程之间传递数据。

[0038] 参考图 1，用于执行基于图形的计算的系统 100 包括耦接到数据存储器 102 的开发环境 104 和耦接到数据存储器 102 的运行时间环境 108。开发者 101 使用开发环境 104 构建应用。应用与由数据存储器 102 中的数据结构所指定的一个或多个计算图形关联，所述数据结构可以是作为开发者使用开发环境 104 的结果而被写入到该数据存储器中的。关于计算图形 105 的数据结构 103 指定例如计算图形的顶点（组件或数据集）和在顶点之间的链接（表示工作元素的流动）。数据结构也能够包括组件的各种特性、数据集 (dataset) 和计算图形的流程（也叫做“数据流图形”）。

[0039] 运行时间环境 108 可以被托管在诸如 UNIX 操作系统的合适操作系统的控制下的一个或多个通用计算机中。例如，运行时间环境 108 能够包括多节点并行计算环境，其包括使用多个中央处理单元 (CPU) 的计算机系统的配置，所述多个中央处理单元可以是本地（如，诸如 SMP 计算机的多处理器系统）或本地分发的（如，耦接为集群或 MPP 的多处理器）、或者远程或远程分发的（如，经由 LAN 或 WAN 网络耦接的多处理器）、或它们的任意组合。

[0040] 运行时间环境 108 被配置来从数据存储器 102 和 / 或用户 107 中接收控制输入以执行和配置计算。控制输入能够包括用于使用在所存储的图形数据结构中指定的相应计算图形来处理特定数据集的命令。用户 107 能够例如使用命令行或图形接口来与运行时间环境 108 交互。

[0041] 运行时间环境 108 包括预执行模块 110 和执行模块 112。预执行模块 110 执行任何预处理过程，并且准备和维持用于执行计算图形的资源，诸如字典 111 和用于近似串匹配的档案 114。字典 111 存储单词和关于在数据集中出现的单词的相关信息。档案 114 存储来自基于单词、短语的预处理的各种结果或者数据集的记录。字典 111 和档案 114 能够以各种格式中的任意一种来实现并且能够被组织为数据的单个集合或作为多个字典和档案。执行模块 112 调度和控制对于分配给计算图形的过程的执行以执行各组件的计算。执行模块 112 能够与耦接到系统 100 的、诸如从数据库系统提供记录的数据源 116 的外部计算资源进行交互，在与图形组件关联的处理期间访问所述外部计算资源。

[0042] 参考图 2，计算图形 105 的简单示例包括执行聚集 (rollup) 操作的聚集组件 200，第一输入数据集 202、第二输入数据集 204 以及输出数据集 206。输入数据集聚集组件 200 向提供工作组件的流（如，数据库记录），以及输出数据集 206 接收由聚集组件 200 产生的工作组件的流（如，聚合数据库记录）。输入数据集 202 和 204 以及输出数据集 206 表示在由运行时间环境 108 可访问的存储介质（诸如数据源 116）中存储的数据（如，数据库文件）。聚集组件 200 比较从输入数据集接收的记录的关键字段值，并且基于在关键字段值之间的近似匹配而产生聚合记录。

[0043] 图 3 示出由预执行模块 110 执行的预处理过程 300，以准备由计算图形使用的字典 111 和档案 114。过程 300 从用户接收指示将处理哪些源和从那些源的哪个字段读取单词的配置信息 (302)。要读取哪个字段的指示可以根据默认设置而是暗示的，如读取全部字段。过程 300 通过读取所选择的源的记录的所选择字段，在字典中存储字段中出现的单词，并且更新诸如单词频率计数的统计来编译字典 111(304)。过程 300 通过对字典中的单词产生接近 (close) 单词 (308) 以及根据各个接近单词的比较而寻找潜在模糊匹配 (310) 来编译要存储在档案 114 中的潜在模糊匹配 (306)。与作为潜在模糊匹配的单词对一起存储分值，所述分值能够在图形执行期间被用来确定该潜在模糊匹配是否为实际模糊匹配。过程 300 通过基于存储在档案 114 中的结果计算关于字典中出现的单词的重要性分值而更新字典 111(312)。例如，对于字典 111 中的每个单词，过程 300 基于潜在模糊匹配来重新规格化 (renormalize) 单词频率计数 (314)，如下面详细描述的。然后，这些重新规格化的单词频率计数能够被用来计算可以在匹配短语和源的记录时使用的重要性分值。

[0044] 在一些实施方式中，预处理过程 300 在每次新的源可用时或在现有源中接收到新的记录时被重复。该过程能够由用户激活或者以重复的间隔或响应于一定事件而被自动激

活。

[0045] 2 模糊匹配

[0046] 许多商业面临的挑战是使用具有可能不完全相同的等价值的如“名称”或“地址”的字段来调和两个（或更多个）数据集。调和数据库可能涉及回答有关数据的各种问题，诸如下述。在一个数据集中的公司名称是否出现在另一个数据集中？如果是，它们是否具有相同的地址？如果来自两个数据集的公司名称完全相同，则能够使用联合操作来比较相应的地址字段，该操作找到具有匹配关键字（这里，是公司名称）的全部记录。但是如果名称不完全相同呢？一个名称可能以单词公司 (COMPANY) 结尾而另一个将其缩写为 CO 并且第三个不过将其完全地漏掉。类似 OF 或 THE 的冠词可以在一个名称中但是不在另一中。单词可能被拼错（如，COMPNY 代替 COMPANY)。在一个源中的名称可能包含附加信息，如联系人名称或账号。

[0047] 对于像企业名称的一些事情，不管在不同出处的数据集中，还是即便在单个数据集中也没有固定的格式规则。该挑战是找到匹配，一般为一组可能的匹配，即便名称是不一致的。能够使用近似串匹配、也被熟知为“模糊”匹配来执行这样的匹配。该匹配是模糊的，因为它们容忍错误或差异。

[0048] 当操作使用模糊匹配时，如果发现两个单词或短语在可接受的不同或差异的一定范围中等价，尽管不必相同，则说它们是匹配的。例如，单词的确切拼写可能不一致或者一个短语可以包含在另一中不出现的单词。能够使用分值来量化一致性的质量。当对应于记录的给定字段中的匹配的关键字不需相同但是为等价时，利用模糊匹配可以执行熟悉的操作，诸如比较、联合、聚集和查找。

[0049] 为了增加模糊匹配的速度，预执行模块 110 定期处理来自数据源的数据集，该数据源已经被识别为计算图形潜在地可访问的数据源。预执行模块 110 读取出现在数据集的记录的所选择字段中的数据。用户能够选择要处理的全部可用的字段或者可用字段的所选择子集。在一些情况中字段中的数据可以对应于单个单词，并且在一些情况中数据可以对应于包含多个单词的短语。（如这里所使用的，“单词”是包括来自某个字符集的一序列字符的任何串，而短语中的多个单词由空格或诸如逗号的其他分界符分离开）用于将字段分解为单词的规则是可配置的并且该规则能够在字段被初始选择时被指定。在一些情况中，尽管存在嵌入式空格（如，城市名称或英国邮政编码），但是可以通过将短语分解为“多个单词”，将嵌入式空格当成字符对待而以与单个单词相似的方式来处理具有多个单词的短语，如下更详细地描述的那样。这允许在模糊匹配中识别级联的和断开的单词。例如，“john allen”将匹配于“johnallen”或甚至是“johnal len”。

