



(12) 发明专利

(10) 授权公告号 CN 101404015 B

(45) 授权公告日 2012. 03. 21

(21) 申请号 200810166177. 3

(22) 申请日 2008. 10. 06

(30) 优先权数据

60/977, 840 2007. 10. 05 US

12/242, 965 2008. 10. 01 US

(73) 专利权人 富士通株式会社

地址 日本神奈川县川崎市

(72) 发明人 大卫·马尔维特 贾瓦哈拉·贾殷

斯特吉奥斯·斯特吉奥

亚历克斯·吉尔曼

B·托马斯·阿德勒

(74) 专利代理机构 北京三友知识产权代理有限公司

公司 11127

代理人 李辉

(51) Int. Cl.

G06F 17/30 (2006. 01)

审查员 鞠博

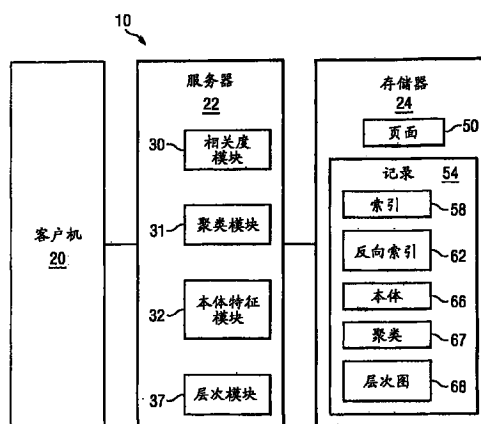
权利要求书 3 页 说明书 16 页 附图 4 页

(54) 发明名称

自动生成词条层次

(57) 摘要

本发明涉及自动生成词条层次。在某些实施方式中,生成词条的层次包括访问包括多个词条的语料库。对一个或多个词条执行以下步骤以产生父子关系:根据有向相关度来标识词条的一个或多个父词条;以及根据该父词条和每个词条来建立一个或多个父子关系。根据这些父子关系而自动生成层次图。



1. 一种用于自动生成词条层次的方法,可应用于搜索数据语料库,该方法包括以下步骤:

访问存储在一个或多个有形介质中的语料库,该语料库包括多个词条;

对所述多个词条中的每个词条执行下述步骤以产生多个父子关系:

根据有向相关度来标识这多个词条中的一个或多个父词条,所述多个词条包括这一个或多个父词条,所述有向相关度是包括两个词条的共现上下文的数量与包括一个词条的共现上下文的数量之比;以及

根据这一个或多个父词条和这每个词条来建立一个或多个父子关系,其中父词条和与其具有父子关系的词条相比在涵义上更加一般化;以及

根据这多个父子关系来自动生成层次图。

2. 根据权利要求1所述的方法,其中,根据有向相关度来标识这多个词条中的一个或多个父词条的步骤还包括以下步骤:

确定从每个词条到候选父词条存在高有向相关度;

确定从该候选父词条到这每个词条存在低非零有向相关度;以及

将该候选父词条标识为这每个词条的父词条。

3. 根据权利要求1所述的方法,其中,根据有向相关度来标识这多个词条中的一个或多个父词条的步骤还包括以下步骤:

确定从每个词条到候选父词条的有向相关度大于0.6;

确定从该候选父词条到这每个词条的有向相关度小于0.4且大于0.1;以及

将该候选父词条标识为这每个词条的父词条。

4. 根据权利要求1所述的方法,其中根据这多个父子关系来自动生成层次图的步骤还包括通过以下步骤来缩减层次图:

标识该层次图的父子关系和冗余父子关系;以及

从该层次图中移除该冗余父子关系。

5. 根据权利要求1所述的方法,其中

该语料库包括多个文档,这多个文档包括所述多个词条;并且
该层次图表示了这多个文档的类别。

6. 根据权利要求1所述的方法,其中

该语料库包括多个搜索结果;

所述多个词条表示了搜索结果的话题;并且

该层次图描述了这多个搜索结果的多个父子关系。

7. 根据权利要求1所述的方法,其中

该语料库包括被组织为多个聚类的多个搜索结果;

所述多个词条表示了聚类的话题;并且

该层次图描述了这多个搜索结果的多个父子关系。

8. 根据权利要求1所述的方法,该方法还包括以下步骤:

将该层次图的多个父词条发送给客户机;

接收父词条的选择,所选父词条具有多个子词条;以及

将这多个子词条发送给该客户机。

9. 根据权利要求 1 所述的方法,该方法还包括以下步骤:
接收包括该层次图的父词条的搜索查询;
标识该父词条的一个或多个子词条;以及
使用该父词条和这一个或多个子词条来搜索该语料库。
10. 根据权利要求 1 所述的方法,其中
该语料库包括多个文档,这多个文档包括所述多个词条;并且
该方法还包括以下步骤:
将该层次图中的每个词条与包括这每个词条的文档相关联;
接收包括父词条的搜索查询;以及
接收与该父词条的子词条相关联的一个或多个文档。
11. 一种用于自动生成词条层次的系统,可应用于搜索数据语料库,该系统包括:
用于访问存储在一个或多个有形介质中的语料库的装置,该语料库包括多个词条;
用于对所述多个词条中的每个词条执行下述步骤以产生多个父子关系的装置:
根据有向相关度来标识这多个词条中的一个或多个父词条,所述多个词条包括这一个或多个父词条,所述有向相关度是包括两个词条的共现上下文的数量与包括一个词条的共现上下文的数量之比;以及
根据这一个或多个父词条和这每个词条来建立一个或多个父子关系,其中父词条和与其具有父子关系的词条相比在涵义上更加一般化;以及
用于根据这多个父子关系来自动生成层次图的装置。
12. 根据权利要求 11 所述的系统,还包括用于通过下述步骤根据有向相关度来标识这多个词条的一个或多个父词条的装置:
确定从每个词条到候选父词条存在高有向相关度;
确定从该候选父词条到这每个词条存在低非零有向相关度;
确定从该候选父词条到这每个词条的有向相关度高于最小阈值;以及
将该候选父词条标识为这每个词条的父词条。
13. 根据权利要求 11 所述的系统,还包括用于通过以下步骤根据有向相关度来标识这多个词条的一个或多个父词条的装置:
确定从每个词条到候选父词条的有向相关度大于 0.6;
确定从该候选父词条到这每个词条的有向相关度小于 0.4 且大于 0.1;以及
将该候选父词条标识为这每个词条的父词条。
14. 根据权利要求 11 所述的系统,根据这多个父子关系来自动生成层次图的装置还包括通过以下步骤来缩减层次图的装置:
标识该层次图的父子关系和冗余父子关系;以及
从该层次图中移除该冗余父子关系。
15. 根据权利要求 11 所述的系统,其中
该语料库包括多个文档,这多个文档包括所述多个词条;并且
该层次图表示了这多个文档的类别。
16. 根据权利要求 11 所述的系统,其中
该语料库包括多个搜索结果;

所述多个词条表示了搜索结果的话题 ; 并且
该层次图描述了这多个搜索结果的多个父子关系。

17. 根据权利要求 11 所述的系统, 其中
该语料库包括被组织为多个聚类的多个搜索结果 ;
所述多个词条表示了聚类的话题 ; 并且
该层次图描述了这多个搜索结果的多个父子关系。

18. 根据权利要求 11 所述的系统, 还包括 :
用于将该层次图的多个父词条发送给客户机的装置 ;
用于接收父词条的选择的装置, 所选父词条具有多个子词条 ; 以及
用于将这多个子词条发送给该客户机的装置。

19. 根据权利要求 11 所述的系统, 还包括 :
用于接收包括该层次图的父词条的搜索查询的装置 ;
用于标识该父词条的一个或多个子词条的装置 ; 以及
用于使用该父词条和这一个或多个子词条来搜索该语料库的装置。

20. 根据权利要求 11 所述的系统, 其中
该语料库包括多个文档, 这多个文档包括所述多个词条 ; 并且
该系统还包括 :
用于将该层次图中的每个词条与包括这每个词条的文档相关联的装置 ;
用于接收包括父词条的搜索查询的装置 ; 以及
用于接收与该父词条的子词条相关联的一个或多个文档的装置。

自动生成词条层次

技术领域

[0001] 本发明总体上涉及词典分析,更具体地涉及自动生成词条层次。

背景技术

[0002] 数据语料库 (corpus) 可保存大量信息,然而找到相关的信息可能很难。可以创建词条层次来方便搜索相关信息。然而,在某些场合下,生成词条层次的已知技术在某些场合下效率不高。

