[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 17/30 (2006.01)

[12] 发明专利申请公布说明书

[21] 申请号 200810166181.X

[43] 公开日 2009 年 4 月 15 日

[11] 公开号 CN 101408886A

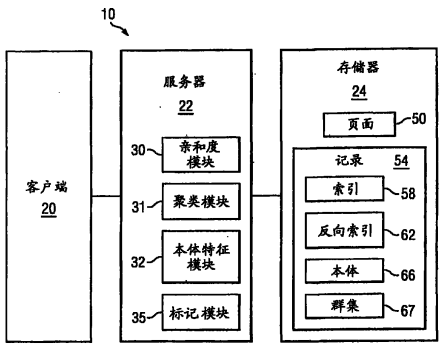
[22] 申请日 2008.10.6
[21] 申请号 200810166181.X
[30] 优先权
[32] 2007.10.5 [33] US [31] 60/977,877
[32] 2008.10.1 [33] US [31] 12/242,984
[71] 申请人 富士通株式会社
地址 日本神奈川县川崎市
[72] 发明人 大卫·马尔维特 贾瓦哈拉·贾殷斯特吉奥斯·斯特吉奥
亚历克斯·吉尔曼
B·托马斯·阿德勒
约翰·J·西多罗维奇
雅尼斯·拉布罗
[74] 专利代理机构 北京三友知识产权代理有限公司
代理人 黄纶伟
权利要求书 5 页 说明书 28 页 附图 5 页

[54] 发明名称

通过分析文档的段落来选择该文档的标签

[57] 摘要

通过分析文档的段落来选择该文档的标签。 在一个实施方式中，为文档指配标签包括访问该文档，其中该文档包括含有词语的文本单元。 针对各文本单元执行以下步骤：文本单元的词语子集被选作候选标签、在所述候选标签之间建立关联性，以及根据所建立的关联性来选择特定候选标签以产生用于该文本单元的候选标签集。 确定各候选标签集的候选标签和其他候选标签集的候选标签之间的关联性。 根据所确定的关联性，为该文档指配至少一个候选标签。



1．一种方法，该方法包括以下步骤：

访问储存在一个或更多个有形介质中的文档，该文档包括含有多个词语的多个文本单元，所述多个词语包括多个关键词；

针对各文本单元执行以下步骤：

在各文本单元的所述关键词之间建立关联性；以及

根据所建立的关联性来选择一个或更多个关键词作为一个或更多个候选标签，以产生所述各文本单元的候选标签集；以及

确定各候选标签集的候选标签和其他候选标签集的候选标签之间的关联性；以及

根据所确定的关联性来为所述文档指配至少一个候选标签。

2．如权利要求 1 所述的方法，该方法还包括：

根据排位技术对所述各文本单元的多个词语进行排位；以及

选择一个或更多个高排位的词语作为所述各文本单元的关键词。

3．如权利要求 1 所述的方法，所述根据关联性选择各文本单元的一个或更多个关键词的步骤还包括：

对所述关键词进行聚类以产生多个群集；以及

指明群集的关键词充分相关。

4．如权利要求 1 所述的方法，所述针对各文本单元执行以下步骤的步骤还包括：

根据排位技术对所述关键词进行排位；以及

选择最高排位的关键词作为根标签。

5．如权利要求 1 所述的方法，所述针对各文本单元执行以下步骤的步骤还包括：

移除与所述其他候选标签不充分相关的一个或更多个候选标签。

6．如权利要求 1 所述的方法，所述确定各候选标签集的候选标签和其他候选标签集的候选标签之间的关联性的步骤还包括：

生成所述各候选标签集的候选标签的概况，所述概况指示所述候选

标签和所述其他候选标签集的候选标签之间的关联性。

7. 如权利要求 1 所述的方法，所述确定各候选标签集的候选标签和其他候选标签集的候选标签之间的关联性的步骤还包括通过以下步骤生成所述各候选标签集的候选标签的概况：

确定包括所述候选标签的候选标签集的数量；以及

根据所述数量生成所述概况。

8. 如权利要求 1 所述的方法，所述确定各候选标签集的候选标签和其他候选标签集的候选标签之间的关联性的步骤还包括通过以下步骤生成所述各候选标签集的候选标签的概况，所述候选标签与权重相关联，所述各候选标签集具有第一根标签：

针对具有第二根标签的各其他候选标签集执行以下步骤以产生多个关联值：

建立在给定所述第一根标签的情况下所述第二根标签的亲和度；以及

通过将所述权重与所述亲和度相乘来计算关联值；以及

根据所述多个关联值来生成所述概况。

9. 如权利要求 1 所述的方法，所述确定各候选标签集的候选标签和其他候选标签集的候选标签之间的关联性的步骤还包括通过以下步骤来生成所述各候选标签集的第一候选标签的概况：

针对各其他候选标签集执行以下步骤：

针对所述其他候选标签集的第二候选标签，建立在给定所述第一候选标签的情况下所述第二候选标签的亲和度，以产生多个亲和度；以及

合并所述亲和度；以及

根据合并后的亲和度来生成所述概况。

10. 如权利要求 1 所述的方法，所述根据所确定的关联性来为所述文档指配至少一个候选标签的步骤还包括：

指配与所述其他候选标签最相关的至少一个候选标签。

11. 一种或更多种编码了软件的计算机可读有形介质，其在被执行

时：

访问储存在一个或更多个有形介质中的文档，该文档包括含有多个词语的多个文本单元，所述多个词语包括多个关键词；

针对各文本单元执行以下步骤：

在各文本单元的所述关键词之间建立关联性；以及

根据所建立的关联性来选择一个或更多个关键词作为一个或更多个候选标签，以产生所述各文本单元的候选标签集；以及

确定各候选标签集的候选标签和其他候选标签集的候选标签之间的关联性；以及

根据所确定的关联性来为所述文档指配至少一个候选标签。

12．如权利要求 11 所述的计算机可读有形介质，其进一步：

根据排位技术对所述各文本单元的多个词语进行排位；以及

选择一个或更多个高排位的词语作为所述各文本单元的关键词。

13．如权利要求 11 所述的计算机可读有形介质，其进一步通过以下步骤来根据关联性选择各文本单元的一个或更多个关键词：

对所述关键词进行聚类以产生多个群集；以及

指明群集的关键词充分相关。

14．如权利要求 11 所述的计算机可读有形介质，其进一步以下步骤来执行针对各文本单元执行以下步骤的所述操作：

根据排位技术对所述关键词进行排位；以及

选择最高排位的关键词作为根标签。

15．如权利要求 11 所述的计算机可读有形介质，其进一步通过以下步骤来执行针对各文本单元执行以下步骤的所述操作：

移除与所述其他候选标签不充分相关的一个或更多个候选标签。

16．如权利要求 11 所述的计算机可读有形介质，其进一步通过以下步骤来执行确定各候选标签集的候选标签和其他候选标签集的候选标签之间的关联性的所述操作：

生成所述各候选标签集的候选标签的概况，所述概况指示所述候选标签和所述其他候选标签集的候选标签之间的关联性。

17. 如权利要求 11 所述的计算机可读有形介质，其进一步通过以下步骤来执行通过生成所述各候选标签集的候选标签的概况来确定各候选标签集的候选标签和其他候选标签集的候选标签之间的关联性的操作：

确定包括所述候选标签的候选标签集的数量；以及

根据所述数量生成所述概况。

18. 如权利要求 11 所述的计算机可读有形介质，其进一步通过以下步骤来执行通过生成所述各候选标签集的候选标签的概况来确定各候选标签集的候选标签和其他候选标签集的候选标签之间的关联性的操作，所述候选标签与权重相关联，所述各候选标签集具有第一根标签，所述步骤包括：

针对具有第二根标签的各其他候选标签集执行以下步骤以产生多个关联值：

建立在给定所述第一根标签的情况下所述第二根标签的亲和度；以及

通过将所述权重与所述亲和度相乘来计算关联值；以及

根据所述多个关联值来生成所述概况。

19. 如权利要求 11 所述的计算机可读有形介质，其进一步通过以下步骤来执行通过生成所述各候选标签集的第一候选标签的概况来确定各候选标签集的候选标签和其他候选标签集的候选标签之间的关联性的操作：

针对各其他候选标签集执行以下步骤：

针对所述其他候选标签集的第二候选标签，建立在给定所述第一候选标签的情况下所述第二候选标签的亲和度，以产生多个亲和度；以及

合并所述亲和度；以及

根据合并后的亲和度来生成所述概况。

20. 如权利要求 11 所述的计算机可读有形介质，其进一步通过以下步骤来根据所确定的关联性向所述文档指配至少一个候选标签：

指配与所述其他候选标签最相关的至少一个候选标签。

21．一种系统，该系统包括：

访问储存在一个或更多个有形介质中的文档的单元，该文档包括含有多个词语的多个文本单元，所述多个词语包括多个关键词；

针对各文本单元执行以下步骤的单元：

在各文本单元的所述关键词之间建立关联性；以及

根据所建立的关联性来选择一个或更多个关键词作为一个或更多个候选标签，以产生所述各文本单元的候选标签集；以及

确定各候选标签集的候选标签和其他候选标签集的候选标签之间的关联性的单元；以及

根据所确定的关联性来为所述文档指配至少一个候选标签的单元。

通过分析文档的段落来选择该文档的标签

技术领域

本发明总体上涉及词法（lexigraphical）分析，更具体地说，涉及通过分析文档的段落来选择该文档的标签。

背景技术

本申请要求 David Marvit 等人于 2007 年 10 月 5 日提交的、发明名称为“Tagging Based on Paragraph and Category Analysis（基于段落和类别分析进行标记）”的美国临时申请 60/977,877 的优先权。