[0050] 预执行模块 110 识别记录中发生的一组单词并且在字典 111 中存储单词的代表（也叫做“单词代表”或“串代表”）。也可以在字典 111 中存储对于每个单词的统计。例如，可以在字典 111 中存储给定数据集或全部数据集的记录中的单词的发生频率。可以在字典 111 中存储关于单词的上下文的信息。例如，上下文能够包括该单词出现的字段或字段组。也能够在字典 111 中存储诸如在短语中它的位置的统计之类的其他关于单词的信息。单词的代表可以是单词自身，或者是以诸如不同字符集或非串代表（如，数字或字母数字编码）的不同形式的单词的编码代表，如下面更详细地描述的。

[0051] 预执行模块 110 的处理包括对给定单词产生一组“接近单词”（或“接近代表”）。

关于给定单词的接近单词与该给定单词相关联地存储在字典 111 中。接近单词被用来产生在单词对之间的潜在模糊匹配，如下面更详细地描述的。

[0052] 处理的结果被存储在档案 114 中以由通过执行模块 112 执行的图形使用。在确定是否应当处理给定记录的过程中（如，在联合或聚集组件中），一些图形可以使用档案 114。测量各记录的相似性是在众多上下文中出现的一种数据质量活动。例如，两个源可以共享公共关键字，该公共关键字被用在联合操作中以将来自两个源的记录拿到一起。需要比较其中可能出现关键字的记录中的字段。多个单词可能存在于一个字段中并且多个字段常被用于保存一定类型的数据，如名称或地址。在字段中的单词的布置在源中可以不一致，额外的单词可以存在在一个源或其他源中，单词可以是乱序的，并且可能存在排字错误。能够使用各种评分函数来计算表现记录之间的匹配的质量的特点的基本分值，然后由针对各种类型的不一致性和错误的惩罚来加权该基本分值。与不同类型的错误相关联的权重是可以调整的，并且每个惩罚对于关于任何比较的分值的贡献是值得报告的。各评分函数在它们如何比较作为模糊匹配的单词以及它们在计算分值中是否对单词进行加权（统计上）上不同。关于记录的分值可以涉及跨越几个单独评分的个体或级联的字段的加权分值。

[0053] 在匹配过程中存在 3 个评分级别：单词、短语和记录。一般在预处理阶段期间进行单词的评分，并且基于预定准则确定两个单词是否是潜在模糊匹配，以及将潜在模糊匹配与“模糊匹配分值”相关联。短语的评分可以不仅考虑模糊匹配的单词，而且可以考虑一个或两个短语中缺少单词的概率、单词乱序出现的概率或在它们之间存在额外单词的概率。评分算法可被配置为使得能够调整不相符的不同源的重要性。第 3 个评分级别是整个记录。通过以上下文敏感的方式组合不同字段的分值来对它们进行评分。缺失或不一致的信息能够给予适当的权重。

[0054] 档案 114 存储找到为潜在模糊匹配的单词对以及对应的模糊匹配分值。能够由用户修改潜在模糊匹配和它们的模糊匹配分值的集合以将所计算的模糊匹配分值替换为不同的模糊匹配分值并且增加与预定准则不相关的潜在模糊匹配单词对。

[0055] 为了短语比较的目的，档案 114 也存储“重要性分值”，其表示单词对于包含该单词的短语的相对重要性。重要性分值使用在数据集中单词出现的反向频率，但是使用（使用分值档案确定的）涉及模糊匹配的变化的频率来调整该值，并且可选地使用从单词的长度、在短语中的位置、源和上下文（如，单词出现的字段）导出的附加信息。例如，因为数据可能没有位于正确的字段中并且需要识别这样的错误，所以基于其中单词出现的上下文来调整该重要性分值可能是有用的。在一些情况下，出现在如地址字段的字段中的串可以以非结构化的格式被接收，但是不过其可以包含结构。这些串能够被解析并且识别单个元素。按上下文的相对重要性能够被用来指示给定的单词是更有可能出现在一个上下文还是另一个上下文中。例如，在地址字段，LONDON（伦敦）作为城市而非街道名称的一部分。前后关系的协调信息有助于解析：如果在 LONDON 后面有另一城市名称，则它可能是街道名称的一部分；如果它紧接在邮政编码的前面，则它可能是城市。

[0056] 在一些实施方式中，档案 114 存储测量两个短语之间的模糊匹配质量的“短语比较分值”。预执行模块 110 将各短语重叠以找到共享的和未使用的单词的列表，并且使用它们的相对重要性、校正 (alignment) 和单词顺序来计算该短语比较分值。

[0057] 存储在档案 114 的信息在它产生之后还能够以各种方式被进一步处理。例如，在

分值档案中，可能的误判(false positive)能够使用黄金参考数据的自我评分的结果来识别并且通过在初始评分期间应用的 n 克 (n-gram) 分析来排除。也可以包括其他分值。例如，来自多个字段的单词或短语能够被独立评分并且组合它们的分值以对整个记录评分。

[0058] 由使用模糊匹配的不同类型的操作访问档案 114。一些操作包括使用精确匹配的“正常版本”以及使用模糊匹配的“模糊版本”。聚集（或聚合）操作的模糊版本将相似的记录关联为群组。这对巩固在原始数据中存在标识唯一实体的关键字的变型形式的记录是有用的。简单定义与唯一实体关联的记录的群组可能是一个显著性结果。一旦已经定义群组，则支持一般的聚合活动。因为模糊匹配是评分的等价关系，所以也将对群组的关联进行评分。这具有这样的效果，与熟悉的情况相反，给定的实体将不必是唯一群组的成员。在群组的成员上的聚合（如数字总和）则将不准确但是仅将在反映群组成员的资格的不确定性的错误范围内知晓。可是，在全部群组上的聚合将仍然保持准确，因为整体上的成员资格是确定的。

[0059] 组件也能够使用该分值档案来识别各个单词中的错误拼写和其他错误。能够由用户（如，根据档案产生的列表）来确认所识别的错误，并且数据校正组件能够校正数据中的错误。基于模糊匹配的该纠错能力能够是对数据全面调节器 (profiler) 能力的扩展，如在标题为“Data Profiling”的美国专利申请号 10/941,402 中更详细描述的，这里通过引用并入。

[0060] 2.1 匹配准则

[0061] 各种准则中的任意一个能够被用于匹配质量的测量。考虑两个单词，MOORGATE 和 MOOGRATE。一种类型的准则是距离度量。存在多个不同的距离度量或用于测量两个单词之间的距离的办法。最简单的一个是汉明 (Hamming) 距离，其对相应字符是不同的位置的数目进行计数。由于该数目依靠于校准，所以使用汉明距离最小的校准。这对应于要求将一个单词转换为另一个的替换的最小数目，如以下示例所示。