附图说明

- [0003] 图 1 图示了自动生成表示词条层次的层次图的系统的一个实施方式 ;
- [0004] 图 2 图示了可用于图 1 的系统的相关度 (affinity) 模块的一个实施方式 ;
- [0005] 图 3 图示了记录基本相关度的相关度矩阵的示例 ;
- [0006] 图 4 图示了记录有向相关度的相关度矩阵的示例 ;
- [0007] 图 5 图示了记录平均相关度的相关度矩阵的示例 ;
- [0008] 图 6 图示了相关度图的示例 ;
- [0009] 图 7 图示了可用于图 1 的系统的聚类模块的一个实施方式 ;
- [0010] 图 8 图示了可用于图 1 的系统的本体 (ontology) 特征模块的一个实施方式 ;
- [0011] 图 9 图示了可用于图 1 的系统的层次模块的一个实施方式 ;
- [0012] 图 10 图示了层次图的示例 ;而
- [0013] 图 11A 和 11B 图示了可从中移除冗余父子关系的节点。
- [0014] 特定实施方式
- [0015] 概述
- [0016] 在某些实施方式中,生成词条层次包括访问包括多个词条的语料库。对一个或多个词条执行下述步骤以生成父子关系:根据有向相关度来标识词条的一个或多个父词条;以及根据这些父词条和每个词条来建立一个或多个父子关系。根据这些父子关系而自动生成层次图。
- [0017] 示例实施方式
- [0018] 在具体实施方式中,创建和查询域本体可包括下述步骤:
- [0019] 1. 收集域中的文档。在具体实施方式中,文档是词条的集合。文档可包括例如《新约圣经》书的可读文本。文档不必包括叙述形式的文本,例如,文档可包括一组用户输入的、单独和共同描述图像内容的标签。文档的集合可被称为“域语料库”。
- [0020] 2. 标识域中感兴趣的词条(“词典词条”)。词条的示例包括词(如“树(tree)”)、短语(如“图形算法(graph algorithm)”)、命名实体(如“纽约(New York)”)等。词条(或概念)可以有不同的形式。在某些情况下,不同的词用于相同的概念,例如,“肾结石(kidney stones)”和“肾石病(kidney calculi)”指相同的概念,“肾结石”。在其他情况下,词干可以有許多词形变化,例如,词干“tree”具有词形变化“tree”和“trees”。在具体

实施方式中,相同词条的各形式可被认为是映射到相同的词条。词典词条的任何适当形式都可能在文档中出现,但是具体词典词条不必在任何文档中出现。

[0021] 用于标识词典词条的示例包括使用对于特定域的人类生成的词典,例如医学词典。在具体实施方式中,可以根据语料库中的一组文本串自动生成词典词条的列表。这些串可按频率来索引和分类,并且可选择频率在阈值之上的串。可以使用其他适当的统计方法来确定词条。在具体实施方式中,“词”可与“词条”和“词典词条”互换。

[0022] 3. 计算在给定共现 (co-occurrence) 上下文中词典词条的共现数量。如果两个词条在相同的共现上下文中都至少出现一次,则它们共现。共现上下文的示例包括文档和段落。

[0023] 4. 创建包括域本体的有向加权图。有向加权图包括作为节点的词典词条和作为边缘权重的相关度。“有向加权图”可用作可由任何适当数据结构表示的相同信息的实际表示,例如矩阵、二进制判决图或二进制判决图的集合。

[0024] 5. 应用查询有向加权图的过程。给定一个或多个词典词条作为输入,该过程输出了与输入词典词条有关的一个或多个词典词条。例如,该过程可输出对一个或多个输入词条具有最高差分有向相关度(下面进行描述)的一个或多个词条的分类列表。在此情况下,输出包括与本体所寻址的域相关的、与输入词条更紧密相关的词条。

[0025] 可使用任何适当的相关度定义。在具体实施方式中,可使用下述定义:

[0026] 1. 基本相关度

[0027] a. 词条 A 与 B 之间的基本相关度 (A) 可定义为包括词条 A 和 B 两者的共现上下文的数量与包括词条 A 或 B 的共现上下文的数量之比:

$$[0028] \quad A(A, B) = |AB| / |A \text{ or } B|$$

[0029] b. 词条 A 与 B 之间的基本相关度 (A) 还可定义为包括词条 A 和 B 两者的共现上下文的数量与包括 A 的共现上下文的数量和包括 B 的共现上下文的数量中的最大值之比:

$$[0030] \quad A(A, B) = |AB| / \max(|A|, |B|)$$

[0031] 2. 有向相关度

[0032] 词条 A 与 B 之间的有向相关度 (DAff) 可定义为假定在共现上下文中观察到 A 的情况下观察到 B 的条件概率:

$$[0033] \quad \text{DAff}(A, B) = |AB| / |A|$$

[0034] 即,有向相关度可以是在包括词条 A 和 B 两者的共现上下文的数量与包括词条 A 的共现上下文的数量之比。通常,DAff(A, B) 与 DAff(B, A) 不同。

[0035] 3. 差分有向相关度

[0036] 词条 A 与 B 之间的差分有向相关度 (DiffDAff) 可定义为词条 A 与 B 之间的有向相关度减去考虑到词条 B 在语料库中的共性的因子。词条 B 在语料库中的共性可以是词条 B 对语料库中的其他词条的基本相关度或有向相关度值的统计值。在具体实施方式中,词条 B 在语料库中的共性可以是词条 B 的平均相关度 (AA),这得出了下面的差分有向相关度的定义:

$$[0037] \quad \text{DiffDAff}(A, B) = \text{DA}(A, B) - \text{AA}(B)$$

[0038] 词条 B 的平均相关度 (AA) 或平均有向相关度可定义为:

$$[0039] \quad \text{AA}(B) = \text{AVERAGE}_x \text{ DAff}(x, B)$$

[0040] 即,平均相关度可以是词条 B 对共现上下文中的其他词条的有向相关度的平均。

[0041] 图 1 图示了自动生成表示词条的层次的层次图的系统 10 的一个实施方式。在具体实施方式中,系统 10 访问包括多个文档(或页面)的语料库。系统 10 根据有向相关度来标识语料库的词条的父词条,并且建立各词条之间的父子关系。系统 10 然后根据父子关系来自动生成层次图。

[0042] 在某些实施方式中,可对词典 D 和词的给定子集的特定反向索引 II 计算有向相关度,其中索引 II 例如包括词 w_i 和 w_j 的条目 $I(w_i)$ 和 $I(w_j)$ 。通常而言,反向索引是一种存储了从词条到其位置(即,词条出现的共现上下文)的映射的索引数据结构。对于 D 中的每对词 w_i 和 w_j ,可将 II 中的条目 $I(w_i)$ 和 $I(w_j)$ 的合取(conjunction)的值除以 $I(w_i)$ 中的值的数量定义为 $DA(i, j)$ 。通常, $DA(i, j)$ 不必等于 $DA(j, i)$ 。结果可按照例如逐行的任何适当方式进行存储,在逐行方式中,先存储 $D(1, i)$,然后存储 $D(2, j)$ 等等。对于每行 i ,可以存储 $|I(w_i)|$,然后存储与 w_j 合取的基数。

[0043] 在某些实施方式中,有向相关度可以按三个阶段来计算。在这些实施方式中,为每个词典词条分配了唯一的整数标识符。反向索引的条目对应于这些整数标识符。在阶段 0,读取对应于 D 的 II 条目。对于参数 (s, o) ,仅保留具有 $ks+o$ 的形式的元素标识符。值 $ks+o$ 定义限定了要检查的 II 条目的子集。这样,就可以并行地计算有向相关度。举例来说,来自参数 $s, o(1, 0)$ 的结果等于对参数 $(3, 0)$ 、 $(3, 1)$ 和 $(3, 2)$ 的计算进行合并而得到的结果。该步骤使得能够针对非常大的反向索引来计算 DA 表。

[0044] 在阶段 1,仅针对 $DA(i, j)$ 逐行计算合取。在阶段 2,读取计算出的上三角 UT DA 阵列。据此获得了作为 UT 的转置的下三角部分。在某些实施方式中,相同维数的多个 DA 阵列可以合并为单个阵列。大 II 上的 DA 阵列可以按带参数 (s, i) 的 $\sum_{i=0..(s-1)} DA$ 来计算。可与计算出的合取一起存储附加信息,使得能够计算有向相关度。在某些情况下,可以存储 II 条目的基数。

[0045] 在某些实施方式中,可逐行存储 DA,因此 AA 条目的计算可以与 DA 条目的计算并行进行。具体来讲,AA 可通过在从盘上读取 DA 时对它们的各行求和而生成,最后用词典条目的数量进行归一化。

[0046] 在图示的实施方式中,系统 10 包括客户机 20、服务器 22 以及存储器 24。客户机 20 允许用户与服务器 22 进行通信以生成语言的本体。客户机 20 可将用户输入发送到服务器 22,并且可以向用户提供(例如,显示或打印)服务器输出。服务器系统 24 对用于生成语言的本体的应用进行管理。存储器 24 存储了服务器系统 24 所使用的数据。

[0047] 在图示的实施方式中,存储器 24 存储了页面 50 和记录 54。页面 50(或文档或共现上下文)可指词的集合。页面 50 的示例包括文档的一个或多个页面、一个或多个文档、一本或多本书、一个或多个 web 页面、信件(例如电子邮件或即时消息)和/或词的其他词的集合。页面 50 可以由页面标识符来标识。页面 50 可以电子地存储在一个或多个有形计算机可读介质中。页面 50 可以与例如文本(如字符、词和/或数字)、图像(如图片、照片或视频)、音频(如录音或计算机生成的声音)和/或软件程序的任何适当的内容相关联。在某些实施方式中,一组页面 50 可属于语料库。语料库可与特定的主题、社区、组织或其他实体相关联。

[0048] 记录 54 用于描述页面 50。在实施方式中,记录 54 包括索引 58、反向索引 62、本

体 66 和聚类 (cluster) 67。索引 58 包括索引列表, 其中页面 50 的索引列表指示了页面 50 的词。反向索引 62 包括反向索引列表, 其中词 (或一组词) 的反向索引列表指示了包括该词 (或该组词) 的页面 50。在一个示例中, 列表 W_i 包括包含词 w_i 的页面 50 的页面标识符。列表 W_i 和 W_j 包括包含词 w_i 和 w_j 二者的合取页面 50 的页面标识符。列表 W_i+W_j 包括包含词 W_i 或 W_j 的析取页面 50 的页面标识符。 $P(W_i)$ 是 W_i 的页面 50 的数量, 即, 包括词 w_i 的页面 50 的数量。