尽管数据语料库（corpus）可以保存大量信息，但是要找出相关信息可能仍然很困难。可以对文档进行标记以便于搜索相关信息。然而，在特定情形中，已知的文档标记技术在定位信息方面不是很有效。类似地，已知的搜索技术在定位相关信息方面也不是很有效。

附图说明

图 1 例示了选择文档的标签的系统的一个实施方式；

图 2 例示了可以与图 1 的系统一起使用的亲和度（affinity）模块的一个实施方式；

图 3 例示了记录基本亲和度的亲和度矩阵的实施例；

图 4 例示了记录有向亲和度的亲和度矩阵的实施例；

图 5 例示了记录平均亲和度的亲和度矩阵的实施例；

图 6 例示了亲和度图的实施例；

图 7 例示了可以与图 1 的系统一起使用的聚类模块的一个实施方式；

图 8 例示了可以与图 1 的系统一起使用的本体（ontology）特征模块的一个实施方式；

图 9 例示了可以与图 1 的系统一起使用的标记模块的一个实施方式；

图 10 例示了用于确定话题的统计分布的方法的实施例；

图 11 例示了用于通过分析文档的段落来向该文档指配标签的方法的实施例；以及

图 12 例示了用于响应于所选择的标签来指配标签的方法的实施例。

具体实施方式

概述

在一个实施方式中，向文档指配标签包括访问该文档，其中该文档包括含有词语（word）的文本单元。针对各文本单元执行以下步骤：文本单元的词语子集被选作候选标签、在这些候选标签中建立关联性，以及根据所建立的关联性的程度来选择特定候选标签以产生该文本单元的的候选标签集。确定各候选标签集的候选标签和其他候选标签集的候选标签之间的关联性。根据确定的关联性将至少一个候选标签指配给文档。例如，可以通过选择与文档最相关的特定数量的候选标签并将这些候选标签指配给该文档来指配标签集。

示例性实施方式

图 1 例示了选择文档的标签的系统 10 的一个实施方式。在具体实施方式中，系统 10 通过分析文档的文本单元（例如，段落）来选择标签。在这些实施方式中，系统 10 根据文本单元的词语来识别用于各文本单元的候选标签集。系统 10 接着对不同候选标签集中的候选标签的关联性进行比较，并根据该关联性来选择该文档的标签。

在具体实施方式中，对于给定的词语子集和词典 D，可以基于特定反向索引（inverted index）II 计算有向亲和度，其中索引 II 包括例如针对词句 w_i 和词句 w_j 的条目 $I(w_i)$和 $I(w_j)$。一般来说，反向索引是储存从词条（term）到其位置（即，呈现词条的同现语境）的映射的索引数据结构。对于 D 中的每对词语 w_i 和 w_j，DA(i,j)可以被定义为 II 中的条目 $I(w_i)$ 和 $I(w_j)$ 的合取（conjunction）除以 $I(w_i)$ 的数目值。一般来说，DA(i,j) 无需等于 DA(j,i)。可以以任何合适的方式（例如，行方式）储存结果，其中储存 D(1,i)，接着存储 D(2,j)，以此类推。对于每行 i，可以存储$|I(w_i)|$，

接着是与 w_j 的合取的基数（cardinality）。

在具体实施方式中，可以按三个阶段计算有向亲和度。在该实施方式中，每一个字典词条都被指配了唯一的整数识别符。反向索引的条目对应于该整数识别符。在阶段 0 中，读取对应于 *D* 的 *II* 条目。对于参数（*s*，*o*）来说，仅保留具有 *ks* + *o* 形式的要素识别符。值 ks + o 限定要检查的 *II* 条目的子集。按这种方式，可以并行计算有向亲和度。作为一个实施例，根据参数 s 的结果，o（1，0）等于根据参数（3，0）、（3，1）、（3，2）合并计算获得的结果。该步骤允许计算用于很大反向索引的 DA 表。

在阶段 1 中，仅针对 *DA*（*i*，*j*）以行方式计算合取。在阶段 2 中，读取计算出的上三角 UT DA 矩阵。据此，获取下三角部分，作为 UT 的转置。在具体实施方式中，可以将相同维的多个 DA 阵列合并成单一阵列。可以将大的 II 上的 DA 阵列计算为具有参数(s，i)的 $sum_{i=0..(s-1)}$ DA。可以利用计算出的合取来存储附加信息，以便计算有向亲和度。在具体情况下，可以存储 *II* 条目的基数。

在具体实施方式中，可以按行方式存储 DA，这样 AA 条目的计算可以与 DA 条目的计算并行进行。具体来说，可以通过在从磁盘中读取 DA 时对 DA 的行进行求和并且最后用词典条目的数量将其归一化而生成 AA。

在所示实施方式中，系统 10 包括：客户端 20、服务器 22 以及存储器 24。客户端 20 允许用户与服务器 22 通信，以使生成语言的本体。客户端 20 可以向服务器 22 发送用户输入，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 管理用于生成语言的本体的应用。存储器 24 存储服务器系统 24 使用的数据。

在所示实施方式中，存储器 24 存储有页面 50 和记录 54。页面 50（或文档或同现语境）可以指词语的集合。页面 50 的示例包括：文档的一个或更多个页面、一个或更多个文档、一本或更多本书、一个或更多个网页、信件（例如，电子邮件或即时消息）和/或其它词语的集合。可以由页面识别符来识别页面 50。页面 50 可以以电子方式存储在一种或更

多种有形计算机可读介质中。页面 50 可以与任何合适的内容相关联，例如，文本（如字符、词语和/或数字）、图像（如图形、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。在具体实施方式中，一组页面 50 可以属于一个语料库。语料库可以与特定主题、团体、组织或其它实体相关联。

记录 54 描述页面 50。在该实施方式中，记录 54 包括：索引 58、反向索引 62、本体 66、以及群集 67。索引 58 包括索引列表，其中，页面 50 的索引列表指示页面 50 的词语。反向索引 62 包括反向索引列表，其中，词语（或词语集）的反向索引列表指示包括该词语（或词语集）的页面 50。在一个实施例中，列表 W_i 包括含有词语 w_i 的页面 50 的页面识别符。列表 W_i & W_j 包括含有词语 w_i 和 w_j 两者的合取（conjunction）页面 50 的页面识别符。列表 $W_i + W_j$ 包括含有词语 w_i 和 w_j 中的任一个的析取（disjunction）页面 50 的页面识别符。$P(W_i)$ 是 W_i 的页面 50 的数量，即，包括词语 w_i 的页面 50 的数量。

在一个实施方式中，可以将列表（如索引列表或反向索引列表）存储为二值判决图（BDD：Binary decision diagram）。在一个实施例中，集合 W_i 的二值判决图 BDD（W_i）代表具有词语 w_i 的页面 50。BDD(W_i)的满足指定计数（satisfying assignment count）Satisf(BDD(W_i))得到具有词语 w_i 的页面 50 的数量 $P(W_i)$：

P（W_i）= Satisf（BDD（W_i））

相应地，

P（W_i & W_j）= Satisf（BDD（W_i）与 BDD（W_j））

P（$W_i + W_j$）= Satisf（BDD（W_i）或 BDD（W_j））

本体 66 表示语言的词语和这些词语之间的关系。在一个实施方式中，本体 66 表示词语之间的亲和度。在所示实施例中，本体 66 包括亲和度矩阵和亲和度图。参照图 3 到图 5，对亲和度矩阵的实施例进行描述。参照图 6，对亲和度图的实施例进行描述。群集 67 记录了彼此相关的词语的群集。参照图 7，对群集进行更详细描述。

在所示实施方式中，服务器 22 包括：亲和度模块 30、聚类模块 31、

本体特征模块 32、以及标记模块 35。亲和度模块 30 可以计算针对词语对的亲和度，在亲和度矩阵中记录该亲和度，和/或报告该亲和度矩阵。亲和度模块 30 也可以生成亲和度图。参照图 2，对亲和度模块 30 进行更详细描述。

在具体实施方式中，聚类模块 31 可以通过识别数据集中的相关要素的群集来发现该数据集中的模式。在具体实施方式中，聚类模块 31 可以识别一组词语的群集（例如，一种语言或一组页面 50）。一般来说，群集的词语彼此高度相关，但与该群集以外的词语不相关。词语的群集可以指定该组词语的主题（或话题）。在具体实施方式中，聚类模块 31 根据词语之间的亲和度来识别相关词语的群集。在该实施方式中，群集的词语彼此高度相关，但与该群集以外的词语不相关。参照图 7，对聚类模块 31 进行更详细地描述。