[0062] M O O R G A T E

[0063] M O O G R A T E

[0064] * *

[0065] 在该示例中，存在相应字符不同的两个位置，要求两次替换来将一个单词转换为另一个。

[0066] 由于在计算汉明距离中校准很重要，所以“打乱”校准的插入可能产生较大的距离，如下示例。

[0067] M O O R G A T E

[0068] M O O R G R A T E

[0069] * * * *

[0070] 在该示例中，在原始单词中的单个插入产生汉明距离 4。对于一些应用，汉明距离夸大了插入的重要性。

[0071] 对汉明距离的替换是“编辑距离”。编辑距离是尽可能地保持校准而将一个单词转换为另一个所需的编辑（其中“编辑”是插入、删除或替换中的一个）的最小数目的计数。单词对 MOORGATE 和 MOOGRATE 具有编辑距离 2，因为其需要一次插入和一次删除（或两次替换）来将一个单词转换为另一个。单词对 MOORGATE 和 MOORGRATE 具有编辑距离 1，因为其

需要一次插入来将一个单词转换为另一个。

[0072] 存在用于计算编辑距离的各种算法。通过对插入、删除和替换（甚至互换）分配不同的权重，可以使得编辑距离分值反映不同类型的错误的重要性。编辑距离的复杂在于其通常在计算上比汉明距离更浩大。

[0073] “接近单词比较”是对上述编辑距离的替换，并且计算上迅速。由于在实际数据中排字和抄录错误的频率相对地较低，所以在单个单词中找到多于一个的错误较少见。（当然，诸如更多的将一个音节替换为语音上相似的一个的系统错误可以涉及多个字母，但是这能够被当成非匹配单词。）代替计算两个给定单词之间的编辑距离，预执行模块 110 使用“删除－联合”过程来实现接近单词比较以确定给定单词是否是潜在模糊匹配。删除－联合过程通过构造从每个单词中删除单个字符而获得的全部变型单词而开始，以获得关于每个单词的被叫做“删除变型”的一组接近单词。然后删除－联合过程比较两组删除变型以找出是否任意的删除变型是否匹配。如果它们匹配，则原始单词被指定为潜在模糊匹配。删除－联合过程的特征可以在于找到与有限组的等价变化相关的单词，该等价变化涉及单次删除、单次插入或伴随插入的删除（这覆盖替换和互换两者）。在以下示例中更详细地描述删除变型的产生和删除－联合过程。接近单词比较的其他形式是可以的。例如，每个单词能够具有根据不同组的等价变化而是“接近的”的一组接近单词。

[0074] 删除－联合过程也能够使用其中删除多个字符的删除变型，但是误判的数目一般随删除的字符的数目而上升。误判是当一个自然发生（“真实的”）单词匹配另一个自然发生单词的时候。例如，CLARKE 和 CLAIRE 是通过插入之后跟随的删除而相关的。的确，错误能够将一个单词转录为另一个，但是因为两个单词均是真实的，所以该可能性是每个单词无错误地自然发生。缺少有关哪个单词是自然发生的信息，则给定模糊匹配算法无法直接判断哪个单词是真实的（尽管存在推断它的一些可能性，如下所述的），从而对于该算法来说要检测误判是困难的。该情形是微妙的，因为一些改变可能是错误的而另一些则不是。用于应对误判的一个办法是给每个单词附加进一步的度量，叫做“重要性分值”，其最初基于如单词在数据中出现的频率的因素（单词频率计数或重新规格化的单词频率计数的反(inverse)）。例如，如果一个自然发生的单词容易与另外一个弄错，则二者的重要性降低，反映在它们的可区分性中的置信度的丧失。

[0075] 关于误判的潜在问题是其可能导致应该被区分的记录被报告成可能的匹配。在删除变型中使用单次删除的一个优点是，随着单词长度的增加（如超过大约 5 个字符），误判的数目迅速下降，因为一般在记录中存在更少的长单词，并且长单词一般被更普遍地分离。在一些实施方式中，用来产生删除变型的删除的数目能够取决于单词的长度。例如，对于具有不超过 5 个字符的单词，删除单个字符，而对于具有超过 5 个字符的单词，多达 2 个字符被删除（如，删除变型包括单个字符被删除的全部单词和两个单词被删除的全部单词）。

[0076] 预执行模块 110 能够包括减少大量评估的总体计算时间并且加速 ad hoc 评估的特征。出现在记录中的单词和它们的删除变型被存储在字典 111 中，并且删除－联合过程的删除比较的结果被预计算并存储在档案 114 中，其中它们是可用的而无需另外的评估。这对关于名称 / 地址数据的简单观察起到了杠杆作用：基于自然语言的数据中单词的数目相对于可能最终被处理的记录的数目是相当地小（如，＜＜ 1 百万）。并非每次其出现时重复相同的相当浩大的计算，档案 114 存储该结果（接近对和相关联的模糊匹配分值）以用于

重新使用。因为可能的单词的数目相对较小，并且仅观察的变型被存储，从而档案 114 的容量是可管理的。

[0077] 档案 114 具有另外的好处。如果在档案 114 中的短语的模糊匹配分值小于预定阈值（假定更低的分值指示更高质量的匹配），则用于短语的评分算法将报告单词之间的模糊匹配。通过使得用户能够手动调整档案 114 中的单词对的模糊匹配分值，能够避开不期望的匹配（如，误判）。另外，通过添加具有合适模糊匹配分值的对，在接近单词比较中不匹配的单词能够被添加为档案 114 中的匹配。例如，添加对 HSBC MIDLAND 创建了在“HSBC BANK”和“MIDLAND BANK”之间的模糊匹配。这提供了一种用于应用基于商业含义（这里，MIDLAND 是 HSBC 从前的名称）的同义词的方法。通过手动添加条目并且设置该模糊匹配分值以指示模糊匹配，多词关联（如对于 INTERNATIONAL BUSINESS MACHINES 的 IBM）也能够被添加为档案中的模糊匹配。在一些实施方式中，多词关联与单个单词之间的关联分开地被存储和 / 或处理。例如，在一些情况中，多词关联可以在标准化期间使用并且在短语比较评分期间不是必需的。

[0078] 对于单词、短语和记录的匹配过程一般包括将候选者识别为潜在模糊匹配并且判断在候选者之间的匹配质量以确定实际模糊匹配。

[0079] 匹配记录可以涉及匹配来自相应字段的单词，或在一些情况中，涉及匹配来自一个或多个字段的单词的短语。例如，由于当匹配名称或地址时来自不同源的记录中的存在的单词数目不必相同，所以能够通过“短语匹配”来执行在该上下文中的模糊匹配。预执行模块 110 能够选择将空格当成一个额外的字符，故短语就简单地变为较长的单词。当使用这种技术时，预执行模块 110 可以执行另外的处理以应对可选的单词，以考虑到在具有被当成字符对待的空格的较长单词中的区域会改变的事实。

[0080] 2.2 标准化和解析

[0081] 在模糊匹配的一些实施方式中，在比较之前短语被标准化。这按可预测的方式减少可变性，例如，通过漏掉通用单词，如 OF 或 THE，或者通过将通用缩写替换为完整的单词，诸如替换 CO 为 COMPANY。该标准化能够增加匹配的力度并且在一些办法中改进匹配过程的性能。困难在于一些信息可能在标准化期间丢失，或者会产生错误的识别 :CORP 很可能是 CORPORATION 的缩写，但是它可能为 CORPS 的错误拼写。

[0082] 标准化可被配置为使得在一些情况中标准化不被使用或者被维持在最小，并且其他机制被用来应对通用单词的问题和缩写的替换。在一些实施方式中，全部短语被标准化为大写字母或小写字母（同时保持如重音符号的字符变型），因为词形一般在不同源之间不一致并且大部分是假的区别。意图是用于保持原始数据的完整性并且为处理中的以后步骤对字段中的系统变型留出任何补偿（如缩写或同义词）。这样的原因是，虽然在地址字段中如 ST 的单词常意味者 STREET，但是这并非总是如此，从而太早地将它替换会是潜在的问题。