[0049] 在一个实施方式中, 列表 (如, 索引列表或反向索引列表) 可被存储为二进制判决图 (BDD)。在一个示例中, 集合 W_i 的二进制判决图 $BDD(W_i)$ 表示包含词 w_i 的页面 50。 $BDD(W_i)$ 的满足指定计数 $Satisf(BDD(W_i))$ 得到了包含词 w_i 的页面 50 的数量 $P(W_i)$:

[0050] $P(W_i) = Satisf(BDD(W_i))$

[0051] 相应地,

[0052] $P(W_i \& W_j) = Satisf(BDD(W_i) \text{ AND } BDD(W_j))$

[0053] $P(W_i + W_j) = Satisf(BDD(W_i) \text{ OR } BDD(W_j))$

[0054] 本体 66 表示语言的词以及词间的关系。在一个实施方式中, 本体 66 表示词间的相关度。在图示的示例中, 本体 66 包括相关度矩阵和相关度图。相关度矩阵的示例将参照图 3 到图 5 来描述。相关度图的示例将参照图 6 来描述。聚类 67 记录了彼此相关的词的聚类。聚类将参照图 7 来更详细地描述。

[0055] 在图示的实施方式中, 服务器 22 包括相关度模块 30、聚类模块 31、本体特征模块 32 和层次模块 37。相关度模块 30 可计算词对的相关度、将相关度记录在相关度矩阵中, 和 / 或报告相关度矩阵。相关度模块 30 还可生成相关度图。相关度模块 30 将参照图 2 来更详细地描述。

[0056] 在特定实施方式中, 聚类模块 31 可通过标识数据集中相关元素的聚类来发现数据集中的模式。在特定实施方式中, 聚类模块 31 可标识一组词的聚类 (例如, 语言或一组页面 50)。通常而言, 聚类的词是彼此高度相关的, 但是与聚类外的词却并非如此。一聚类词可指定一组词的主题 (或话题)。在特定实施方式中, 聚类模块 31 根据词间的相关度来标识相关词的聚类。在这些实施方式中, 聚类的词彼此高度相关, 但与聚类外的词却并非如此。聚类模块 31 将参照图 7 来更详细地描述。

[0057] 在特定实施方式中, 本体特征模块 32 可确定一个或多个词 (例如, 特定词或包括多词的文档) 的集合的一个或多个本体特征, 然后可在各种情况下应用本体特征。本体特征是可将词组放置在语言的本体空间中的词组的特征。本体特征的示例包括深度和具体性 (specificity)。在特定实施方式中, 深度可表示词组的字面复杂度。较深的词组可以是更技术性和更专业性的, 而较浅的词组可以是更普通的。在特定实施方式中, 词组的具体性与词组的主题数量相关。较具体的词组可以有较少的主题, 而较不具体的词组可以有较多的主题。

[0058] 本体特征模块 32 可在任何适当的场合应用本体特征。适当场合的示例包括: 根据本体特征来搜索、分类或选择文档; 报告文档的本体特征; 以及确定一个或多个用户的文档的本体特征。本体特征模块 32 将参照图 8 来更详细地描述。

[0059] 在特定实施方式中, 层次模块 37 自动生成表示词条的层次的层次图。在这些实施方式中, 层次模块 37 访问包括多个文档 (或页面) 的语料库。层次模块 37 根据有向相关

度来标识语料库的词条的父词条,并且建立词条间的父子关系。层次模块 37 随后根据父子关系来自动生成层次图。层次模块 37 将参照图 9 来更详细地描述。

[0060] 系统 10 的组件可包括接口、逻辑、存储器和 / 或其他适当的元件。接口接收输入、发送输出、处理输入和 / 或输出,和 / 或执行其他适当的操作。接口可包括硬件和 / 或软件。

[0061] 逻辑执行组件的操作,例如执行根据输入来生成输出的指令。逻辑可包括硬件、软件和 / 或其他逻辑。逻辑可编码在一个或多个有形介质中,并且在被计算机执行时可执行操作。诸如处理器的某些逻辑可管理组件的操作。处理器的示例包括一个或多个计算机、一个或多个微处理器、一个或多个应用和 / 或其他逻辑。

[0062] 存储器存储有信息。存储器可包括一个或多个有形的计算机可读和 / 或计算机可执行存储介质。存储器的示例包括计算机存储器(例如,随机存取存储器(RAM)或只读存储器(ROM))、大容量存储介质(例如,硬盘)、可移除存储介质(例如,致密盘(CD)或数字视频盘(DVD))、数据库和 / 或网络存储(例如,服务器)和 / 或其他计算机可读介质。

[0063] 可以对系统 10 进行修改、添加或省略而不会偏离本发明的范围。系统 10 的组件可以是集成的或分立的。此外,系统 10 的操作可由更多、更少或其他组件来执行。例如,生成器 42 和 46 的操作可由一个组件来执行,或者相关度计算器 34 的操作可由多于一个组件来执行。此外,系统 10 的操作可使用包括软件、硬件和 / 或其他逻辑的任何适当逻辑来执行。在本文中使用时,“每个”是指集合的每个成员或集合的子集的每个成员。

[0064] 可对矩阵的示例进行修改、添加或省略而不会偏离本发明的范围。矩阵可包括更多、更少或其他值。此外,矩阵的值可按任何适当的顺序排列。

[0065] 图 2 图示了可用于图 1 的系统 10 的相关度模块 30 的一个实施方式。相关度模块 30 可计算词对的相关度、将相关度记录到相关度矩阵中和 / 或报告相关度矩阵。相关度模块 30 还可生成相关度图。

[0066] 在图示的实施方式中,相关度模块 30 包括相关度计算器 34、本体生成器 38 和词推荐器 48。相关度计算器 34 计算词 w_i 或包括第一词 w_i 和第二词 w_j 的词对的任何适当类型的相关度。相关度的示例包括基本、有向、平均、差分和 / 或其他相关度。

[0067] 在一个实施方式中,词推荐器 48 接收种词(seed word),并且标识与种词的相关度大于阈值相关度的词。阈值相关度可以为任何适当的值,如大于等于 0.25、0.5、0.75 或 0.95。阈值相关度可预编程或由用户指定。

[0068] 基本相关度可根据包括词 w_i 和 / 或 w_j 的页面 50 的量(例如,数量)来计算。合取页面量表示了包括词 w_i 和词 w_j 两者的页面 50 的量,而析取页面量表示了包括词 w_i 或词 w_j 的页面 50 的量。基本相关度可通过合取页面量除以析取页面量来给出。在一个示例中,合取页面的数量表示了包括词 w_i 和词 w_j 的页面的数量,而析取页面的数量表示了包括词 w_i 或词 w_j 的页面的数量。基本相关度可通过合取页面的数量除以析取页面的数量来给出:

[0069]
$$\text{Affinity}(w_i, w_j) = P(W_i \& W_j) / P(W_i + w_j)$$

[0070] 图 3 图示了记录基本相关度的相关度矩阵 110 的示例。在图示的示例中,相关度矩阵 110 记录了词 $w_1, \dots, w_5$ 的逐对相关度。根据相关度矩阵 110,词 w_0 和 w_1 之间的相关度是 0.003,词 w_0 和 w_2 之间的相关度是 0.005,等等。

[0071] 回来参照图 1,相关度组包括彼此具有高相关度的词对,并且可用于捕获词 w_1 和 w_2 之间相对于页面内容的关系。高相关度可以被指定为高于相关度组阈值的相关度。阈值可

设置为任何适当的值,如大于等于 0.50、0.60、0.75、0.90 或 0.95。词可属于多于一个的相关度组。在一个实施方式中,相关度组可表示为 BDD。BDD 的指针可与该组的每个词一起存储在反向索引 62 中。

[0072] 有向相关度可用于测量词 w_i 相对于词 w_j 的重要性。相关度计算器 34 根据包括词 w_i 和 w_j 的页面 50 的量(例如,数量)来计算在给定词 w_j 的情况下词 w_i 的有向相关度。词 w_j 页面量表示了包括词 w_i 的页面 50 的量。给定词 w_j 的情况下词 w_i 的有向相关度可由合取页面量除以词 w_j 页面量来给出。例如,词 w_j 页面的数量表示了包括词 w_i 的页面 50 的数量。给定词 w_j 的情况下词 w_i 的有向相关度可由合取页面 50 的数量除以词 w_i 页面 50 的数量来给出:

$$[0073] \quad \text{DAffinity}(w_i, w_j) = P(W_i \& W_j) / P(W_i)$$

[0074] $\text{DAffinity}(w_i, w_j)$ 与 $\text{DAffinity}(w_j, w_i)$ 不同。词 w_i 和 w_j 之间的高有向相关度 $\text{DAffinity}(w_i, w_j)$ 表示在页面 50 包括词 w_j 的情况下页面 50 包括词 w_i 的较高概率。在一个示例中,页面 [1 2 3 4 5 6] 包括词 w_i ,而页面 [4 2] 包括词 w_j 。包括词 w_j 的页面也包括词 w_i ,因此从词 w_j 的角度来看,词 w_i 是非常重要的。仅三分之一包括 w_i 的页面也包括词 w_j ,因此从词 w_i 的角度来看,词 w_j 是不重要的。

[0075] 图 4 图示了记录了词 $w_0, \dots, w_5$ 的有向相关度的相关度矩阵 120 的示例。在该示例中,词 124 是 A 词,而词 128 是 B 词。矩阵 120 的各行记录了在给定 A 词的情况下 B 词的相关度,并且相关度矩阵 120 的各列记录了在给定 B 词的情况下 A 词的相关度。