在具体实施方式中，本体特征模块 32 可以确定一个或更多个词语构成的组（例如，具体词语或包括词语的文档）的一个或更多个本体特征，并接着可以在多种情形中的任一种中应用该本体特征。本体特征是将词语集放置在一种语言的本体空间中的词语集特征。本体特征的实施例包括深度和专度。在具体实施方式中，深度可以指示词语集的文本复杂性（sophistication）。较深的词语集可能较技术化并且专业化，而较浅的词语集可能较通用。在具体实施方式中，词语集的专度与词语集的主题的数量有关。较专的词语集可能具有较少的主题，而不太专的词语集可能具有较多的主题。

本体特征模块 32 可以在任何合适的情形下应用本体特征。合适情形的实施例包括根据本体特征进行的搜索、分类或选择文档；报告文档的本体特征；以及确定一个或更多个用户的文档的本体特征。参照图 8 对本体特征模块 32 进行更详细地描述。

在具体实施方式中，标记模块 35 可以选择标签来标记文档。可以以任何合适的方式选择标签。在具体实施方式中，标记模块 35 将话题建模为话题的相关词语的统计分布。标记模块 35 使用该统计分布来识别文档的所选词语具有最高出现概率的话题，并且标记模块 35 根据识别的话题

来选择该文档的标签。在其他实施方式中，标记模块 35 识别文档的段落的候选标签。标记模块 35 确定这些候选标签与文档的其他候选标签的关联性，并根据该确定来选择该文档的标签。在再一实施方式中，标记模块 35 推荐文档的标签。可以基于与用户或计算机输入或选择的目标标签的亲和度（例如，有向和/或差分亲和度）来推荐标签。一旦选择了最终标签，标记器 314 就可以向文档指配选择的标签。参照图 9 对标记模块 35 进行更详细地描述。

系统 10 的组件可以包括接口、逻辑、存储器和/或其他合适元件。接口接收输入、发送输出、对输入和/或输出进行处理，和/或执行其他合适操作。接口可以包括硬件和/或软件。

逻辑执行对组件的操作，例如，执行指令以根据输入来生成输出。逻辑可以包括硬件、软件和/或其他逻辑。逻辑可以以一种或更多种有形介质中进行编码，并且可以当计算机执行该逻辑时执行操作。诸如处理器的特定逻辑可以管理组件的操作。处理器的实施例包括一个或更多个计算机、一个或更多个微处理器、一个或更多个应用程序，和/或其他逻辑。

存储器存储信息。存储器可以包括一个或更多个有形的、计算机可读的和/或计算机可执行的存储介质。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），）、海量存储介质（例如，硬盘）、可移动存储介质（光盘（CD）或数字视频光盘（DVD））、数据库和/或网络存储器（例如，服务器）以及/或其他计算机可读介质。

在不脱离本发明的范围的情况下，可以对系统 10 进行改进、添加或省略。系统 10 的组件可以是集成或分离的。此外，可以通过更多、更少或其他组件来执行系统 10 的操作。例如，可以通过一个组件执行生成器 42 和生成器 46 的操作，或者可以通过一个以上的组件来执行亲和度计算器 34 的操作。另外，可以使用包括软件、硬件的任何合适逻辑和/或其他逻辑来执行系统 10 的操作。如本文档中所使用的，“每个”指集合中各成员，或集合的子集中的各成员。

在不脱离本发明的范围的情况下，可以对矩阵的实施例进行改进、

添加或省略。矩阵可以包括更多、更少或其他值。另外，可以以任何合适的顺序来排列矩阵的值。

图 2 例示了可以与图 1 的系统 10 一起使用的亲和度模块 30 的一个实施方式。亲和度模块 30 可以计算词语对的亲和度、将该亲和度记录在亲和度矩阵中、和/或报告该亲和度矩阵。亲和度模块 30 还可以生成亲和度图。

在所示的实施方式中，亲和度模块 30 包括亲和度计算器 34、本体生成器 38 以及词语推荐器 48。亲和度计算器 34 计算针对词语 w_i 或针对包括第一词语 w_i 和第二词语 w_j 的词语对的任何合适类型的亲和度。亲和度的实施例包括基本亲和度、有向亲和度、平均亲和度、差分亲和度和/或其他亲和度。

这一个实施方式中，词语推荐器 48 接收词根，并且识别其与词根的亲和度大于阈值亲和度的词语。阈值亲和度可以具有任何合适值，例如大于或等于 0.25、0.5、0.75 或 0.95。阈值亲和度可以被预编程的或者由用户指定。

可以根据包括词语 w_i 和/或 w_j 的页面 50 的量（例如，数量）来计算基本亲和度。合取页面量是指既包括词语 w_i 又包括词语 w_j 的页面 50 的量，而析取页面量是指包括词语 w_i 或词语 w_j 中的任一个的页面 50 的量。可以通过合取页面量除以析取页面量来给出基本亲和度。在一个实施例中，合取页面数量指示包括词语 w_i 和词语 w_j 的页面的数量，而析取页面数量指示包括词语 w_i 或词语 w_j 的页面的数量。可以通过合取页面的数量除以析取页面的数量来给出基本亲和度。

$$\text{亲和度}(w_i, w_j) = P(W_i \,\&\, W_j) / P(W_i + W_j)$$

图 3 例示了记录基本亲和度的亲和度矩阵 110 的实施例。在所示的实施例中，亲和度矩阵 110 记录词语 w_1.....w_5 的逐对亲和度。根据亲和度矩阵 110，词语 w_0 和 w_1 之间的亲和度为 0.003，词语 w_0 和 w_2 之间的亲和度为 0.005 等。

返回参照图 1，亲和度组包括彼此具有高亲和度的词语对，并且可以被用来针对页面内容来捕捉词语 w_1 和 w_2 之间的关系。高亲和度可以

被指定为高于亲和度组阈值的亲和度。阈值可以被设置为任何合适的值（例如，大于或等于 0.50、0.60、0.75、0.90 或 0.95）。一个词语可以属于一个以上的亲和度组。在一个实施方式中，亲和度组可以表示为 BDD。BDD 的指针与该组的每个词语一起被储存在反向索引 62 中。

有向亲和度可以被用来测量词语 w_i 对于词语 w_j 的重要性。亲和度计算器 34 根据包括词语 w_i 和词语 w_j 的页面 50 的数量（例如，数目）来计算在给定词语 w_j 的情况下词语 w_i 的有向亲和度。词语 w_j 页面数量是指包括词语 w_i 的页面 50 的数量。可以通过合取页面数量除以词语 w_j 页面数量来提供在给定词语 w_j 的情况下 w_i 的有向亲和度。例如，词语 w_j 页面数量指示包括词语 w_i 的页面 50 的数量。可由合取页面 50 的数量除以词语 w_i 的页面 50 的数量来提供在给定词语 w_j 的情况下 w_i 的有向亲和度：

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

DAffinity(w_i,w_j)和 DAffinity(w_j,w_i)不同。词语 w_i 和 w_j 之间的高有向亲和度 DAffinity(w_i,w_j)表示页面 50 在包括词语 w_j 的情况下包括词语 w_i 的较高概率。在一个实施例中，页面[1 2 3 4 5 6]包括词语 w_i，而页[4 2]包括词语 w_j。包括词语 w_j 的页面也包括词语 w_i，因此从词语 w_j 的角度，词语 w_i 具有高重要性。包括词语 w_i 的页面中仅有三分之一的页面也包括词语 w_j，因此从词语 w_i 的角度，词语 w_j 具有较低的重要性。

图 4 例示了记录针对词语 $w_0, \ldots, w_5$ 的有向亲和度的亲和度矩阵 120 的实施例。在该实施例中，词语 124 是 A 词语，而词语 128 是 B 词语。矩阵 120 的各行记录了在给定 A 词语的情况下 B 词语的亲和度，而亲和度矩阵 120 的各列记录了在给定 A 词语的情况下 B 词语的亲和度。

返回参照图 1，针对其他词语 w_j 计算词语 w_i 的平均亲和度。在一个实施方式中，平均亲和度可以是词语 w_i 和各其他词语 w_j 之间的亲和度的平均值。词语 w_i 在 N 个词语中的平均亲和度可以通过下式给出：

$$\text{AveAff}(w_i) = \frac{1}{N}\sum_{j=1}^{N} P(w_i \mid w_j)$$

图 5 例示了记录平均亲和度的亲和度矩阵 140 的实施例。行 142 记录了针对词语 1 到词语 50,000 的基本亲和度。行 144 记录了词语 1 到词语 50,000 的平均亲和度。

返回参照图 1，词语的平均亲和度可以指示词语的深度。具有较低平均亲和度的词语可以被认为是较深的词语，而具有较高平均亲和度的词语可以被认为是较浅的词语。较深的词语倾向于更技术化、专属并且更准确。具有较高百分比的较深词语的页面 50 被认为是较深页面，而具有较低百分比的较深词语的页面 50 可以被认为是较浅页面。在一个实施方式中，用户可以指定要检索的词语和/或页面 50 的深度。