[0083] 能够指定对于标点字符的特别对待，因为这些在各源之间往往是不一致的。例如，许多标点字符或者是可选择的、或者指示在另外的未结构化的字段中的格式化。标点的处理类型的三种示例是 :替换为空的、替换为空格、以及替换为断线。其他形式的标准化能够包括将一些预定的单词或短语替换为其他单词或短语，诸如替换一个单词为同义词。用户能够使用在可配置的查找文件中的规则来控制标点和单词的替换。

[0084] 解析向来自任意数目的字段的不同单词或短语的部分分配含义。例如，解析可以在处理地址时使用，该地址在一些源中可以显露为结合到一个或两个字段而同时在其他源中它们被分散到 8 个或 10 个字段。能够使用参考源来促进解析期望包括已知元素的字段，该参考源提供用于识别和验证元素的辅助信息。例如，能够使用邮政地址文件 (PAF 文件) 作为识别和验证地址的各个元素的参考源来处理解析地址字段。

[0085] 2.3 单词频率和上下文

[0086] 预执行模块 110 针对出现在记录中的单词扫描给定源的记录，并且，在一些情况中，将该扫描限制到记录的所选择字段。在给定源的记录的所选择字段中出现的单词被存储在字典 111 中。在一些实施方式中，字典中的每个条目存储单词、单词的频率、单词的位置统计，以及单词的上下文。频率是单词出现在该源的记录的次数的计数 (如，单词可以在给定记录中出现多次)。频率能够是全部字段上的聚合计数或者是多次计数 (其中每个计数表示单词在给定字段中出现多少次)。频率也能够被重新规格化，如下面更详细地描述的。如果单词出现在给定字段的短语中，则对该短语计算该单词在该短语中的位置。在字典 111 中对于给定单词的位置统计包括，例如，在该单词出现的所有短语中该位置的平均和标准偏差。

[0087] 一个叫做“上下文”的类别词被存储在字典 111 中以支持字段的逻辑组合。用户能够在选择要处理的字段时指定上下文。上下文的使用使得可以比较具有不同记录结构的各个源，而不要求标准化到通用格式。这提高了各源之间的比较质量，因为能够从特定字段中的单词的存在中导出的源特定信息通过到通用格式的标准化而不会丢失。给定字段可以出现在多个上下文中，允许比较的粒度和数据的模糊布置的容差二者的控制。

[0088] 例如，“地址”上下文可以被用来对包含作为地址的一部分的单词的字段进行分组。与上下文关联的该字段名称在各个源之间可以不同。例如，在一个源中的字段 address_line1、address_line2、address_line3 可以都是地址上下文的一部分，而在另一源中地址上下文可以包括建筑物 _ 编号、街道、城市和邮编。单词频率计数和出现在不同字段的给定单词的重要性分值能够被聚集以表示相关上下文的频率。

[0089] 字典 111 被用于识别代表潜在模糊匹配的档案 114 中的单词并且识别由重要性分值表示的每个单词的重要性。能够使用可有效访问的多种文件格式的任意一种来维持该字典 111。例如，能够以索引压缩 (级联的) 平实文件格式来存储字典 111，诸如在美国申请序列号 11/555,458 中描述的格式，这里通过引用并入。或者为了不同的目的、源、字段编码或者为了例如加速访问，字典 111 能够保持在单个文件或多个文件中。

[0090] 关于单词的其他信息能够被包括在字典 111 中。例如，如果字段中的单词根据重要性按照降序顺序排序 :这将重要单词更早地置于短语中，则“重要性位置”是单词在字段中将具有的位置。例如，按照原始顺序的短语可以是“Bank of America”。具有按照重要性排序的单词的短语可以是“AmericaBank of”。单词“Bank”在原始短语中的位置是第 1 个 (3 个之中)。单词“Bank”的重要性位置是第 2 个 (3 个之中)。

[0091] 2.4 模糊匹配分值

[0092] 在源中和各源之间的单词的潜在模糊匹配是预计算的并且与描述匹配的质量的特点的关于在单词对之间的每个潜在匹配的模糊匹配分值一起被存储在档案 114 中。由于源中独特的单词的数目一般比源中全部单词的数目少很多，所以该预计算步骤通过消除单

词的冗余模糊比较来加速稍后的比较和字段的评分。最初，仅构成潜在模糊匹配（根据诸如接近单词比较技术的预定准则）的单词被存储在档案 114 中。用户能够基于预定准则通过手动调整模糊匹配分值、或添加档案 114 中关于单词的模糊匹配分值的最初群体中未识别的匹配对来修改和扩展档案 114。

[0093] 在一些实施方式中，档案 114 使用删除联合过程来增加模糊匹配分值。并非计算每个单词对之间的全部编辑距离，这将在计算上很浩大，而是在删除联合过程中仅附近的单词被比较。这是按以下方式实现的。对单词字典 111 中的每个单词（或对于字典 111 的一部分，如对于给定源、上下文和 / 或字段），获得通过删除单个字符形成的每个变型。关于给定原始单词的“删除集”包含条目的列表，每个条目具有关于原始单词的关键字（“word_key”）、原始单词（“original”）、删除变型（“delection_var”）、和已经从原始单词中删除的字符的位置（“deletion_pos”）。该删除集能够与原始单词一起存储在字典 111 中，或者可以在由预执行模块 110 使用以产生被存储在档案 114 中的潜在模糊匹配之后被丢弃。原始单词和删除变型一起被包括在该删除集中并且具有删除字符位置为 0。例如，以下是对于单词 LONDON 的删除集：

[0094]

word_key	deletion_pos	deletion_var	original
[0095] 1	0	LONDON	LONDON
[0096] 1	1	ONDON	LONDON
[0097] 1	2	LNDON	LONDON
[0098] 1	3	LODON	LONDON
[0099] 1	4	LONDN	LONDON
[0100] 1	5	LONDO	LONDON

[0101] 注意，word_key、deletion_pos 是识别给定删除变型的唯一“关键字”。

[0102] 该删除联合过程能够被扩展到更多删除，记录删除位置的序列，但是，由于多于一个的删除序列可能导向相同的单词，以致对于给定删除变型的“关键字”不再是唯一的。（可是存在通过要求按照特定顺序完成删除而确定的规范的关键字，如从原始单词的左边开始的顺序，并且总是指示根据先前变型的删除位置。）因此，从原始单词 LONDON 中删除两个字符（两次均在第三位置）产生的删除变型 LOON 将具有删除集条目：

[0103]

word_key	deletion_pos	deletion_var	original
[0104] 1	3,3	LOON	LONDON

[0105] 删除联合过程通过对删除 _ 变型单词执行联合操作来从一个或多个字典中的单词中确定潜在模糊匹配。通过比较所删除的字符的位置来对模糊匹配的质量进行评分。在用于计算模糊匹配的过程的一个示例中，分值点被分配给如下的不同类型的改变：每次删除为 1 个点、改变第一个字母为 1 个点、改变最后一个字母为 1 个点、如果所删除的字符被多于一个的位置分离为 1 个点。与每个类型的改变关联的权重是可调整的。如果一个单词的删除位置是 0 而其他不是 0，则其为单次插入或删除。如果删除位置相同，则其为代替。具有相同 word_key 和 deletion_pos 的匹配被忽视，因为这些是确切的匹配。指示成对字母的删除的匹配也被忽视，因为其不提供信息（如，通过删除字符 2 或 3 得到 MEET- ＞ MET）。