[0076] 回来参照图 1,词 w_i 的平均相关度是针对其他词 w_j 而计算的。在一个实施方式中,平均相关度可以是词 w_i 和每个其他词 w_j 之间的相关度的平均。词 w_i 对 N 个词的平均相关度可如下给出:

$$[0077] \quad \text{AveAff}(w_i) = \frac{1}{N} \sum_{j=1}^N P(w_i | w_j)$$

[0078] 图 5 图示了记录了平均相关度的相关度矩阵 140 的示例。行 142 记录了词 1 到词 50,000 的基本相关度。行 144 记录了词 1 到词 50,000 的平均相关度。

[0079] 回来参照图 1,词的平均相关度可表示词的深度。具有较低平均相关度的词可被认为是较深词,而具有较高平均相关度的词可被认为是较浅词。较深词趋于更技术性、具体和精确。具有较高百分比的较深词的页面 50 可被认为是较深页面,而具有较低百分比的较深词的页面 50 可被认为是较浅页面。在一个实施方式中,用户可指定要检索的词和 / 或页面 50 的深度。

[0080] 页面 50 的较深词可形成较高相关词的一个或多个聚类。聚类可表示共同的想法或主题。页面 50 的主题数量可以表示页面 50 的具体性。具有较少主题的页面 50 可被认为是更具体,而具有较多主题的页面 50 可被认为较不具体。

[0081] 词 w_i 相对于词 w_j 的差分相关度是词 w_i 与词 w_j 之间的有向相关度减去词 w_j 对于所有其他词的平均相关度。差分相关度可以表达为:

$$[0082] \quad \text{DiffAff}(w_i, w_j) = \text{DAffinity}(w_i, w_j) - \text{AveAff}(w_j)$$

[0083] 差分相关度移除了由词 w_j 在页面 50 中的出现的一般倾向引起的偏置。在特定环境中,差分相关度可提供在页面包括词 w_j 的情况下该页面包括词 w_i 的概率的更精确表示。

[0084] 差分相关度可用在各种应用中。在一个示例中,可以使用人名之间的差分相关度

来研究社会关系网。在另一示例中,可以使用语言元素之间的差分相关度来研究自然语言处理。在另一示例中,可以使用产品之间的差分相关度来研究市场。

[0085] 相关度计算器 34 可使用任何适当的技术来搜索反向索引列表以计算相关度。例如,为了标识包括词 w_i 和 w_j 两者的页面,相关度计算器 34 可在词 w_i 的列表 W_i 和词 w_j 的列表 W_j 中搜索共同元素,即,共同页面标识符。

[0086] 在特定实施方式中,本体生成器 38 生成语言的本体 66,如相关度矩阵或相关度图。本体可从诸如基本、有向、平均、差分和 / 或其他相关度中的任何适当相关度生成。本体 66 可从以任何适当方式从语言中选择的词来生成。例如,可以选择来自与一个或多个特定主题相关的词或语言的共同使用部分的词。

[0087] 在图示的实施方式中,本体生成器 38 包括相关度矩阵生成器 42 和相关度图生成器 46。相关度矩阵生成器 42 生成记录了词之间的相关度的相关度矩阵。相关度图生成器 46 生成表示词之间的相关度的相关度图。在相关度图中,节点表示词,而各节点之间的有向边缘的权重表示各节点所表示的词之间的相关度。相关度图可以有任何适当的维数。

[0088] 图 6 图示了相关度图 150 的示例。相关度图 150 包括节点 154 和链接 158。节点 154 表示词。在该示例中,节点 154a 表示词“二进制”。各节点 154 之间的有向边缘的权重表示各节点 154 所表示的词之间的相关度。例如,较大的权重表示较大的相关度。节点之间的链接 158 表示节点 154 所表示的词之间的相关度大于相关度阈值。相关度阈值可以为任何适当的值,例如大于等于 0.25、0.5、0.75 或 0.95。

[0089] 图 7 图示了可用于图 1 的系统 10 的聚类模块 31 的一个实施方式。在特定实施方式中,聚类模块 31 通过标识数据集中相关元素的聚类来发现数据集中的模式。在特定实施方式中,聚类模块 31 可标识词的集合的聚类(例如,页面 50 或语言的集合)。通常,聚类的词彼此高度相关,但是与聚类外的词却并非如此。一聚类词可指定词的集合的主题(或话题)。

[0090] 在特定实施方式中,聚类模块 31 根据词之间的相关度来标识相关词的聚类。在该实施方式中,聚类的词彼此高度相关,但是与聚类外的词却并非如此。在一个实施方式中,如果词足够相关,则它们可以被认为是高度相关的。如果它们满足一个或多个相关度准则(如阈值),则它们可以是足够相关的,其示例在下面给出。

[0091] 任何适当的相关度都可用于标识聚类。在特定实施方式中,聚类模块 31 使用了有向相关度。词相对于其他词的有向相关度表征了词的共现。聚类包括具有相似共现的词。在某些实施方式中,聚类模块 31 使用了差分相关度。差分相关度趋于移除由词在页面 50 中出现的一般倾向造成的偏置。

[0092] 在图示的实施方式中,聚类模块 31 包括聚类引擎 210 和聚类分析器 214。聚类引擎 210 根据相关度来标识词的聚类,而聚类分析器 214 应用相关度聚类以分析各种情况。

[0093] 聚类引擎 210 可按任何适当方式根据相关度来标识词的聚类。标识聚类的方法的三个示例给出如下:根据词的集合来构建聚类,将词排序为聚类,以及比较词的相关度矢量。在一个实施方式中,聚类引擎 210 根据词的集合来构建聚类。在一个示例中,聚类引擎 210 根据相关度为 $*Aff(w_i, w_j)$ 的词 $\{w_i\}$ 的集合 W 构建出聚类 S 。相关度值 $*Aff(w_i, w_j)$ 表示词 w_i 相对于词 w_j 的任何适当类型的相关度,例如有向相关度 $DAffinity(w_i, w_j)$ 或差分相关度 $DiffAff(w_i, w_j)$ 。这里提供的相关度值的某些示例可以被认为是归一值。在该示例

中, $\text{Aff}_{\text{for}}(w_i, w_j)$ 表示前向相关度, 而 $\text{Aff}_{\text{back}}(w_j, w_i)$ 表示后向相关度。

[0094] 在该示例中, 聚类 S 以种词 w_q 开始。当前词 w_x 表示聚类 S 的在当前迭代中与来自集合 W 的词进行比较的词。一开始, 当前词 w_x 被设置为种词 w_q 。

[0095] 在迭代过程中, 当前词 w_x 被设置为聚类 S 的词。根据集合 W 的词 w_i 与当前词 w_x 的前向相关度 $\text{Aff}_{\text{for}}(w_i, w_x)$ 对它们进行排序。以排序过的集合 W 的起点开始, 标识满足相关度准则的候选词 w_c 。相关度准则可包括与当前词 w_x 准则的前向相关度:

[0096] $\text{Aff}_{\text{for}}(w_c, w_x) > \text{Th}_{\text{cf}}$

[0097] 和与种词 w_q 准则的后向相关度:

[0098] $\text{Aff}_{\text{back}}(w_q, w_c) > \text{Th}_{\text{cb}}$

[0099] 其中 Th_{cf} 表示候选词的前向阈值, 而 Th_{cb} 表示候选词的后向阈值。候选词 $\{w_c\}$ 的有序集合的第一个词被添加到聚类 S , 添加的词的数量由参数 Size_c 给定。阈值 Th_{cf} 和 Th_{cb} 可以是范围从最小值到最大值的任何适当值的浮点参数。在某些示例中, Th_{cf} 和 Th_{cb} 的适当值可以根据实际相关度的有序列表来确定。例如, 可使用该列表中的第 200 个值。参数 Size_c 可以是任意适当值的整数参数。适当值的示例包括 1、2、3 或 4 的缺省值。在特定实施方式中, 参数可以在某些迭代时发生变化。

[0100] 可执行任何适当次数的迭代。在一个示例中, 可在方法开始之前指定迭代的次数。在另一示例中, 可以在执行方法期间计算次数。例如, 可以根据聚类 S 的大小的增长率来计算次数。

[0101] 在另一实施方式中, 聚类引擎 210 通过将词的集合中的词排序为聚类来标识聚类。在一个示例中, 根据相关度 $*\text{Aff}(w_i, w_j)$, 例如差分或有向相关度, 对集合 W 的词 $\{w_i\}$ 进行排序。在另一示例中, 根据词 w_i 与不同的词的集合 Q 的每个成员的相关度的合计功能 (如求和) 来对词 $\{w_i\}$ 进行排序。集合 W 可按任何适当方式来选择。例如, 集合 W 可以是与查询最相关的 X 个词, 其中 X 可以是任何适当的值, 如范围从 10 到 100、100 到 200, 或者 200 或更大的值。

[0102] 在该示例中, 聚类初始是空的。集合 W 中的第一个词 w_i 被放置在聚类中。在每次迭代时, 从集合 W 中选择当前词 w_x 。如果 $*\text{Aff}(w_x, w_f)$ 满足由相关度阈值 Th 给定的相关度准则, 则将当前词 w_x 放置到聚类中, 其中 w_f 表示放置在聚类中的第一个词。阈值 Th 可以为任何适当的值, 例如 0.1 到 0.5 范围内的值, 其最小值为 0.0, 最大值为 1.0。如果 $*\text{Aff}(w_x, w_f)$ 不满足阈值 Th , 则将当前词 w_x 放置到空聚类中。对集合 W 的每个词重复迭代。