页面 50 的较深词语可以形成高度相关词语的一个或更多个群集。群集可以表示共同的思想或主题。页面 50 的主题的数量可以指示页面 50 的专度。具有较少主题的页面 50 可以被认为较专属页面，而具有较多主题的页面 50 可以被认为欠专属（less specific）页面。

词语 w_i 对于词语 w_j 的差分亲和度是词语 w_i 和 w_j 之间的有向亲和度减去词语 w_j 对于所有其他词语的平均亲和度。差分亲和度可以被表达为：

$$\mathrm{DiffAff}(w_i, w_j) = \mathrm{DAffinity}(w_i, w_j) - \mathrm{AveAff}(w_j)$$

差分亲和度排除了由词语 w_j 在页面 50 中出现的一般趋势而造成的偏差（bias）。在具体情况下，假定页面包括词语 w_j 时，差分亲和度可提供该页面包括词语 w_i 的概率的更精确指示。

差分亲和度可以被用于各种应用中。在一个实施例中，人名之间的差分亲和度可以被用来研究社会网络。在另一实施例中，语素之间的差分亲和度可以被用来研究自然语言处理。在另一实施例中，产品之间的差分亲和度可以被用来研究市场策略。

亲和度计算器 34 可以使用任何合适的技术来搜索反向索引列表以计算亲和度。例如，为了识别既包括词语 w_i 又包括 w_j 的页面，亲和度计算器 34 可以针对公共要素（即，公共页面识别符）搜索词语 w_i 的列表 W_i 和词语 w_j 的列表 W_j。

在具体实施方式中，本体生成器 38 生成语言的本体 66（例如，亲和度矩阵或亲和度图）。可以根据诸如基本亲和度、有向亲和度、平均亲和度、差分亲和度和/或其他亲和度中的任何合适亲和度来生成本体。可以以任何合适的方式根据从语言中选择的词语来生成本体 66。例如，可以选择来自语言的公用部分的词语或者与一个或更多个具体主题区域相

关的词语。

在所示的实施方式中，本体生成器 38 包括亲和度矩阵生成器 42 和亲和度图生成器 46。亲和度矩阵生成器 42 生成记录词语之间的亲和度的亲和度矩阵。亲和度图生成器 46 生成表示词语之间的亲和度的亲和度图。在亲和度图中，节点表示词语，而节点之间的有向边的权重表示由节点所表示的词语之间的亲和度。亲和度图可以具有任何合适的维数。

图 6 例示了亲和度图 150 的实施例。亲和度图 150 包括节点 154 和链路 158。节点 154 表示词语。在本实施例中，节点 154a 表示词语“binary（二进制）”。节点 154 之间的有向边的权重表示由节点 154 表示的词语之间的亲和度。例如，较大的权重表示较大的亲和度。节点之间的链路 158 指示由节点 154 表示的词语之间的亲和度高于亲和度阈值。亲和度阈值可以具有任何合适的值（例如，大于或等于 0.25、0.5、0.75 或 0.95）。

图 7 例示了可以与图 1 的系统 10 一起使用的聚类模块 31 的一个实施方式。在具体实施方式中，聚类模块 31 通过识别数据集中的相关要素的群集来发现数据集中的模式。在具体实施方式中，聚类模块 31 可以识别词语集的群集（例如，语言或页面 50 的集合）。一般来说，群集的词语彼此高度相关，而与群集以外的词语不高度相关。词语的群集可以指定词语集的主题（或话题）。

在具体实施方式中，聚类模块 31 根据词语之间的亲和度来识别相关词语的群集。在这些实施方式中，群集的词语彼此高度相关，但是与群集以外的词语不高度相关。在一个实施方式中，如果词语充分相关，则可以认为它们高度相关。如果词语满足一个或更多个亲和度标准（例如，阈值），则词语充分相关，下面提供了其实施例。

可以使用任何合适的亲和度来识别群集。在具体实施方式中，聚类模块 31 使用有向亲和度。一词语相对于其他词语的有向亲和度表征了词语的共现。群集包括具有类似共现的词语。在具体实施方式中，聚类模块 31 使用差分亲和度。差分亲和度旨在去除词语在页 50 中出现的一般趋势所导致的偏差。

在所示的实施方式中，聚类模块 31 包括聚类引擎 210 和聚类分析器

214。聚类引擎 210 根据亲和度来识别词语的群集，并且群集分析器 214 应用亲和度聚类以分析各种情形。

聚类引擎 210 可以根据亲和度以任何合适的方式来识别词语的群集。提出了用于识别群集的方法的三种实施例：根据词语集构建群集、将词语分类成群集，以及比较词语的亲和度矢量。在一个实施方式中，聚类引擎 210 根据词语集构建群集。在一个实施例中，聚类引擎 210 根据具有亲和度*Aff(w_i, w_j)的词语{w_i}的集 W 建立群集 S。亲和度值*Aff(w_i, w_j)代表词语 w_i 相对于 w_j 的任意合适类型的亲和度，诸如有向亲和度 DAffinity(w_i, w_j)或差分亲和度 DiffAff (w_i, w_j)。这里提供的亲和度值的某些实施例可以被认为是归一化值。在该实施例中，Aff_{for} (w_i, w_j)代表前向亲和度，且 Aff_{back} (w_j, w_i)代表后向亲和度。

在本实施例中，群集 S 以词根 W_q 开始。当前词语 w_x 表示在当前迭代处群集 S 的正与来自集合 W 的词语比较的词语。最初，将当前词语 w_x 设置为词根 w_q。

在迭代期间，当前词语 w_x 被设置为群集 S 的词语。根据它们与当前词语 w_x 的前向亲和度 $Aff_{for}(w_i, w_x)$来对集合 W 的词语 w_i 进行分类。从分类集合 W 的起点开始，识别满足亲和度标准的候选词语 w_c。亲和度标准可以包括对于当前词语 w_x 的前向亲和度标准：

$$Aff_{for}(w_c, w_x) > Th_{cf}$$

和对于词根 w_q 的后向亲和度标准：

$$Aff_{back}(w_q, w_c) > Th_{cb}$$

其中，Th_{cf} 表示用于候选词语的前向阈值，而 Th_{cb} 表示用于候选词语的后向阈值。候选词语（w_c）的有序集合的第一词语被添加到群集 S，添加的词语的数量由参数 $Size_c$ 给出。阈值 Th_{cf} 和 Th_{cb} 可以是具有从最小值到最大值的任何合适值的浮点参数。在特定实施例中，可以根据实际亲和度的级别有序列表来确定 Th_{cf} 和 Th_{cb} 的合适值。例如，可以使用列表中第 200 个值。参数 $Size_c$ 可以是具有任何合适值的整数参数。合适值的实施例包括默认值 1、2、3 或 4。在具体实施方式中，这些参数在具体迭代中可以不同。

可以执行任何合适数量的迭代。在一个实施例中，可以在开始执行该方法之前指定迭代数量。在另一实施例中，可以在方法执行期间计算该数量。例如，可以根据群集 S 的大小的增长率来计算该数量。

在另一实施方式中，聚类引擎 210 通过将词语集的词语分类成群集来识别群集。在一个实施例中，根据亲和度*Aff(w_i, w_j)（例如，差分亲和度或有向亲和度）来对集合 W 的词语（w_i）进行分类。在另一实施例中，根据词语 w_i 与不同词语集 Q 的各成员的亲和度的累积函数（例如，求和）来分类词语（w_i）。可以以任何合适方式选择集合 W。例如，集合 W 可以是与查询最相关的 X 个词语，其中 X 可以具有任何合适值（例如，从 10 到 100、100 到 200 或者等于或大于 200 的值）。

在本实施例中，群集初始为空。来自集合 W 的第一词语 w_i 被放置在群集中。在每次迭代中，从集合 W 选择当前词语 w_x。如果*Aff(w_x, w_f)满足亲和度阈值 Th 给出的亲和度标准，则当前词语 w_x 被放置在群集中，其中 w_f 表示该群集中放置的第一词语。阈值 Th 可以具有任何合适值（0.1 至 0.5 范围的值（最小值为 0.0 和最大值为 1.0））。如果*Aff(w_x, w_f)不满足阈值 Th，则当前词语 w_x 被放置在空群集中。针对集合 W 的各词语重复这些迭代。

在处理了集合 W 的词语之后，可以消除小群集。例如，可以消除具有少于 Y 个词语的群集。Y 可以是任何合适值（例如范围在 3 到 5、5 到 10、10 到 25、25 到 50 或者大于等于 50 的范围中的值）。

如果群集的数量不在令人满意的范围内，则可以利用针对群集布置生成更严格或更宽松的标准的不同阈值 Th 来重复该处理。可以通过具有任何合适值的群集数量最小值和群集数量最大值给出令人满意的范围。合适值的实施例包括最小值在 1 到 5、5 到 10 或者 10 或大于或等于 10 范围的值，以及最大值在 10 到 15、15 到 20 或者 20 或大于或等于 20 的范围中的值。可以增加阈值 Th 的值，以增加群集的数量，并且可以减小阈值 Th 的值以减小群集的数量。