[0106] 以下是对于原始单词 LONDON、LODON、LOMDON 和 LODNON 的来自各个删除集的一系列所选择条目的示例。

[0107] word_key deletion_pos deletion_var original

[0108] 1 0 LONDON LONDON

[0109] 1 3 LODON LONDON

[0110] 1 4 LONON LONDON

[0111] 2 0 LODON LODON

[0112] 3 0 LOMDON LOMDON

[0113] 3 3 LODON LOMDON

[0114] 4 0 LODNON LODNON

[0115] 4 3 LONON LODNON

[0116] 4 4 LODON LODNON

[0117] 在该示例中，一些删除变型条目已经被抑制，因为它们不会导致有意思的匹配。联合操作将第一条目与具有 deletion_var 的相同值的第二条目配对。结果得到的在原始单词对之间的潜在模糊匹配是：

[0118] 第一条目 第二条目 潜在模糊匹配

[0119] 1 3 LODON LONDON 2 0 LODON LODON LONDON LODON

[0120] 1 3 LODON LONDON 3 3 LODON LOMDON LONDON LOMDON

[0121] 1 3 LODON LONDON 4 4 LODON LODNON LONDON LODNON

[0122] 1 4 LONON LONDON 4 3 LONON LODNON LONDON LODNON

[0123] 2 0 LODON LODON 3 3 LODON LOMDON LODON LOMDON

[0124] 分别地，以上示范的潜在模糊匹配表示单词 0- 删除、替换、互换、通过不同路径获得的互换和单词 0- 插入（单词 1- 删除）。代表潜在模糊匹配的档案 114 中的每个单词对具有指示匹配的质量的相关的模糊匹配分值。

[0125] 使用上述的过程，对于这些对的模糊匹配分值如下：

[0126] 潜在模糊匹配 模糊匹配分值

[0127] LONDON LODON 1

[0128] LONDON LOMDON 1

[0129] LONDON LODNON 2

[0130] LONDON LODNON 2

[0131] LODON LOMDON 1

[0132] 作为将被识别为潜在模糊匹配的对进行评分的另一示例，单词 ONDON 和 LONDON 将具有模糊匹配分值为 4(1 对于第一个字母、1 对于删除 L、1 对于删除 0，1 对于非相邻删除)。

[0133] 能够对具有来自包括单个源的任意数目的源的单词的字典 111 执行联合操作。在一些情况中，字典 111 包括具有来自第一源 (source0) 的单词的部分和具有来自第二源 (source1) 的单词的部分。能够使用所选择的源执行删除 - 联合过程并且结果能够被存储在档案 114 中，一起存储在档案 114 中的还有哪个源被选择的指示和诸如它们按照什么顺序被比较的其他信息。

[0134] 例如，代表潜在模糊匹配的档案 114 中的每个单词对也能够与 entry_descriptor（条目描述符）相关联。entry_descriptor 能够被实现为比特映射字段，其指示该对的来源。一个比特指示该对是通过使用该删除联合方法的图形产生的。其他比特指

示该分值是否已经被用户修改、该对是否已经被用户插入（如引入同义词）或者该对是否是通过清除嵌入的标点字符（叫做“清除的”）而产生的。

[0135] 标点字符的清除（潜在地包括外语的发音符号）起因于因为经验表明多个标点字符能够发生嵌入在单个单词中。有时这些是有效可选择的，如姓名首字母的句点；有时它们看起来像拖尾的无用信息字符。由于往往涉及多于一个的字符，所以这样的清除不会被单次删除方法获得而多次删除将带来太多的误判。以下是清除后单词的一对示例。

[0136] 原始单词 清除后单词

[0137] S.I.ID. SIID

[0138] HOLDINGS### HOLDINGS

[0139] 来自单个源的单词能够彼此相对来评分。当存在多于一个源时，来自每个源的单词能够相对其他源（如，独立的字段或上下文）的那些单词而被评分。这识别可能不在自身源内出现的潜在模糊匹配。由于拼写错误和排印错误是相对隔离的事件，故它们一般随各个源而改变。

[0140] 档案 114 能够存储遍及所有源的潜在模糊匹配结果，例如以索引压缩（级联的）平实文件格式来存储，诸如在美国申请序列号 11/555,458 中描述的格式，这里通过引用并入。以多文件格式并用于不同的目的的多个档案 114 的使用是可能的（如，用于下述的多词分值）。当新源被引入时，针对添加到字典 111 的它们自己的删除集并针对已经在字典 111 的现有删除集来对单词进行评分。新的潜在模糊匹配对和分值能够按照排序顺序级联到现有档案 114 的末尾。稍后的融合处理能够为了性能而重新组织该档案。

[0141] 在一些实施方式中，如果关于单词对的档案 114 的条目已经存在，则该对被丢弃并且不再评分。例如，这样做是为了性能并且允许用户修改通过删除联合过程而产生的分值。由于档案 114 是累积的，因为产生了大量所观察到的对，所以需要被评分的单词数目随时间下降。

[0142] 修改条目的分值对于个别地避开误判来说是有用。计算图形的组件能够使用档案 114 中的分值来基于分值档案中它们的分值是否低于给定阈值来确定单词对是否是模糊匹配。将档案 114 中的给定单词对的分值提高到阈值以上有效地避开了匹配（指示所识别的潜在模糊匹配不是实际模糊匹配）。单词对的明智的评分使得可以通过调整阈值来选择性地开启和避开一组单词。在一些实施方式中，一个或多个阈值被配置为取决于上下文，并且给定对的分值也能够取决于上下文（如，使用存储在档案 114 中的附加信息）。

[0143] 用户也能够添加不会在删除 - 联合过程中识别的单词对到档案 114。这对于添加同义词或缩写词到分值档案是有用的。例如，STREET 和 ST 不会被删除 - 联合过程识别为潜在模糊匹配，但是这可能是期望的等同。也可以添加绰号识别，如 ROBERT 和 BOB（这是取决于上下文分值的自然示例——这在个人名称上下文中可以当成匹配但是在其他上下文则不然）。

[0144] 在档案中的模糊匹配分值能够通过来自进一步处理的结果（如，在那些单词出现的短语之间的已接受的匹配）的反馈而更新。

[0145] 2.5 单词频率重新规格化和重要性分值

[0146] 在档案 114 增加了至少一些单词对之后，其能够被用来“重新规格化”在字典 111 中的单词频率计数。通过添加作为潜在模糊匹配而涉及该单词的全部单词的计数来调整每

个单词的频率。结果得到的重新规格化频率被用来计算单词的“重要性分值”，其进而将在匹配短语时被使用。在数据中单词频率越小，则其在其与其他单词可区分的方面更加重要。

[0147] 通过错误拼写的单词来说明将频率重要性的概念应用到原始单词频率计数的困难。拼写错误较少见因此实在是太为重要。通过利用与之匹配的更频繁发生的单词的计数来调整它们的计数，显示出它们真实的相对重要性。高频率匹配单词不该必须地被认作是“正确拼写的”，因为这暗示可能不应用的对匹配单词的校正。并非全部低频率单词都被错误拼写以及并非全部匹配的高频率单词都是正确拼写的甚至是错误拼写的单词。例如，NORTE 可能是 NORTH 的错误拼写或它可能只是西班牙语的 North。LABLE 可能是 LABEL 的错误拼写但是它也可能是 TABLE 的错误拼写，二者将作为高频率匹配发生。