[0103] 在处理了集合 W 的各词后, 可消除小的聚类。例如, 可消除少于 Y 个词的聚类。 Y 可以为任何适当的值, 如 3 到 5、5 到 10、10 到 25、25 到 50, 或者 50 或更大的范围内的值。

[0104] 如果聚类的数量不在令人满意的范围内, 则可用对放置在聚类中产生更严格或更松散的准则的阈值 Th 的不同值来重复该处理。令人满意的范围可由具有任何适当值的聚类号最小值和聚类号最大值来给定。适当值的示例包括对于最小值而言的 1 到 5、5 到 10, 或者 10 或更大范围内的值, 以及对于最大值而言的 10 到 15、15 到 20, 或者 20 或更大范围内的值。可增大阈值 Th 的值以增大聚类的数量, 并且可减小阈值 Th 的值以减小聚类的数量。

[0105] 在另一实施方式中, 聚类引擎 210 通过比较词的相关度矢量来标识聚类。在某些实施方式中, 相关度矩阵的行和列可产生相关度矢量 $\langle w_i, *\text{Aff}(w_i, w_1), \dots, *\text{Aff}(w_i,$

$w_j), \dots, \text{Aff}(w_i, w_n) \rangle$, 其表示词 w_i 相对于词 w_j 的相关度, $j = 1, \dots, n$ 。相关度值 $\text{Aff}(w_i, w_j)$ 表示词 w_i 相对于词 w_j 的任何适当类型的相关度, 例如, 有向相关度或差分相关度。

[0106] 在特定实施方式中, 具有相似相关度值的相关度矢量可表示一聚类。仅出于描述目的, 可以认为相关度矢量是相关度空间中词的相关度的坐标。即, 每个相关度值 $\text{Aff}(w_i, w_j)$ 都可被认为是特定维的坐标。具有相似相关度值的相关度矢量表示矢量与其相关联的词在相关度空间中彼此接近。即, 矢量表示词与其他词具有相似相关度关系, 因此可适用于相同聚类中的成员关系。

[0107] 如果一个相关度矢量邻近由适当距离函数确定的另一相关度矢量, 则相关度矢量可能是类似的。距离函数可以在相关度矢量上被定义为例如给定大小的矢量的标准欧几里得距离, 或被定义为给定大小的矢量的余弦。距离函数可以由聚类引擎 210 或用户来指定。

[0108] 在特定实施方式中, 聚类引擎 210 应用聚类算法来标识彼此接近的相关度矢量。聚类算法的示例包括直接、重复对分 (bisection)、聚集、偏置聚集和 / 或其他适当算法。在一个示例中, 聚类引擎 210 可包括如 CLUTO 的聚类软件。

[0109] 聚类分析器 214 可使用相关度聚类在任何适当应用中进行分析。在一个实施方式中, 聚类分析器 214 可使用相关度聚类来归类页面 50。类别可与聚类标识符或聚类的一个或多个成员相关联。在一个示例中, 可标识页面 50 的聚类, 然后可根据这些聚类来对页面 50 进行归类。在另一示例中, 可选择页面 50 的重要词, 然后可定位包括该词的聚类。然后可以根据所定位的聚类对页面 50 进行归类。

[0110] 在一个实施方式中, 聚类分析器 214 可使用相关度聚类来分析页面 50 的语料库。语料库可与特定主题、一个或多个个体的社区、组织或其他实体相关联。在一个示例中, 聚类分析器 214 可标识语料库的聚类, 并根据这些聚类来确定语料库的语料库特性。语料库特性可表示涉及与语料库相关联的实体的词。如果一个或多个页面 50 具有语料库特性的聚类, 则该页面 50 可能与该实体相关。

[0111] 在一个实施方式中, 聚类分析器 214 可使用相关度聚类来搜索查询解疑和扩展。在该实施方式中, 聚类分析器 214 标识包括给定搜索查询的搜索词条的聚类。聚类提供了与给定搜索查询相关的词和 / 或分类。在一个示例中, 可以将聚类中的词报告给搜索人以帮助下一次搜索查询。在另一示例中, 聚类分析器 214 可从这些聚类中选择词, 并且自动形成一个或多个新的搜索查询。聚类分析器 214 可串行或并行地运行新的查询。

[0112] 在一个实施方式中, 聚类分析器 214 可使用相关度聚类来研究社会网络。在一个示例中, 页面 50 可提供对社会网络的了解。这种页面的示例包括通讯 (如信件、电子邮件以及即时消息)、备忘录、文章和会议记录。这些页面 50 可包括包含社会网络的人的用户标识符 (如姓名) 的词。可以标识姓名的聚类以分析网络的人之间的关系。在一个示例中, 可使用差分相关度聚类来滤出页面 50 中出现最多的姓名, 而不提供诸如系统管理者姓名的信息。

[0113] 在特定实施方式中, 聚类分析器 214 可通过组合和 / 或比较数据集的聚类来分析数据集。在一个实施方式中, 对重叠数据集的聚类进行比较。可以将一个数据集的聚类映射到另一数据集的聚类, 这可以提供对数据集之间的关系的了解。例如, 数据集可来自对同事组的文档的分析和来自对该组的社会关系网的研究。社会网络聚类可被映射到文档主题聚类以分析社会网络与主题之间的关系。

[0114] 图 8 图示了本体特征模块 32 的一个实施方式。本体特征模块 32 可确定一个或多个词的集合的一个或多个本体特征（例如，特定词或包括多个词的文档），然后可在多种情况下的任何情况下应用本体特征。一个或多个词的集合可包括文档的必要词条。如果与词条 t 相关的前 k 个词中的至少一个在文档中也存在，则词条 t 可能是必要词条。否则，该词条对文档可能是非必要的。

[0115] 本体特征是在语义含义上而言，沿可将文档与给定区域中的其他文档相区分的一个或多个特征轴来表征文档的可计量的测量结果。例如，文档的深度可针对其可理解度来区分文档，文档的具体性可针对其关注度来区分文档，而文档的主题可针对其主题的可寻址范围来区分文档。本体特征可按任何适当方式来定义。例如，计算语言中的独立算法可用来表征文档的可读性或深度。

[0116] 在图示的实施方式中，本体特征模块 32 包括深度引擎 230、主题引擎 240、具体性引擎 244 以及本体特征 (OF) 应用引擎 250。深度引擎 230 可确定一个或多个词（例如，特定词或包括多个词的文档）的深度。通常，深度可表示词的字面复杂度。较深的词可能是较技术和专业化的，而较浅的词可能是较普通的。在特定实施方式中，深度模块 32 可计算文档的词深度，然后根据词深度来计算文档深度。在特定实施方式中，深度引擎 230 可将深度值和 / 或深度等级分配给文档和 / 或词。较深的文档或词可被分配较高的深度值或等级，而较浅的文档或词可被分配较浅的深度值或等级。

[0117] 深度引擎 230 可按任何适当方式来计算词深度。在特定实施方式中，深度引擎 230 根据平均相关度来计算词深度。在这些实施方式中，词深度是该词的平均相关度的函数。较深的词可具有较低的平均相关度，而较浅的词可具有较高的平均相关度。在特定示例中，深度引擎 230 可通过根据多个词的平均相关度对它们进行排序来计算这些词的深度。具有较低平均相关度的词可被赋予较高的深度等级，而具有较高平均相关度的词可被赋予较低深度等级。

[0118] 在特定实施方式中，深度引擎 230 可使用聚类分析来计算词深度。在这些实施方式中，聚类的词彼此高度相关，但与聚类外的词却并非如此。聚类空间中的距离可根据作为深度的指示符的相关度来测量。在特定实施方式中，属于较少的聚类或较小的和 / 或距其他聚类较远的聚类的词可被认为较深，而属于较多的聚类或较大和 / 或距其他聚类较近的聚类的词可被认为较浅。

[0119] 在其他特定实施方式中，深度引擎 230 可通过向相关度图 150 应用链接分析来计算词深度。链接分析可通过例如 PAGERANK 的任何适当的链接分析算法来执行。仅出于描述目的，可使用图 6 的相关度图 150 来计算词深度。相关度图 150 包括节点 154 和链接 158。节点 154 表示词。节点 154 之间的链接 158 表示节点 154 所表示的词之间的相关度在相关度阈值之上，即，这些词令人满意地相关。

[0120] 在特定实施方式中，深度引擎 230 计算节点 154 的流行度。较流行的节点 154 可表示较浅的词，而较不流行的节点 154 可表示较深的词。从第一节点 154 到第二节点 154 的链接 136 被认为是第一节点 154 对第二节点 154 的流行度表决 (vote)。此外，来自较流行节点 154 的表决可比来自较不流行节点 154 的表决具有更高的权重。此外，第一节点 154 对第二节点 154 的相关度加权了该表决。深度引擎 230 根据节点 154 的加权表决来计算节点 154 的流行度。较不流行的词可被认为较深，而较流行的词可被认为较浅。

[0121] 深度引擎 230 可按任何适当方式来计算文档深度。在特定实施方式中,深度引擎 230 根据文档的至少一个、一些或全部词的深度来计算文档的深度。在某些实施方式中,词深度由平均相关度给出,因此文档深度可根据文档的词的的平均相关度来计算。例如,文档的浅度可以是文档的词的的平均相关度的平均,即,文档中每个词的平均相关度的和除以文档中词的总数。文档的深度则可以按文档的浅度的倒数来计算得出。