在另一实施方式中，聚类引擎 210 通过比较词语的亲和度矢量来识别群集。在具体实施方式中，亲和度矩阵的行和列可以产生亲和度矢量<

w_i, *Aff(w_i, w_1), ..., *Aff(w_i, w_j), ..., *Aff(w_i, w_n)>,该亲和度矢量表示词语 w_i 相对于词语 w_j（j=1，...，n）的亲和度。亲和度值*Aff(w_i, w_j)表示词语 w_i 相对于词语 w_j 的任何合适类型的亲和度（例如，有向亲和度或差分亲和度）。

在具体实施方式中，具有类似亲和度值的亲和度矢量可以表示一个群集。仅出于描述性目的，可以将亲和度矢量看作是亲和度空间中的词语的亲和度的坐标。即，每个亲和度值*Aff(w_i, w_j)可以被认为是针对具体维的坐标。具有类似亲和度值的亲和度矢量表示与这些矢量相关联的词语在亲和度空间中彼此接近。即，这些矢量指示这些词语具有与其他词语类似的亲和度关系，并因此可以适于同一群集中的成员关系。

如果一个亲和度矢量近似于由合适距离函数确定的另一亲和度矢量，则这些亲和度矢量类似。可以通过亲和度矢量上例如将该距离函数定义为针对给定大小的矢量的标准欧几里得距离，或者定义为给定大小的矢量的余弦。该距离函数还可以通过聚类引擎 210 或者由用户指定。

在具体实施方式中，聚类引擎 210 应用聚类算法来识别具有彼此近似的值的亲和度矢量。聚类算法的实施例包括直接、重复二等分（repeated bisection）、聚合（agglomerative）、偏置聚合、和/或其它合适算法。在一个实施例中，聚类引擎 210 可以包括诸如 CLUTO 的聚类软件。

群集分析器 214 可以在任何合适的应用中使用亲和度聚类来进行分析。在一个实施方式中，群集分析器 214 可以使用亲和度聚类来归类页面 50。类别可以与群集识别符或者群集的一个或更多个成员相关联。在一个实施例中，页面 50 的群集可以被识别，并且接着可以根据该群集对页面 50 进行归类。在另一实施例中，可以选择页面 50 的重要词语，并接着对包括这些词语的群集进行定位。然后可以根据定位后的群集对页面 50 归类。

在一个实施方式中，群集分析器 214 可以使用亲和度聚类来分析页面 50 的语料库。语料库可以与具体主题、一个或更多个个体的集合（community）、组织或其他实体相关联。在一个实施例中，群集分析器 214 可以识别语料库的群集，并根据该群集确定语料库的语料库字符。语

料库字符可以指示与和该语料库相关联的实体相关的词语。如果一个或更多个页面 50 具有语料库字符的群集，则页面 50 与该实体相关。

在一个实施方式中，群集分析器 214 可以使用亲和度聚类来搜索查询歧义消除和查询扩展。在本实施方式中，群集分析器 214 识别包括给定搜索查询的搜索词条的群集。群集提供与给定搜索查询相关的可替换词语和/或类别。在一个实施例中，来自群集的词语可以被报告给搜索者，以帮助下一搜索查询。在另一实施例中，群集分析器 214 可以从群集中选择词语，并自动形成一个或更多个新的搜索查询。群集分析器 214 可以串行或并行运行新的查询。

在一个实施方式中，群集分析器 214 可以使用亲和度聚类来研究社会网络。在一个实施例中，页面 50 可以提供对社会网络的深刻见解。这些页面的实施例包括信件（例如信件、电子邮件以及即时消息）、便笺、文章以及会议记录。这些页面 50 可以包括含有社会网络的人员的用户识别符（例如，姓名）的词语。可以识别姓名的群集，以分析网络的人员之间的关系。在一个实施例中，差分亲和度聚类可用于过滤大多数页 50 中的出现的名字，而不提供诸如系统管理员的名字之类的信息。

在具体实施方式中，群集分析器 214 可以通过组合和/或比较数据集的群集来分析数据集。在一个实施方式中，对重叠的数据集的群集进行比较。来自一个数据集的群集可以被映射到另一数据集的群集上，这样可以提供对这些数据集之间的关系的深刻见解。例如，该数据集可以来自对同事组的文档的分析和来自对该组的社会网络研究。可以将社会网络群集映射至文档主题群集，来分析社会网络与该主题之间的关系。

图 8 例示了本体特征模块 32 的一个实施方式。本体特征模块 32 可以确定一个或更多个词语（例如，具体词语或包括词语的文档）的集合的一个或更多个本体特征，并且接着可以在任何不同情形中应用该本体特征。一个或更多个词语的集合可以包括文档的必要词条。如果与词条 t 相关的前 k 个词条中的至少一个也呈现在该文档中，则词条 t 可以是必要词条。否则，该词条对于该文档可能不是必不可少的。

本体特征是沿一个或更多个特征轴表征文档的可量化测量，所述特

征轴可以在给定区域中从语义上对该文档与其他文档进行区分。例如，文档的深度可以针对它的可理解性来区分文档、文档的专度可以针对它的关注点来区分文档，而文档的主题可以针对其关注的话题范围来区分文档。可以以任何合适方式定义本体特征。例如，计算机语言中的独立算法可以被用来表征文档的可读性或深度。

在所示的实施方式中，本体特征模块 32 包括深度引擎 230、主题引擎 240、专度引擎 244 以及本体特征（OF：ontology feature）应用引擎 250。深度引擎 230 可以确定一个或更多个词语(例如，具体词语或包括词语的文档)的深度。一般来说，深度可以指示词语的文本复杂性。越深的词语可以是更加技术化的并且更专业的，而越浅的词语可以是更通用的。在具体实施方式中，深度模块 32 可以计算文档的词语的深度，并接着根据词语的深度来计算文档的深度。在具体实施方式中，深度引擎 230 可以为文档和/或词语指配深度值和/或深度级别。越深的文档或词语可以被指配越高的深度值或级别，而越浅的文档或词语可以被指配越低的深度值或级别。

深度引擎 230 可以以任何合适的方式计算词语深度。在具体实施方式中，深度引擎 230 根据平均亲和度来计算词语深度。在这些实施方式中，词语的深度是词语的平均亲和度的函数。较深的词语可以具有较低的平均亲和度，而较浅的词语可以具有较高的平均亲和度。在具体实施例中，深度引擎 230 可以通过根据它们的平均亲和度对词语进行排位来计算词语的深度。对具有较低平均亲和度的词语给予较高的深度级别，而对具有较高平均亲和度的词语给予越低的深度级别。

在具体实施方式中，深度引擎 230 可以使用聚类分析来计算词语深度。在这些实施方式中，群集的词语相互高度相关，而与群集以外的词语较低相关。可以根据能够作为深度指示符的亲和度来测量群集空间中的距离。在具体实施方式中，属于较少数群集或者属于较小群集和/或离其他群集较远的群集的词语可以被认为较深，而属于较多数群集或者属于较大群集和/或离其他群集较近的群集的词语被认为较浅。

在其他具体实施方式中，深度引擎 230 可以通过向亲和度图 150 应

用链路分析来计算词语深度。可以通过任何合适的链路分析算法（例如，PAGERANK）来执行该链路分析。仅出于描述性目的，图 6 的亲和度图 150 可以被用来计算词语深度。亲和度图 150 包括节点 154 和链路 158。节点 154 表示词语。节点 154 之间的链路 158 指示由节点 154 表示的词语之间的亲和度高于亲和度阈值，即，这些词语令人满意地相关。

在具体实施方式中，深度引擎 230 计算节点 154 的通用性。较通用的节点 154 可以表示较浅的词语，而较不通用的节点 154 可以表示较深的词语。从第一节点 154 到第二节点 154 的链路 136 被认为第一节点 154 对第二节点 154 的通用性选票。另外，来自较通用节点 154 的选票（vote）可以具有比来自较不通用节点 154 的选票更大的权重。此外，第一节点 154 与第二节点 154 的亲和度加权了该选票。深度引擎 230 根据节点 154 的加权后的选票来计算节点 154 的通用性。较不通用的词语被认为较深词语，而较通用的词语可以被认为较浅词语。

深度引擎 230 可以以任何合适方式来计算文档深度。在具体实施方式中，深度引擎 230 根据文档中的至少一个、一些或所有词语的深度来计算文档的深度。在具体实施方式中，词语深度根据平均亲和度给出，因此可以根据文档的词语的平均亲和度来计算文档深度。例如，文档的浅度可以是文档的词语的平均亲和度的平均值（即，文档中各词语的平均亲和度的和除以用文档中的词语的总数）。接着，文档的深度可以被计算为文档的浅度的倒数。