[0148] 重要性携带可区分的强烈涵义。如果单词与多个高频率单词匹配，如 LABLE 这样，则它被看成不那么重要的，因为存在对于它被错误地当成另一单词的较大的范围。

[0149] 在档案 114 中，重新规格化的单词频率计数能够被存储，并且诸如单词列表的其他信息用来执行该重新规格化（如，为了诊断的目的）。以下是作为单词 AVENUE 的潜在模糊匹配的单词的示例，以及示出单词出现的上下文的示范短语及单词出现在示范数据源的次数的计数。在该示范数据源中，单词 AVENUE 自身出现 10,500 次。

	单词	上下文	计数
[0150]			
[0151]	AVENE	237 Park Avene	1
[0152]	AVNENUE	255 5th Avnenue SW	1
[0153]	AVENUNE	306 MAROON AVENUNE	2
[0154]	AVEUNUE	236 TATE AVEUNUE	1
[0155]	AVENUES	Philadelphia & Reading Avenues	11
[0156]	AVENUUE	57 NORTH TWELFTH AVENUUE	1
[0157]	AVENUS	1900 9th Avenus	1
[0158]	AVEUNE	1010 BELLEVERNON AVEUNE	4
[0159]	AVENIE	3401 Hillview Avenie	1
[0160]	AVNUE	540 GODDARD AVNUE	1
[0161]	AVEBUE	10319 FIRMONA AVEBUE	1
[0162]	AVENYE	132 JEFFERSON AVENYE	1
[0163]	VENUE	725 North Mathilda Venue	1
[0164]	ANENUE	3330 Evergreen Anenue	1
[0165]	VENDUE	ZI LA VENDUE	1
[0166]	AVENUE	5200 NW 33rd Avenue.Suite 215	1

[0167] 单词 AVENUE 将示出为具有 16 个潜在模糊匹配并且将通过那些匹配单词的计数的总和来调整 10500 的计数。这些匹配单词的每个使得它们的计数由通过与 AVENUE 相关联的 10500 加上对于匹配单词被识别为潜在模糊匹配的任意其他单词来调整。一般，错误拼写的单词比正确拼写的单词链接有更少的模糊匹配。

[0168] 在以下示例中说明通过其来重新规格化单词的频率计数的过程。在该示例中，关于 source0 的字典包括出现在被叫做“legal_address（合法地址）”的字段中的单词 MÉXICO（原始频率计数 11）和 MEXICO（原始频率计数 259），而关于 source1 的字典包

括出现在被叫做“taddress3”的字段中的单词 MEXCIO(原始频率计数 2) 和 MEXICO(原始频率计数 311)(注意 source0 中加重音符的 E)。

[0169] 在该示例中,基于在关于 source0 的字典的删除集和关于 source1 的字典的删除集之间的联合操作,存储在档案 114 中的单词对包括 source0 和 source1 之间的潜在模糊匹配。(以下示例将利用来自每个源的删除集的自联合的潜在模糊匹配来增加档案。)因此,以下两个潜在模糊匹配对出现在档案 114 中 :

[0170] MÉXICO MEXICO

[0171] MEXICO MEXCIO

[0172] 重新规格化过程的示例如下。处理关于 source0 和 source1 二者的字典,例如,开始于 source0 中的单词MÉXICO。在档案 114 中查找该单词以寻找在 source1 中出现的潜在模糊匹配的列表。然后,在关于 source0 的原始字典中查找每个潜在模糊匹配。结果得到的计数被添加到原始计数。应用于上述示例的过程产生关于在 source0 的字典中的单词频率计数的重新规格化的以下结果 :

[0173] 来自 source0 首次输入单词和原始计数 :MÉXICO 11 在档案中查找,返回 {MEXICO} 在 source0 字典中查找每个,添加计数 :

[0174] source0 :MEXICO 259

[0175] 找到 :{MEXICO}

[0176] 对于MÉXICO的重新规格化计数 = 11+259 = 270 来自 source0 的第二次输入单词和原始计数 :MEXICO 259 在档案中查找,返回 { MÉXICO, MEXCIO} 在 source0 字典中查找每个,添加计数 :

[0177] source0 :MÉXICO 11

[0178] source0 :MEXCIO 未找到

[0179] 找到 :{MÉXICO}

[0180] 对于 MEXICO 的重新规格化计数 = 259+11 = 270

[0181] 假定对于具有原始单词频率计数 5 的单词MÉXICA在 source0 字典中存在附加条目。

[0182] 该单词MÉXICA在 source1 中不具有对于任何单词的潜在模糊匹配,故它不出现在档案 114 中,因此将不参与对于 source0(相对于 source1) 的重新规格化。可是,如果已经利用 source0 字典的删除集上的自身联合来扩展档案 114 为具有该附加条目,则在档案 114 中将有以下附加条目 :

[0183]

MÉXICA MÉXICO

[0184] 然后,对第一单词MÉXICO的查找将添加MÉXICA到该组所找到的潜在模糊匹配中。对于MÉXICO的重新规格化则将执行如下 :

[0185] source0 :MEXICO 259

[0186] source0 :MÉXICA 5

[0187] 找到 :{ MÉXICA, MEXICO}

[0188] 对于MÉXICO的重新规格化计数 = 11+5+259 = 275 现在,重新规格化的单词频

率更高，反映出附加潜在模糊匹配MÉXICA的存在。然后计算基于重新规格化的单词频率计数计算的重要性分值，例如对用非空记录的总数除以重新规格化的单词频率计数求对数。在重要性分值的这个版本中，单词及其变型出现越频繁，则重要性分值越低。负值指示单词出现常多于每个记录一次。

[0189] 重新规格化的单词频率计数能够被用来识别可能的错误拼写或相反地可能的误判。简言之，期望错误拼写预计是少见的而误判则不然。简单的比率测试指示哪些单词具有比重新规格化的单词频率计数少得多的计数。这些很可能是错误拼写。如果参考 ngram 频率则可以取得甚至更高的置信度。ngram 是 n 字母单词片段。遍及全部数据的 ngram 的频率分布指示不同字母组合的发生频率。（该分布当然是语言特定的。）该想法在于在档案的产生期间，在两个单词之间的改变的位置是已知的。横越 (span) 改变的位置的两字母和三字母（以及更多的）单词片段能够被识别并查找它们的频率。如果在一个单词中与该改变相关联的 ngram 频率与在另一个单词中的 ngram 频率相比低很多，则指示前一单词很可能被错误拼写。

[0190] 另一方面，其每个具有相对较高计数的多个变型很可能是自然发生的变型——也即，误判。从评分的观点看，误判匹配的存在由于对其添加了相对更大的计数而降低了单词的重要性。在一些情况中，因为其指示该单词可能被弄错，所以这会是理想的。但是在另外的情况中，各单词样子如此不同，以致不可能是错误产生的它们 --- 当然不是在相对计数的级别。