[0122] 在特定实施方式中,深度可根据文档的所选词的集合来计算平均深度。所选集合可包括文档的必要的词,如前(最深的)X%的词,其中 X 可以小于 10,在 10 到 20、20 到 30、30 到 40、40 到 50、50 到 60、60 到 70 之间或大于 10。所选集合可排除 P%的标准语法词和 / 或 Q%的无用词,其中 P 和 Q 为任何适当的值,如小于 10,在 10 到 20、20 到 30、30 到 40、40 到 50、50 到 60、60 到 70 之间或大于 10。

[0123] 在特定实施方式中,深度引擎 230 根据文档中词深度的分布来计算文档的深度。在特定实施方式中,较深的文档可具有较高百分比的较深词。

[0124] 在特定实施方式中,深度引擎 230 根据文档相关度来计算文档的深度。文档之间的相关度描述了文档之间的关系。在某些实施方式中,平均文档相关度可以类似于平均词相关度如何表示词深度的方式来表示文档深度。文档相关度可按任何适当方式来定义。在一个示例中,共同词 $P(D_1 \& D_2)$ 的数量表示了文档 D_1 和 D_2 中都有的词的数量,而独特词 $P(D_1 + D_2)$ 的数量表示了文档 D_1 或 D_2 中的词的数量。文档 D_1 和 D_2 之间的文档相关度 DocAff 可定义为:

$$[0125] \quad \text{DocAff}(D_1, D_2) = P(D_1 \& D_2) / P(D_1 + D_2)$$

[0126] 深度引擎 230 可按类似于计算平均词相关度的方式来计算平均文档相关度。具有较低平均相关度的文档可被认为较深,而具有较高平均相关度的文档可被认为较浅。

[0127] 在某些实施方式中,深度引擎 230 可通过将链接分析应用到文档相关度图来计算文档深度。除了文档相关度图的节点表示文档而不是词以外,文档相关度图可类似于相关度图 150。深度引擎 230 用在给定第一文档的情况下的第二文档的文档相关度,来对从表示第一文档的第一节点到表示第二文档的第二节点的链接进行加权。然后可以对得到的链接的权重进行归一化。

[0128] 在某些实施方式中,深度图可显示在用户界面上以示出文档的深度。还可以显示可用于选择深度级别的深度滑块。在某些实施方式中,如果文档包括较大文档的一些部分,则深度图可表示这些部分的深度。

[0129] 在某些实施方式中,深度引擎 230 可按任何其他适当方式计算文档深度,如处理文档的相关度的直方图和 / 或基于深度来删减独特的词的百分比然后处理直方图。其他方法包括 Gunning-Fog、Flesch 或 Fry 方法。

[0130] 在某些实施方式中,深度引擎 230 可通过将深度值映射到特定深度级别来校准深度。在某些实施方式中,范围 R_i 中的深度值可被映射到级别 L_i 。例如, $R_0 = \{r_0: r_0 < c_0\}$ 可被映射到级别 L_0 , $R_1 = \{r_1: c_0 < r_1 < c_1\}$ 可被映射到级别 $L_1, \dots$, 而 $R_n = \{r_n: c_n < r_n\}$ 被映射到级别 L_n 。范围可包括任何适当的深度值,并且不必为相同大小。可存在任何适当数量的级别,如小于 5、5 到 7、7 或 8、8 到 10、10 到 20、20 到 50、50 到 100,或大于 100。

[0131] 主题引擎 240 可确定文档的主题(或主题)。在某些实施方式中,主题引擎 240 根据文档中词的聚类来确定主题,这些聚类可通过聚类模块 31 来标识。如上所述,一聚类词

可指示词的集合的主题（或主题）。文档的主题可提供关于文档的内容的有用信息。例如，包括聚类 { 肾脏, 肾, 蛋白质, 问题 } 的文档可能关于由于肾脏功能弱化而从肾导致的蛋白质的问题，而不是云豆的蛋白质含量。

[0132] 在某些实施方式中，主题引擎 240 根据主题地图来确定主题。在这些实施方式中，使用例如词条频率反向文档频率 (TF-IDF) 技术的任何适当技术从文档中提取关键词。使用关键词从主题地图中选择候选主题。将候选主题与文档进行比较以确定主题与文档有多匹配。在某些示例中，可以将候选主题的直方图与文档的直方图进行比较。如果候选主题与文档匹配，则这些主题可提供该文档的主题的类型和数量的估计。

[0133] 具体性引擎 240 可计算文档的具体性。在特定实施方式中，具体性引擎 240 可为文档分配具体性值和 / 或具体性等级。较具体的文档可被分配较高的具体性值或等级，而较不具体的文档可被分配较低的具体性值或等级。

[0134] 在特定实施方式中，具体性引擎 240 根据文档的主题的数量来计算具体性。在某些示例中，较具体的文档可以有较少的主题，而较不具体的文档可以有较多的主题。在特定实施方式中，具体性引擎 240 根据文档的主题的数量和主题之间的相关度来计算具体性。在某些示例中，较具体的文档可以有较少的主题，且各主题之间具有较高的相关度，而较不具体的文档可以有较多的主题，且各主题之间具有较低的相关度。

[0135] 在特定实施方式中，主题的数量可依赖于深度（或级别）。例如，处于较浅深度的单独主题可表示处于较大深度的多个主题。在某些实施方式中，深度可由用户使用深度滑块来选择或可以是预定的。在某些实施方式中，级别可由用户选择或可以是预定的。例如，可定义任何适当数量的级别，并且可相对于级别来计算深度。例如，级别可以是基于领域的（例如，工程、医疗、新闻、体育或财经领域）；基于专业的（例如，心脏病学、眼科学或肾脏学专业）；基于主题的（例如，高血压、胆固醇、支路手术或动脉阻塞主题）；基于细节的（例如，姿态血压过低、慢性高血压或急性高血压细节）；基于解决方案的（例如，老年病因学、医疗或通用解决方案）；基于个人的（例如，用户查询级别）。

[0136] 本体特征应用引擎 250 可应用本体特征（如，深度、主题或具体性）在任何适当场合下执行本体特征分析。适当场合的示例包括：根据本体特征搜索、排序、推荐或选择文档；报告文档的本体特征；以及确定一个或多个用户的文档（或文档的集合）的本体特征。在特定实施方式中，本体特征应用引擎 250 可使用包括关于本体特征的信息的索引。在一个示例中，本体特征应用引擎 250 使用根据深度排序而生成和 / 或维持的文档深度 (DD) 反向索引 62。DD 反向索引 62 包括 DD 反向索引列表，其中词的 DD 反向索引列表列出了包括该词的文档（或页面 50）的文档标识符。文档的文档标识符可表示文档的深度。例如，用于编码文档标识符的二进制编码可表示深度。在一些情况下，DD 反向索引列表可仅列出深度令人满意的文档。在另一示例中，本体特征应用引擎 250 除了反向索引 62 外还使用排序表和深度表。深度表可表示文档的深度。

[0137] 在特定实施方式中，本体特征应用引擎 250 在文档中搜索本体特征的特定值，如文档深度或具体性的特定值。特定值可以由用户预定、计算或选择。在特定实施方式中，可使用深度滑块和 / 或具体性滑块来选择这些值。

[0138] 在特定实施方式中，本体特征应用引擎 250 可使用本体特征作为排序准则来排序文档。例如，本体特征应用引擎 250 可根据关于主题的文档深度和 / 或具体性以及其它排

序准则来排序文档。在某些示例中,本体特征应用引擎 250 搜索 DD 反向索引 62 以获得根据文档深度而排序过的文档。在一些示例中,本体特征应用引擎 250 使用非 DD 反向索引 62 来搜索文档,然后根据深度来排序文档。

[0139] 在特定实施方式中,本体特征应用引擎 250 可将本体特征的值图形显示给客户机 20。可对一些或全部文档,例如搜索结果的前 X% 的文档,提供图形显示。本体特征值可按任何适当方式呈现。在一些示例中,如数字、词或图标图形指示符可表示值。图形指示符可放置得邻近于例如搜索结果的列表中的项目、在线新闻的主题或文档图标。在一些示例中,现有插图的修改可表示值。例如,文本的大小、字体、样式、颜色或图形指示符可表示值。在另一示例中,图形可表示值。本体特征直方图可包括文档量轴和本体特征轴,并且可表示特定本体特征值的文档的量。例如,包括文档量轴和文档深度轴的文档深度直方图可表示特定文档深度的文档量。

[0140] 在特定实施方式中,本体特征应用引擎 250 可允许用户请求搜索具有特定本体特征值的文档。用户可被允许为不同词的查询指定值。在某些示例中,本体特征引擎 250 可为用户提供选择深度的选项,然后用户可输入所选深度。选项可按任何适当方式呈现,如:
(i) 绝对词条(例如,表示深度的数字或数字范围);(ii) 相对词条(例如,搜索结果关于深度的一部分,如“最深的 X%”);(iii) 语义词条(例如,“介绍性的”、“浅”、“深”、“非常深”和/或“技术含量高”);(iv) 图形词条(例如,滑块、按钮和/或其他图形元素);或(v) 词条的任何适当组合(例如,具有语义标签的滑块)。在一些情况下,滑块可包括浅端和深端。用户可将滑块朝一端或另一端移动以表示所选深度。在提供搜索结果时,文档深度直方图可通过滑块来表现,并且可使用滑块作为文档深度轴。