在具体实施方式中，可以根据文档的所选词语集的平均深度来计算深度。所选的集合可以包括文档必要词语（例如，前（最深）X%的词语，其中 X 可以小于 10、10 到 20、20 到 30、30 到 40、40 到 50、50 到 60、60 到 70，或者大于 10）。所选的集合可以排除 P%的标准语法词语和/或 Q%的停顿词（stop word），其中 P 和 Q 具有任何合适值（例如小于 10、10 到 20、20 到 30、30 到 40、40 到 50、50 到 60、60 到 70，或者大于 10）。

在具体实施方式中，深度引擎 230 根据文档中词语深度的分布来计算文档的深度。在具体实施方式中，较深的文档可以具有较高百分比的

较深词语。

在具体实施方式中，深度引擎 230 根据文档亲和度来计算文档的深度。文档之间的亲和度描述文档之间的关系。在具体实施方式中，平均文档亲和度可以以类似于平均词语亲和度怎样指示词语深度的方式来指示文档深度。可以以任何合适方式来定义文档亲和度。在一个实施例中，通用词语的数量 $P(D_1\&D_2)$指示既存在于文档 D_1 中又存在于文档 D_2 中的词语的数量，而分立词语数量 $P(D_1 + D_2)$指示存在于文档 D_1 或 D_2 中词语的数量。文档 D_1 和 D_2 之间的文档亲和度 DocAff 可以被定义为：

$$DocAff(D_1, D_2) = P(D_1 \& D_2) / P(D_1 + D_2)$$

深度引擎 230 可以以与计算平均词语亲和度类似的方式来计算平均文档亲和度。具有较低平均亲和度的文档被认为较深，而具有较高平均亲和度的文档被认为较浅。

在具体实施方式中，深度引擎 230 可以通过向文档亲和度图应用链路分析来计算文档深度。除文档亲和度图的节点表示文档而不是词语之外，文档亲和度图可以与亲和度图 150 类似。深度引擎 230 使用第二文档相对于给定的第一文档的文档亲和度来加权从代表第一文档的节点到代表第二文档的第二节点的链路。接着，可以对外发链路（outgoing link）的权重进行归一化。

在具体实施方式中，深度图可以被显示在用户接口上以示出文档的深度。也可以显示可以用来选择深度级别的深度滑块。在具体实施方式中，如果文档包括多个部分的较大文档，则深度图可以指示各部分的深度。

在具体实施方式中，深度引擎 230 可以以任何其他合适方式来计算文档深度（例如，处理文档的亲和度直方图，和/或基于深度截取不同词语的百分比，接着处理直方图）。其他方法包括 Gunning-Fog、Flesch 或 Fry 方法。

在具体实施方式中，深度引擎 230 可以通过将深度值映射到具体深度级别来计算深度。在具体实施方式中，范围 R_i 中的深度值可以被映射到级别 L_i。例如，$R_0 = \{r_0: r_0 < c_0\}$可以被映射到级别 L_0、$R_1 = \{r_1: c_0 <$

$r_1 < c_1\}$可以被映射到级别L_1, ...,以及$R_n = \{r_n: c_n < r_n\}$可以被映射到级别L_n。该范围可以包括任何合适深度值并且不需要具有相同大小。可以存在任何合适数量的级别（例如小于 5、5 到 7、7 或 8、8 到 10、10 到 20、20 到 50、50 到 100，等于或大于 100）。

主题引擎 240 可以确定文档的主题（或话题）。在具体实施方式中，主题引擎 240 根据由聚类模块 31 识别的、文档中词语的群集来确定主题。如上面所讨论的，词语的群集可以指定词语集的主题（或话题）。文档的主题可以提供关于文档的内容的有用信息。例如，包括群集{肾脏的（renal）、肾（kidney）、蛋白质、问题}的文档可能关于由于肾功能衰退而导致的蛋白质流失，而不是芸豆的蛋白质含量。

在具体实施方式中，主题引擎 240 根据主题映射来确定主题。在这些实施方式中，使用任何合适技术（例如，词条频率-逆文档频率（TF-IDF：term frequency-inverse document frequency）技术）从文档中提取关键词。关键词被用来从主题映射中选择候选主题。候选主题与文档进行比较，以确定该主题多大程度上与文档匹配。在具体实施例中，候选主题的直方图可以与文档的直方图进行比较。如果候选主题与文档匹配，则这些主题可以提供文档的主题的类型估计和数量估计。

专度引擎 240 可以计算文档的专度。在具体实施方式中，专度引擎 240 可以对文档指配专度值和/或专度级别。较专属的文档可以被指配较高的专度值或级别，而较不专属的文档可以被指配较低的专度值或级别。

在具体实施方式中，专度引擎 240 根据文档的主题数量来计算专度。在具体实施例中，较专属的文档可以具有较少的主题，而较不专属的文档可以具有较多主题。在具体实施方式中，专度引擎 240 根据文档的主题数量和这些主题之间的亲和度来计算专度。在具体实施例中，较专属的文档可以具有较少的主题，且这些主题之间具有较高的亲和度，而较不专属的文档可以具有较多的主题，且这些主题之间具有较低的亲和度。

在具体实施方式中，主题数量可以取决于深度（或级别）。例如，较浅深度处的单个主题可以表示较大深度处的多个主题。在具体实施方式中，可以通过用户使用深度滑块来选择深度，或者深度可以是预先确定

的。在具体实施方式中，级别可以由用户选择或者可以被预先确定。例如，可以定义任何合适数量的级别，并且可以根据该级别计算深度。例如，级别可以是基于领域（例如，工程、医学、新闻、体育或金融领域）、基于专业（例如，心病学、眼科学或肾脏专业）、基于课题（例如，高血压、胆固醇、搭桥手术或动脉阻断题目）、基于细节（例如，体位性低血压、慢性高血压或急性高血压细节）、基于消退（resolution）（例如，老年病因、药学、或遗传消退）、基于个人的（例如，用户查询级别）。

本体特征应用引擎250可以应用本体特征（例如深度、主题或专度），来在任何合适情形中执行本体特征分析。合适的情形的实施例包括：根据本体特征来搜索、分类、推荐或选择文档；报告文档的本体特征；以及确定一个或更多个用户的文档（或文档集）的本体特征。在具体实施方式中，本体特征应用引擎 250 可以使用包括关于本体特征的信息的索引。在一个实施例中，本体特征应用引擎 250 使用根据深度级别生成和/或维护的文档深度（DD：document depth）反向索引 62。DD 反向索引 62 包括 DD 反向索引列表，其中词语的 DD 反向索引列表列出了包括该词语的文档（或页面 50）的文档识别符。文档的文档识别符可以指示文档的深度。例如，用来编码文档识别符的二进制编码可以指示深度。在一些情况下，DD 反向索引列表可以仅列出具有令人满意的深度的文档。在另一实施例中，除反向索引 62 之外，本体特征应用引擎 250 还使用级别表和深度表。该深度表可以指示文档的深度。

在具体实施方式中，本体特征应用引擎 250 搜索具有本体特征的指定值（例如，文档深度或专度的指定值）的文档。该指定值可以由用户预先确定、计算或者选择。在具体实施方式中，可以使用深度滑块和/或专度滑块来选择这些值。

在具体实施方式中，本体特征应用引擎 250 可以将本体特征用作分类标准来分类文档。例如，本体特征应用引擎 250 可以针对主题以及其它分类标准根据文档深度和/或专度来分类文档。在具体实施例中，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分类的文档。在一些实施例中，本体特征应用引擎250使用非DD反向索引62来搜索

文档，并接着根据深度对这些文档分类。

在具体实施方式中，本体特征应用引擎250可以向客户端20以图形方式显示本体特征的值。可以为一些或所有文档（例如，为来自搜索结果的前X%的文档）提供图形显示。该本体特征值可以以任何合适方式呈现。在一些实施例中，图形指示符（例如，数量、词语或图标）可以指示该值。例如，图形指示符例如可以靠近搜索结果列表中的项、在线新闻的标题或文档图标放置。在某些实施例中，现有的插图（iconograph）的变更可以表示值。例如，图形指示符或文本的大小、字体、类型、颜色可以指示值。在另一实施例中，图表可以指示值。本体特征直方图可以包括文档数量轴和本体特征轴，且可以指示特定本体特征值的文档数量。例如，包括文档数量轴和文档深度轴的文档深度直方图可以指示特定文档深度的文档数量。

在具体实施方式中，文档特征应用引擎250可以允许用户请求搜索具有特定本体特征值的文档。可以允许用户指定用于查询的不同词语的值。在特定实施例中，本体特征应用引擎250可以为用户提供选项以选择深度，用户然后可以输入所选的深度。这些选项可以以任意合适的方式呈现，诸如以：（i）绝对词条（例如，代表深度的数值或数值范围）；（ii）相对词条（例如，搜索结果相对于深度的比例，诸如，“最深的X%”）；（iii）语义学词条（例如，‘介绍性的’、‘浅’、‘深’、‘很深’和/或‘高度专业’）；（iv）图形词条（例如，滑块、按钮和/或其他图形元素）或（v）词条的任意合适的组合（例如具有语义学标签的滑块）。在某些情况下，滑块可以包括浅端和深端。用户可以移动滑动器朝向一端或另一端以指示所选的深度。当提供搜索结果时，文档深度直方图可以通过滑块呈现，且可以使用滑动器作为文档深度轴。