[0191] 重要性可以完全是相对的。当涉及多个源时，一些潜在模糊匹配单词可能不会在全部源中出现。这意味着调整后的计数能够在各源之间变化。类似地，与单词相关联的字段或上下文可以是相关的。例如，做出的该调整能够适用于源、字段和上下文的解析。在相对的情况中，只有当在比较两个源时在合适的源 / 字段 / 上下文中实际出现的潜在模糊匹配才将被用于调整该计数。期望来自源特定变型的贡献一般是较小的影响。

[0192] 2.6 编码

[0193] 在删除 - 联合过程中找到的接近单词基于本质上未从它们在原始数据集中的样子中改变的单词中的字符的排列。接近单词比较也能够在通过使用“单词编码”改变的“单词的间隔”中进行。当使用单词编码时所找到的接近集合可以不同。单词编码将单词映射到新的表示。映射可以是一对一、一对多或多对一。一些编码可以将单词变换为不同的字符集，而一些编码可以将单词变换为数字表示。单词编码修改单词的间隔从而根据给定度量的单词之间的距离可能变化。在应用单词编码之后，就它们的自然字符表示来说并不接近的单词可以是接近的。

[0194] 例如，在一些实施方式中，预执行模块 110 执行“素数编码”，其中字符集中的每个字符被编码为素数（如，在字母表的每个字母被映射到不同的素数，不管大小写），并且编码的单词是字符的素数的乘积。因为乘法是累积的（即独立于因子的阶），所以只要它们包括相同的字符集，而不管它们的顺序如何，则两个编码的单词是相同的。该编码对互换不敏感或者真正地对加扰不敏感，并且是多对一映射的示例。

[0195] 对于给定编码，删除 - 联合过程的变型在编码以产生接近单词之前能够执行字符的删除，或者在编码以产生接近单词之后能够执行接近单词操作。能够执行用于素数编码的删除 - 联合过程的变型，其中模块 110 将编码乘积除以素数以删除删除变型中的对应字

符。对于一些编码（如素数编码），编码之后的接近单词操作产生如同在编码之前已进行了字符删除的结果，但是对于其他编码，如果在编码之前已进行了字符删除，则接近单词操作可以产生不同的结果。

[0196] 对于使用多个字母表或字符集（如，识别字符的计算机字节代码）的一些语言，如日语，编码可以包括在编码之间标准化字母表或字符集的选择。

[0197] 2.7 多词

[0198] 多词 (multiword) 是被当成单词的、包含嵌入空格的短语。在前一示例中，在评分之前短语已经被解析为没有嵌入空格的单词。这忽略了两个潜在的错误源：空格能够被插入单词中而单词之间的空格能够被丢弃。另一示例是将与短语相关的同义词处理为单个单词，如首字母缩写词。

[0199] 允许嵌入空格以削弱作为分隔符的空格的识别。这通过将短语的解析扩展为不仅包含单个单词，而且包括全部相邻的一对词和三个词等来实现。短语被分解为比特定长度短的它的全部子短语（多词）。多词 (mword) 中的全部嵌入空格被删除以形成级联单词 (cword)。这类似于通过删除形成的单词。cword 变为多词字典和多词档案的关键字。当比较多词时，匹配它们的 cword 然后给原始 mword 评分。目前，忽视在 mword 中错误拼写的单词的概率。为对待这种情况，当对 mword 对进行评分时参考该档案。

[0200] 作为示例，考虑具有以下 3 个名称的源

[0201] JOHN A SMITH

[0202] JO HNA SMITH

[0203] JOHNA SMITH

[0204] 将第一条目 mword 分解为长度 3 将给出一组 mword：{JOHN，A，SMITH，JOHN A，A SMITH，JOHN A SMITH}。

[0205] 3 实现方式

[0206] 能够使用用于在计算机上执行的软件来实现这里描述的近似串匹配方法。例如，该软件形成一个或多个计算机程序中的进程，所述一个或多个计算机程序在一个或多个编程的或可编程的计算机系统（其可以具有各种架构，诸如分布式、客户端 / 服务器或网格）上执行，每个系统包括至少一个处理器、至少一个数据存储系统（包括易失性和非易失性存储器和 / 或存储元件）、至少一个输入设备或端口以及至少一个输出设备或端口。软件可以形成更大程序的一个或多个模块，例如，提供与计算图形的设计和配置有关的其他服务。图形的节点和元素能够被实现为存储在计算机可读介质中的数据结构或遵循存储在数据仓库中的数据模型的其他组织数据。

[0207] 软件可以被提供在存储介质上，诸如 CD-ROM，可通过通用或专用可编程计算机读取或者经过网络的通信介质传递到（以传播信号编码）在其中被执行的计算机上。全部功能可以在专用计算机上执行，或者使用专用硬件（诸如协处理器）来执行。可以以分布式方式实现软件，其中由软件指定的计算的不同部分由不同的计算机执行。每个这样的计算机程序优选地被存储在或下载到通用或专用可编程计算机可读的存储媒体或设备（如，固态存储器或媒体，或磁或光媒体），用于当存储媒体或设备被计算机系统读取以执行这里所述的程序时配置和操作计算机。发明性系统也可以考虑实被现为计算机可读存储介质，配置有计算机程序，其中存储介质被配置来使得计算机系统以特定的和预定的方式操作以执

行这里所述的功能。

[0208] 已经描述了本发明的多个实施例。然而，将理解可以做出各种修改而不脱离本发明的精神和范围。例如，上述的一些步骤可以是顺序独立的，因此能够按照不同于所述的顺序来执行。

[0209] 应该理解，前述说明意欲是说明性的而非限制本发明的范围，该范围由所附权利要求的范围来限定。例如，可以以不同的顺序来执行上述的多个功能步骤，而不会在本质上影响整个处理。其他实施例在以下权利要求的范围中。