[0141] 在特定实施方式中,本体特征应用引擎 250 可计算一个或多个用户的集合的本体特征特性。本体特征特性可包括主题的上下文中的用户深度和用户具体性。本体特征特性描述了与用户集合相关联的文档的本体特征。例如,科学家可使用比三年级学生将使用的文档更深的文档。本体特征特性可关于一个或多个主题而给定。例如,遗传学者在遗传学领域可使用比他在诗歌领域中使用的更深的文档。本体特征特性可用于确定用户的专门技能,自动建立用户的简历,并分析用户的社会网络。

[0142] 与用户关联的任何适当文档都可被分析以估计本体特征特性,例如,通讯(如,电子邮件和即时消息)、web 页面以及搜索历史(如,搜索查询和选择的页面)。在特定实施方式中,本体特征应用引擎 250 可随时间跟踪本体特征特性,并且可使用过去的特性来预测未来的特性。在某些示例中,本体特征应用引擎 250 可假定用户深度和/或具体性通常随时间和/或该领域的活动而增加。

[0143] 在特定实施方式中,本体特征应用引擎 250 可组合某些操作。例如,本体特征应用引擎 250 可监视用户的深度,然后根据该用户深度来搜索文档。在一个示例中,对用户深度进行监视,并根据深度向用户提供新闻。对未来的用户深度进行预测,并且提供符合预测的用户深度的新闻。

[0144] 图 9 图示了可用于图 1 的系统 10 的层次模块 37 的一个实施方式。在特定实施方式中,层次模块 37 生成表示词条的层次的层次图。在这些实施方式中,层次模块 37 访问诸如页面 50(或文档)的语料图。层次模块 37 根据有向相关度来标识语料库的词条的父词条,并且建立父子关系。层次模块 37 然后根据这些父子关系来自动生成层次图。在图示

的实施方式中,层次模块 37 包括图引擎 320 和层次应用模块 322。图引擎 320 生成层次图 350。层次图的示例参照图 10 来描述。

[0145] 图 10 图示了层次图 350 的示例。在特定实施方式中,层次图 350 表示了词条的层次。在这些实施方式中,层次图 350 包括表示词条的节点 354 和表示从父词条到子词条的父子关系的弧 358。在特定实施方式中,在父子关系中,父词条和子词条通过诸如有向相关度的相关度而相关。父词条可以比子词条更一般,相反,子词条可能比父词条更具体。

[0146] 在特定实施方式中,祖父词条(或曾祖父词条等)与子词条之间的关系也可被称为父子关系。即,父子关系可从父词条开始,经过零、一或更多父/子词条(充当关系中的父词条和子词条),并且以子词条结束。换句话说,父子关系可以有任何适当的代数。

[0147] 回来参照图 9,在图示的实施方式中,图引擎 320 包括父子标识符 324 和图缩减器(graph reducer)328。在特定实施方式中,父子标识符 324 标识了词条之间的父子关系,而图缩减器 328 缩减层次图 350。

[0148] 在特定示例中,图引擎 320 为语料库 C 的文档的集合 S 生成层次图 350。文档可包括一个或多个词条。页面(X,C)表示集合 S 的包括至少有词条 X 的一个共现的文档的子集。页面(X+Y,C)表示包括集合 S 的至少有词条 X 的一个共现和词条 Y 的至少一个共现的文档的子集。DocFreq(X,C)表示包括词条 X 的至少一个共现的文档的数量,并可根据下式来计算:

$$[0149] \quad \text{DocFreq}(X, C) = |\text{Pages}(X, C)|$$

[0150] 如上所述,有向相关度可用于测量给定词条对另一词条有多重要。在特定实施方式中,假定文档有第一词条,有向相关度表示了文档有第二词条的概率。如果概率高,则存在高有向相关度,而如果概率低,则存在低有向相关度。例如,淋巴瘤是一类癌症。包括“淋巴瘤”的文档还可能包括“癌症”。因此,从淋巴瘤到癌症存在高有向相关度。然而,反过来就未必如此。存在若干种类的癌症,所以包括“癌症”的许多文档可能并不包括“淋巴瘤”。因此,从癌症到淋巴瘤存在低有向相关度。

[0151] 有向相关度可按任何适当方式(例如使用上述表达)来表达。在其他实施方式中,DAff(X,Y,C)表示在语料库 C 中给定 X 的情况下 Y 的有向相关度,并且可根据下式来计算:

$$[0152] \quad \text{DAff}(X, Y, C) = \text{DocFreq}(X+Y, C) / \text{DocFreq}(X, C)$$

[0153] 在特定实施方式中,父子标识符 324 通过对于给定词条 X 确定候选父词条 Y_i 的集合来标识父子关系。在这些实施方式中,从给定词条 X 到父词条 Y_i 存在高有向相关度,而父词条 Y_i 出现在没有给出词条 X 的文档中比出现在有词条 X 的文档中更多,这表示词条 X 与父词条 Y_i 相关,但并非对父词条 Y_i 必不可少。

[0154] 在特定实施方式中,从给定词条 X 到父词条 Y_i 存在高有向相关度的条件可表达为:

$$[0155] \quad \text{DAff}(X, Y_i, S) \geq \text{Th}_{\text{high}}$$

[0156] 其中 Th_{high} 代表表示较高有向相关度的较高有向相关度阈值。较高的有向相关度阈值可以为任何适当的值,如 0.5 或更大、0.6 或更大、0.7 或更大的范围内的值。

[0157] 在特定实施方式中,父词条 Y_i 出现在没有给出词条 X 的文档中比出现在有词条 X 的文档中更多的条件且词条 X 与父词条 Y_i 相关但并非对父词条 Y_i 必不可少的条件可表达为:

[0158] $\text{DAff}(Y_i, X, S) \leq \text{Th}_{\text{low}}$

[0159] $\text{DAff}(Y_i, X, S) \geq \text{Th}_{\text{min}}$

[0160] 其中 Th_{low} 代表表示较低有向相关度的较低有向相关度阈值,而 Th_{min} 代表最小有向相关度阈值。较低有向相关度阈值可以为任何适当的值,如 0.5 或更低、0.4 或更低、或 0.1 或更低的范围内的值。最小有向相关度阈值可以为任何适当的值,如 0.1 或更低、0.01 或更低、或 0.001 或更低的范围内的值。满足给定词条的上述条件的候选父词条 Y_i 可被认为是给定词条的父词条,该给定词条可被认为是子词条。

[0161] 可根据父子关系将表示词条的节点 354 组织为层次图 350。在特定实施方式中,图缩减器 328 通过移除冗余父子关系来缩减层次图 350。参照图 11A 和 11B 来描述移除冗余父子关系的示例。

[0162] 图 11A 和 11B 图示了可从中移除冗余父子关系的节点。冗余父子关系可以是与给定父子关系冗余的父子关系。冗余父子关系可提供比给定父子关系所提供的相同或更少的信息。例如,给定父子关系可具有父词条、零个、一个或多个父 / 子词条以及子词条。冗余父子关系可具有相同父词条、相同父 / 子词条的子集以及相同子词条,而没有其他词条。换句话说,子词条有彼此链接的两个父词条。

[0163] 在图示的示例中,给定父子关系有父词条“健康”、父 / 子词条“疾病”以及子词条“糖尿病”。冗余父子关系有父词条“疾病”与子词条“糖尿病”。图缩减器 328 可移除该冗余父子关系。

[0164] 回来参照图 9,在特定实施方式中,层次应用模块 322 应用层次图 350。层次模块 37 可根据任何适当的语料库来生成任何适当的层次图,层次应用模块 322 可按任何适当方式来应用该语料库。

[0165] 在特定实施方式中,语料库包括包含词条的文档,并且层次图 350 充当可用于搜索文档的文档的类别。在实施方式的特定示例中,层次应用模块 322 将层次图 350 的父词条发送给客户机。用户可从父词条中选择作为搜索词条的词条。层次应用模块 322 接收父词条的选择,并将所选父词条的子词条发送给客户机。用户可从这些词条中进行选择,并且可重复这种处理,直到达到层次图 350 的终点或用户已经指示了最终选择为止。

[0166] 在实施方式的其他示例中,层次应用模块 322 接收包括层次图的父词条的搜索查询。层次应用模块 322 标识父词条的子词条,并使用父词条和子词条来搜索语料库。在实施方式的另一示例中,层次图 350 的每个图词条都可与包括该图词条的文档相关联。层次应用模块 322 接收包括层次图 350 的父词条的搜索查询。层次应用模块 322 随后检索一个或多个与该父词条或该父词条的子词条相关联的文档。

[0167] 在特定实施方式中,语料库包括搜索结果,并且层次图 350 根据搜索结果的父子关系来组织搜索结果。层次图 350 的词条可表示搜索结果或者搜索结果所出现的聚类的主题。层次应用模块 322 可将层次图 350 和 / 或组织好的搜索结果呈现给用户。

[0168] 可对该方法进行修改、添加或省略而不偏离本发明的范围。该方法可包括更多、更少或其他的步骤。此外,各步骤可按任何适当顺序来执行。

[0169] 本发明的某些实施方式可提供一个或更多技术优点。一个实施方式的技术优点可能在于,表示词条层次的层次图是对文档的语料库自动生成的。层次图描述了词条之间的父子关系并且可用于对语料库的文档进行归类。另一实施方式的技术优点可能在于对搜索

结果自动生成层次图。层次图可用于组织搜索结果。

[0170] 尽管就某些实施方式描述了本发明,但是本领域技术人员可以想到这些实施方式的更改和替换。因此,上述实施方式的描述并不限制本公开。其他改变、替代和更改都是可能的,而不会偏离由权利要求书定义的本公开的精神和范围。