在具体实施方式中，本体特征应用引擎250可以计算一个或更多个用户的集合的本体特征字符。本体特征字符可以包括主题上下文中的用户深度和用户专度。本体特征字符描述了文档的与用户集相关联的本体特征。例如，科学家可以使用比三年级学生更深的文档。可以针对一个或更多个主题给出本体特征字符。例如，遗传学家可以在遗传学领域中

使用比他在诗歌领域中使用的文档更深的文档。本体特征字符可以被用来确定用户的专长、为用户自动构建简历，以及分析用户的社会网络。

可以分析与用户相关联的任何合适的文档，以估计本体特征字符（例如，信件（例如，电子邮件和即时消息）、网页、以及搜索历史（例如搜索查询和选择的页面））。在具体实施方式中，本体特征应用引擎 250 可以随着时间跟踪本体特征字符，并且可以使用过去的字符来预测未来的字符。在具体实施例中，本体特征应用引擎 250 可以假设用户深度和/或专度总体上随时间和/或区域中的活动而增加。

在具体实施方式中，本体特征应用引擎 250 可以组合某些操作。例如，本体特征应用引擎 250 可以监视用户的深度，并且接着根据该用户深度来搜索文档。在一个实施例中，监视用户深度，并且接着根据该深度向用户提供新闻。预测未来的用户深度，并且提供适合预测的用户深度的新闻。

图 9 例示了可以选择标签来标记文档的标记模块 35 的一个实施方式。可以以任何合适的方式来选择标签。在具体实施方式中，标记模块 35 将话题（或主题）建模为话题的相关词语的统计分布。标记模块 35 使用统计分布来识别文档的高排位的词语中具有最高出现概率的话题，并且根据识别出的话题来选择文档的标签。在所示的实施方式中，标记模块 35 包括话题建模器 310 和文档标记器 314。在具体实施方式中，话题建模器 310 生成建模话题的统计分布，而文档标记器 314 基于该统计分布来选择标签。话题建模器 310 和文档标记器 314 可以利用任何合适方法来建模话题和选择标签。参照图 10 来描述方法的实施例。

在其他实施方式中，标记模块 35 通过分析文档的段落来指配标签。在这些实施方式中，标记模块 35 识别文档的段落的候选标签。标记模块 35 确定候选标签与文档的其他候选标签的关联性，并根据该关联性来选择该文档的标签。参照图 11 对通过分析文档的段落来指配标签的方法的实施例进行更详细地说明。

在再一实施方式中，标记模块 35 可以基于由用户或计算机选择的推荐标签来指配标签。在这些实施方式中，标记模块 35 推荐文档的标签。

推荐词条可以与目标标签具有较高的亲和度，而彼此之间具有较低的亲和度，以减小文档的本体空间。标记模块 35 可以响应于选择的标签来连续推荐标签。一旦已经选择了最终标签，标记模块 35 就可以对文档指配所选择的标签。参照图 12 对用于指配标签的方法的实施例进行更详细地说明。

图 10 例示了用于根据话题的统计分布来指配标签的方法的实施例。可以根据词语的总体来生成统计分布。可以使用任何合适的总体（例如语言或语料库（例如，因特网）的词语）。相对于其他词语，与话题相称的词语可能具有相对较高的出现概率。例如，对于话题“自行车”，相对于“砖块”、“桶”以及“披萨”等词语的出现概率来说，“轮胎”、“链条”、以及“骑乘”等词语可以具有相对更高的出现概率。

在步骤 410 处开始该方法，其中使用任何合适的排位技术来为语料库的文档的词条进行排位。在排位技术的一个实施例中，根据频率（例如词条频率或者词语频率-逆文档频率（TF-IDF））对词条进行排位。较高的频率可以产生较高的级别。在排位技术的另一实施例中，根据在以上随机机会中与其他词条共现的词条的标准差数量来对词条进行排位。较高的标准差数量可以产生较高的级别。

在步骤 414，将一个或更多个高排位的词条选作文档的关键词。在一些实施例中，可以使用排位的前 N 项，其中 N 可以是 1 到 5、5 到 10 或者大于等于 10 的值。在其他实施例中，可以使用具有高于文档的平均级别的预定距离（例如，一个标准差）的词条。

在步骤 418，根据它们的关键词来对文档进行聚类，其中各群集与关键词相关联。针对群集定义的关键词是该群集的话题。如果文档具有 N 个关键词，则将在 N 个群集中呈现该文档。在步骤 422 移除小群集。小群集可以是未满足大小阈值的群集（例如，表现为低于 M 个文档的群集，其中 M 可以是在范围 0 到 50、50 到 100，或者大于等于 200 的值）。在一些实施例中，可以根据语料库的大小来计算 M。例如，M 可以是在范围 0%到 3%、3%到 5%或者大于等于 5%的值。

在步骤 426 收集群集的统计，并在步骤 428 根据该统计来生成群集

的统计分布。可以收集任何合适的统计来生成任何合适的统计分布（例如频率分布和/或概率分布）。在具体实施例中，针对群集的各词语计算指示群集中的词语频率的词条频率。可以根据群集中词语出现的数量或者根据在包括该词语的群集中文档的数量来计算词条频率。根据该词条频率来生成词条分布。在其他实施例中，针对各其他群集计算指示群集的话题与另一群集的话题的共现的共现值。根据该共现值来生成共现分布。如果在步骤 430 存在下一群集，则方法返回到步骤 426 来收集下一群集的统计。如果在步骤 430 不存在下一群集，则方法前进到步骤 434。

在步骤 434 处合并具有类似统计分布的群集。可以对统计分布进行比较，并且类似的统计分布可以被合并入单个频率分布。例如，话题“轿车”和“汽车”的群集可以具有类似统计分布，因此将它们合并到单个群集。如果分布之间的差异小于差异阈值（difference threshold），则可以认为统计分布是类似的。差异阈值可以具有任何合适值（例如，在小于或等于 1%、5%到 10%或者大于等于 10%的范围中的值）。较大群集的话题可以被选作合并后的群集的话题。

在步骤 438，基于产生的群集将话题重新指配为文档的标签。因为一些群集已经被合并，而其他群集已经被移除，所以指配给文档的话题可能改变。重新指配的话题可以用作文档的信息量更大、重复性更少的标签。接着该方法结束。可以在更新语料库的文档时执行该方法。

在步骤 442，向文档指配标签。文档标记器 314 可以根据统计分布以任何合适方式为文档指配标签。在一些实施例中，文档标记器 314 可以根据在步骤 438 处执行的话题的重新指配来为语料库中的文档指配标签。

在其他实施例中，文档标记器 314 可以为语料库中非必要的文档指配标签。可以使用统计分布来识别文档的所选词语中具有较高的出现概率的话题，并且识别的话题可以被选作标签。在这些实施例中，文档标记器 314 根据任何合适的排位技术（例如上面所讨论的技术）来对文档的词语排位。从最高排位的词语开始，文档标记器 314 根据话题的统计分布来确定该词语对于各话题的频率。文档标记器 314 接着可以从词语

最频繁出现的话题到词语最少出现的话题对话题进行排位。可以生成该词语针对这些话题的统计分布。

在这些实施例中，文档标记器 314 接着可以针对文档的一个或更多个其他高排位的词语以类似方式生成统计分布。在具体实施例中，可以例如等同地或者根据词语的级别来对词语的统计分布进行加权。例如，较高排位的的词语可以具有包括较高权重的统计分布。统计分布可以被合并，以产生合并的统计分布。在具体实施方式中，可以对加权后的统计分布求和。例如，对与具体话题相关联的值求和以产生一值，该值表示该话题在文档的给定高排位的词语中的可能性。文档标记器 314 可以将一个或更多个可能的话题指配为文档的标签。

图 11 例示了用于通过分析文档的段落来为文档指配标签的方法的实施例。该方法可以用于包括微观点（micro-idea）、观点以及学说（hypothesis）的文档。在具体实施方式中，微观点包括独立、完整的表达单元。一个或更多个相关微观点可以形成观点。一个或更多个相关观点可以形成学说。在具体实施例中，语句表达微观点、段落表达观点，而一系列相关段落表达学说。在这些实施例中，段落是相关的，所以段落的核心词条可以具有相对高的有向亲和度。多个核心词条的一些交集可以被用作标签。

在步骤 506 开始该方法，其中文档的段落 P_i 被识别为用于分析的文本单元。段落可以指以任何合适方式（例如，通过固定数量或可变数量的词语、通过段落标志或通过群集）指定的字符、词语和/或语句的任何合适集合。段落可以被定义为例如包括足够数量的足够复杂的词语。