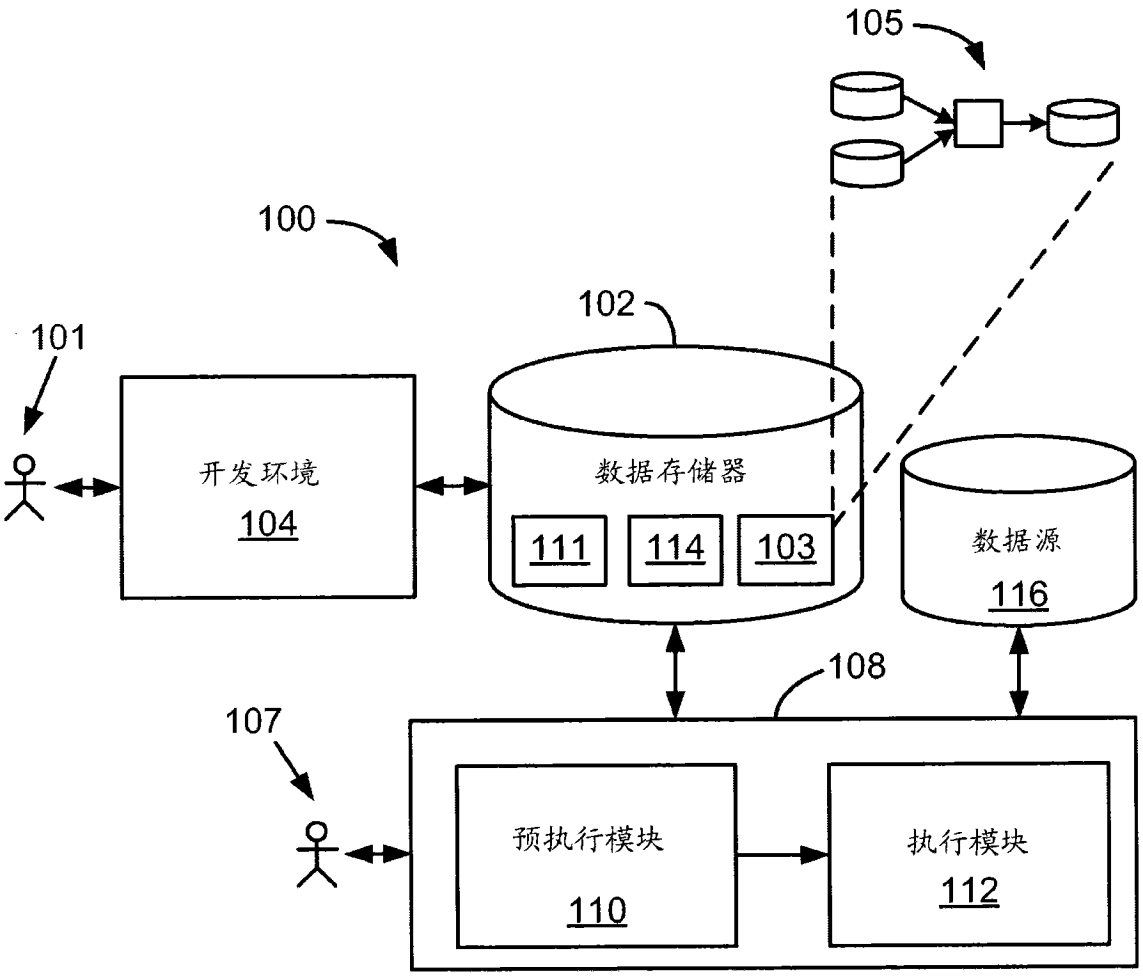

105
100
102
101
开发环境
104
数据存储器
111
114
103
数据源
116
108
107
预执行模块
110
执行模块
112

图 1

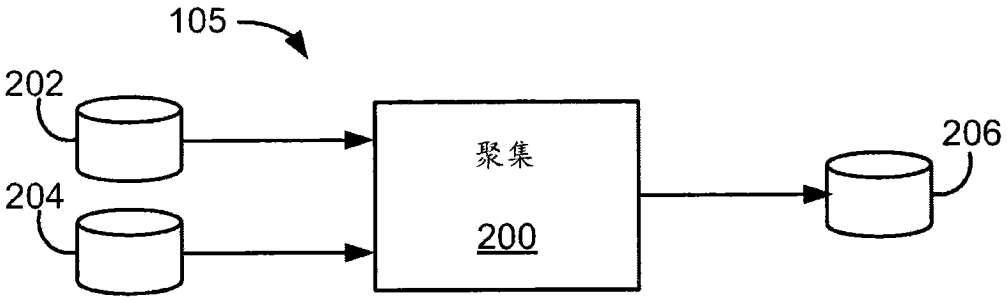

105
202
204
聚集
200
206

图 2

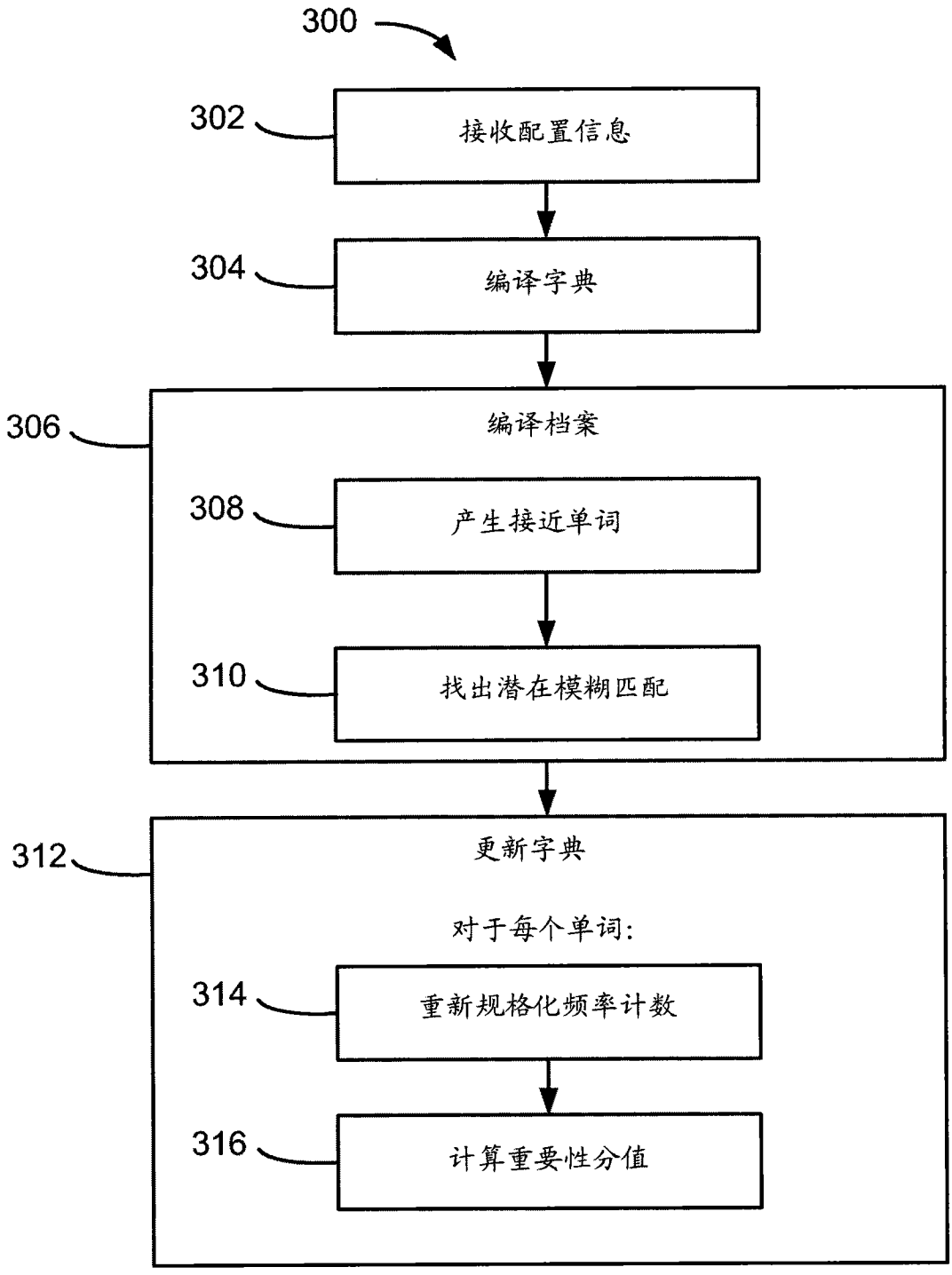

300
302
接收配置信息
304
编译字典
306
编译档案
308
产生接近单词
310
找出潜在模糊匹配
312
更新字典
对于每个单词:
314
重新规格化频率计数
316
计算重要性分值

图 3