[0171] 本申请在 35U.S.C. § 119(e) 下要求 David Marvit 等人于 2007 年 10 月 5 日提交的题为“Automatic Generation of a Hierarchy of Terms”的美国临时申请第 60/977,840 号(律所案号:073338.0551)的优先权。

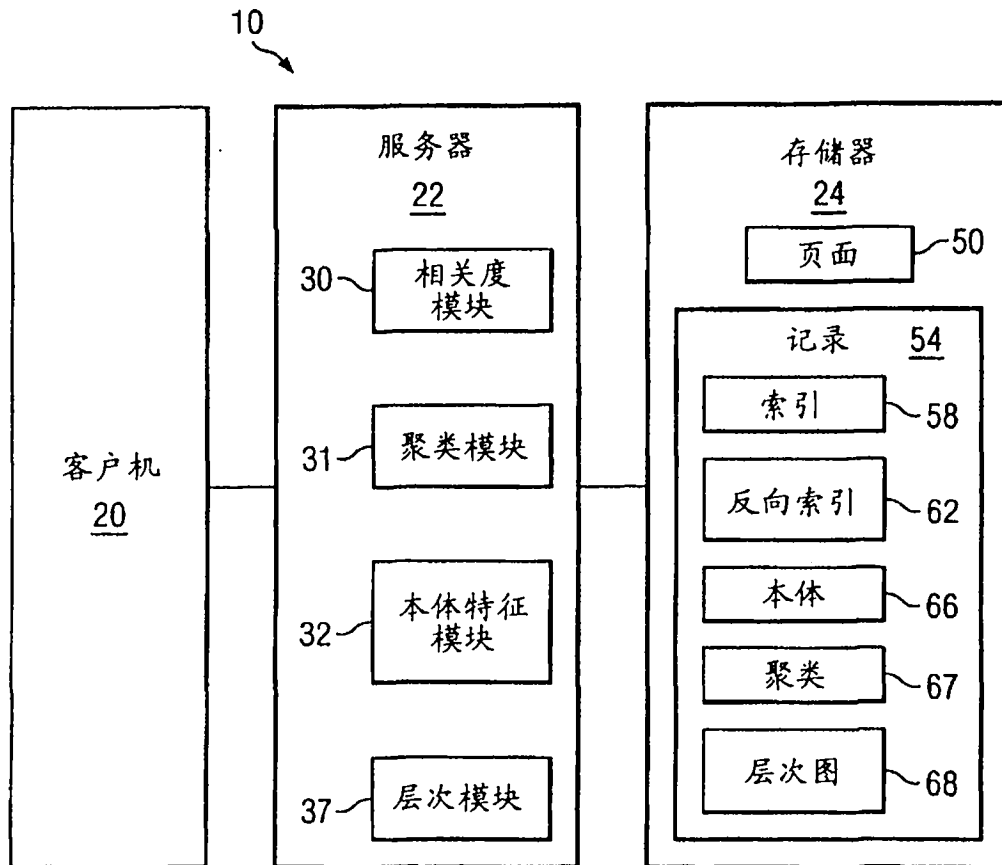


图 1

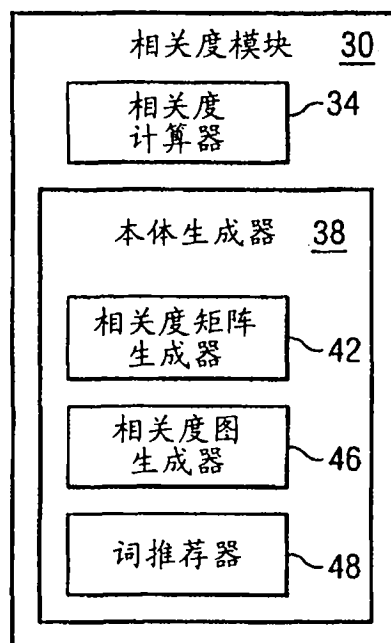


图 2

2/4

110

	w ₁ 狗	w ₂ 森林	w ₃ 树	w ₄ 图片	w ₅ 计算机
w ₀ 二进制	0.003	0.005	0.037	0.021	0.066
w ₁ 狗		0.024	0.033	0.017	0.049
w ₂ 森林			0.092	0.004	0.052
w ₃ 树				0.042	0.056
w ₄ 图片					0.222

图 3

A
124

120

	w ₀ 二进制	w ₁ 狗	w ₂ 森林	w ₃ 树	w ₄ 图片	w ₅ 计算机
w ₀ 二进制	1	0.004	0.005	0.016	0.020	0.037
w ₁ 狗	0.018	1	0.022	0.026	0.016	0.047
w ₂ 森林	0.013	0.013	1	0.055	0.008	0.026
w ₃ 树	0.071	0.029	0.102	1	0.034	0.060
w ₄ 图片	0.071	0.013	0.012	0.026	1	0.255
w ₅ 计算机	0.360	0.112	0.103	0.128	0.716	1

A → B

图 4

140

	词1	词2	词3	[...]	词 50,000
R ₀ 142	词1	----- 0.005	0.037	[...]	0.066
	词2		----- 0.033	[...]	0.049
	词3			----- [...]	0.052
	[...]			----- [...]	
	词 50,000				-----
144	平均	AA1	AA2	AA3	[...] AA50,000

图 5

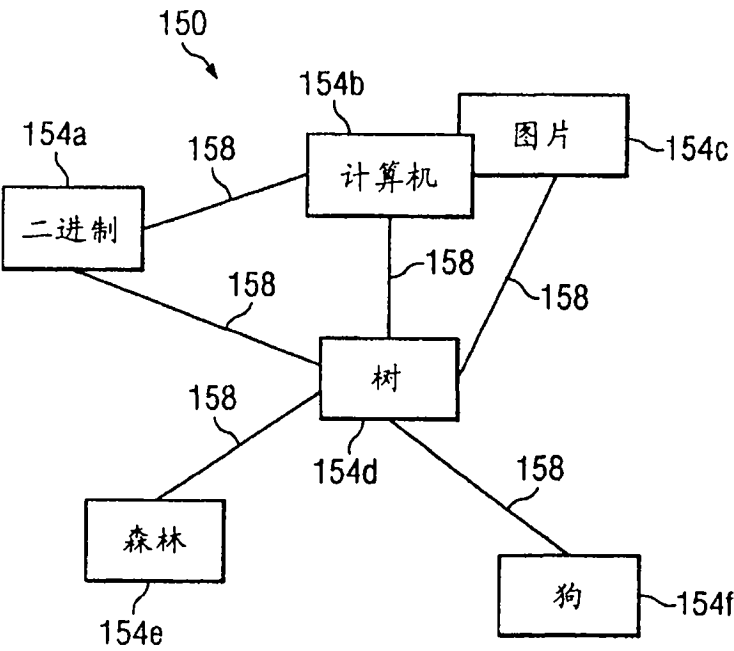


图 6

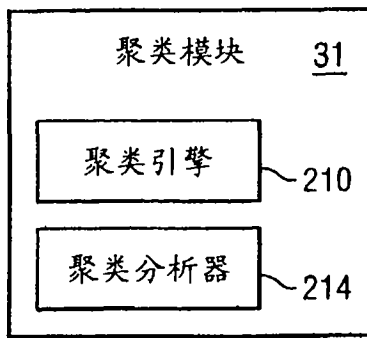


图 7

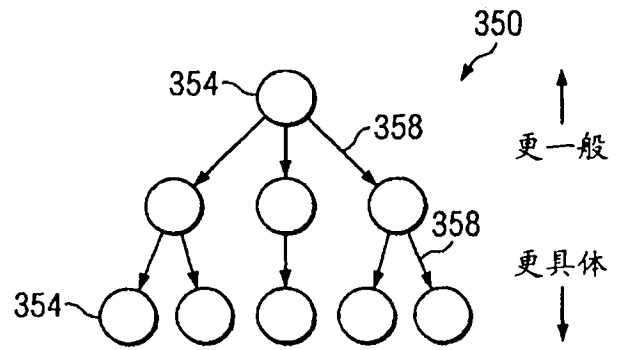


图 10

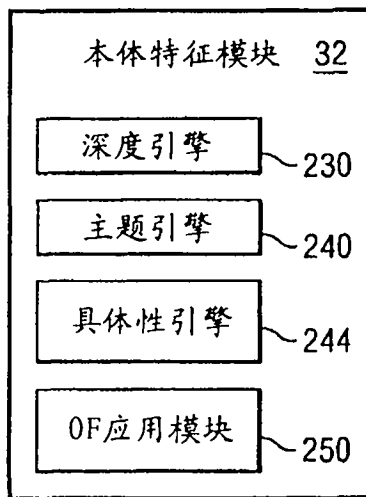


图 8



图 11A

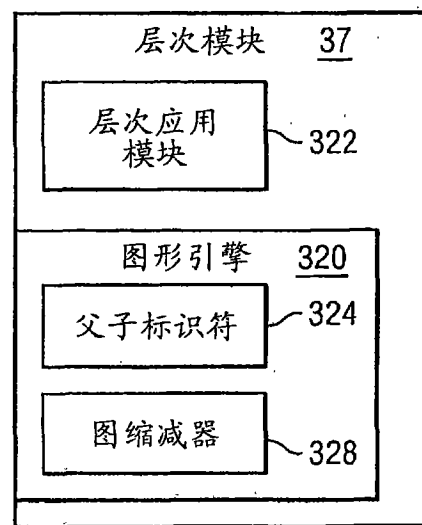


图 9



图 11B