在步骤 510 选择段落 P_i。在步骤 514 处，针对段落 P_i 建立标签 t_k 的候选标签集 $S_i=<t_1, t_2, \ldots, t_m>$。在具体实施方式中，更高排位的的词语可以被选作候选标签。（可以根据任何合适的排位技术对词语排位。）在具体实施方式中，可以根据期望产生的候选标签的数量来选择候选标签的初始数量。例如，如果期望产生的数量为 k，则初始数量可以为 c*k，其中 c>1。参数 c 可以具有任何合适值（例如，c=2、3、4 或 5）。最高排位的的候选标签可以被选作用于集合 S_i 的根 r_i。

在步骤 518，根据任何合适的关联性技术来确定候选标签彼此间的关联性。一般来说，可以以任何合适方式（例如，使用任何合适亲和度）来测量关联性。例如，与目标标签更相关的标签可以被认为更相关标签，而与目标标签较少相关的标签可以被认为较不相关标签。在具体实施方式中，可以（例如，使用有向亲和度和/或差分亲和度）对标签进行聚类，从而认为群集的标签是相关的。

在步骤 520 为候选标签指配偏好权重。可以根据任何合适排位技术来指配偏好权重。例如，较大的偏好权重可以被赋予在段落中具有较高频率和/或在文档中具有较大平均亲和度的倒数的候选标签。在步骤 524，从候选标签集中移除与其他候选标签不充分相关的候选标签。任何合适关联性阈值都可以指明标签是否与其他标签充分相关。在步骤 530，询问是否存在下一段落。如果存在下一段落，则方法返回到步骤 510 以选择下一段落。如果不存在下一段落，则方法前进到步骤 534。

在步骤 534，确定不同段落的候选标签集的关联性。可以根据任何合适关联性技术来确定关联性。在具体实施方式中，与段内分析的情况类似，可以对候选标签进行聚类，并且可以认为群集中的候选标签充分相关。在其他实施方式中，可以针对各候选标签生成互相关概况（profile）。互相关概况指示候选标签与其他候选标签（例如，其他候选标签集的标签）的互关联性。较大的互关联性代表了较大关联性。

可以以任何合适方式计算互相关概况。在一些实施例中，根据包括候选标签的候选标签集的数量生成候选标签的互相关概况，并且候选标签的互相关概况可以考虑候选标签集中的候选标签的频率。在较多具有较高频率的候选标签集中出现的候选标签可以具有较高互关联性。

在其他实施例中，可以根据根 r_i 和根 r_j 的有向亲和度来确定集合 S_i（具有根 r_i）相对于其它集合 S_j（具有根 r_j）的候选标签的互相关概况。在这些实施例中，可以通过将候选标签的偏好权重与集合 S_j 上的根 $r_i \rightarrow r_j$ 的有向亲和度相乘来计算集合 S_i 和具体集合 S_j 的候选标签的互关联值。可以通过合并（例如，求和）具体集合的互关联值来计算候选标签与集合 S_j 的互关联性。

在再一实施例中，可以根据单独标签 t_i 和 t_j 的有向亲和度来确定集合 S_i 的候选标签 t_i 相对于其他集合 S_j（具有标签 t_j）的互相关概况。在这些实施例中，通过确定集合 S_j 上的标签 $t_i \rightarrow t_j$ 的有向亲和度并对这些有向亲和度求和来计算集合 S_i 和具体集合 S_j 的候选标签的互关联值。可以通过合并具体集合的互关联值来计算候选标签和集合 S_j 的互关联性。

在步骤 538，从候选标签中选择标签。在具体实施方式中，选择与其他候选标签最高度相关的候选标签。在一些实施例中，可以选择群集的高排位的候选标签。在其他实施例中，可以根据互相关概况来选择具有最高互关联性的候选标签。选择的标签的数量 k 可以是预定常数，或者可以是根据查询词条的深度确定的值。例如，对于具有较深词条的查询，可以使用更小或更大的 k。随后该方法结束。

图 12 例示了用于响应于选择的标签指配标签的方法的实施例。该方法在步骤 450 处的初始阶段启动。在初始阶段，文档标记器 314 接收初始标签作为文档的目标标签。初始标签可以来自任何合适源。例如，可以通过用户或逻辑（例如，计算机）输入初始标签。逻辑可以输入从文档的分析产生的标签、与用户相关联的其他文档，或者针对文档选择的其他标签。文档标记器 314 可以记录标签的源。

在具体实施方式中，文档标记器 314 可以在客户端 20 处启动图形用户接口的初始显示，该图形用户接口允许用户与文档标记器 314 交互。在一些实施例中，接口可以允许用户请求添加或删除标签。在其他实施例中，接口可以包括图形要素，该图形要素允许用户对具体词条指示标签应该具有的期望亲和度。例如，接口可以包括滑块，该滑块可以更靠近词条移动以指示较高亲和度，或者远离词条移动以指示较低亲和度。

在步骤 454 处的候选阶段，文档标记器 314 响应于输入的标签来推荐词条。可以选择所推荐的词条以将文档与最小本体空间量相关联。例如，所推荐的词条可以与输入标签具有较高的亲和度，而彼此间具有较低的亲和度。例如，如果输入标签为“树”，则推荐标签可以为“植物”、“族”或者“计算机科学”。

推荐词条可以避免多余指定（over specification）和不足指定（under

specification)。该多余指定是由于实质上提供了未提供很多额外信息的本体冗余标签而引起的。例如，如果文档具有标签“树”和“木材”，则添加“森林”就没有提供很多额外信息。该不足指定是由于提供无法消除文档歧义的标签而引起的。例如，文档的标签“bank（银行、河堤、台边）”无法指明该文档是涉及金融机构、河流还是台球桌的边沿。

在步骤 458 处的测试阶段中，文档标记器 314 监视（例如，由用户）已选择的推荐词条和未选择或淘汰的词条。例如，文档标记器 314 接收“流体”，并推荐“可适应的”、“灵活的”、“液体”、“溶液”以及“融化的”。文档标记器 314 注意到“液体”和“融化的”被淘汰，因此文档标记器 314 没有推荐“溶液”。已选词条被添加到目标标签的集合中。

在具体实施方式中，文档标记器 314 可以记录标签的源（例如，用户或逻辑（例如计算机））。源可以具有任何合适的应用。例如，源可以被用来排位搜索结果。在一个实施例中，对于具有由用户选择的标签的搜索结果分配比具有由逻辑生成的标签的结果更高的级别。

在步骤 462 的演进阶段，文档标记器 314 评估推荐词条和选择词条之间的差异，以推荐新的词条。文档标记器 314 可以推荐与选择词条具有较高亲和度（例如有向亲和度和/或差分亲和度）和/或与被淘汰词条具有较低亲和度的词条，并且可以避免推荐与被淘汰词条具有较高亲和度和/或与选择词条具有较低亲和度的词条。在具体实施方式中，文档标记器 314 可以移除一个或更多个本体冗余标签。可以针对任何合适数量的迭代（例如 1 到 5、6 到 10 或者大于等于 10 个迭代）来推荐和选择标签。

在步骤 466 处的指配阶段，文档标记器 314 向文档指配一个或更多个标签。在具体实施方式中，文档标记器 314 可以响应于测试阶段指配标签，或者独立于测试阶段而指配一个或更多个初始标签。接着该方法结束。

在不脱离本发明的范围的情况下，可以对这些方法进行修改、添加或省略。这些方法可以包括更多、更少或其他步骤。另外，可以以任何合适顺序来执行这些步骤。

在具体实施方式中，这些方法可以被执行以选择搜索词条而不是选

择标签。可以在本文档中包括的说明（具体地说，与用于指配标签的方法相关联的说明）中通过用“搜索词条”替换“标签”来描述这些实施方式。

例如，方法可以在初始阶段启动。在初始阶段，初始搜索词条被接收为搜索的目标搜索词条。初始搜索词条可以来自例如可以由用户或逻辑（例如计算机）输入的任何合适源。在候选阶段，可以响应于输入的搜索词条来推荐词条。推荐词条可以被选择以将搜索与最小本体空间量相关联。在测试阶段，可以监视（例如由用户）已选择的推荐词条和未被选择或被淘汰的词条。在演进阶段，可以评估推荐词条和选择词条之间的差异，以推荐新的词条。可以针对任何合适数量的迭代（例如 1 到 5、6 到 10 或者大于等于 10 个迭代）来推荐和选择搜索词条。可以响应于已选择的搜索词条来选择搜索词条。

在不脱离本发明的范围的情况下，可以对这些方法进行修改、添加或省略。这些方法可以包括更多、更少或其他步骤。另外，可以以任何合适顺序来执行这些步骤。

本发明的具体实施方式可以提供一个或更多个技术优点。一个实施方式的技术优点可以是：通过分析文档的段落来为文档选择标记。针对各段落识别候选标记的集合，并且建立不同候选标记集合之间的候选标记的关联性。高度相关的候选标记可以被有效地识别，并被选作标记。

尽管根据具体实施方式对该公开进行了说明，但本领域技术人员应当清楚这些实施方式的改变例和置换例。因此，这些实施方式的上述描述不对本公开构成限制。在不脱离如下列权利要求所限定的本公开的精神和范围的情况下，可以对本发明进行其它改变、代替以及变更。

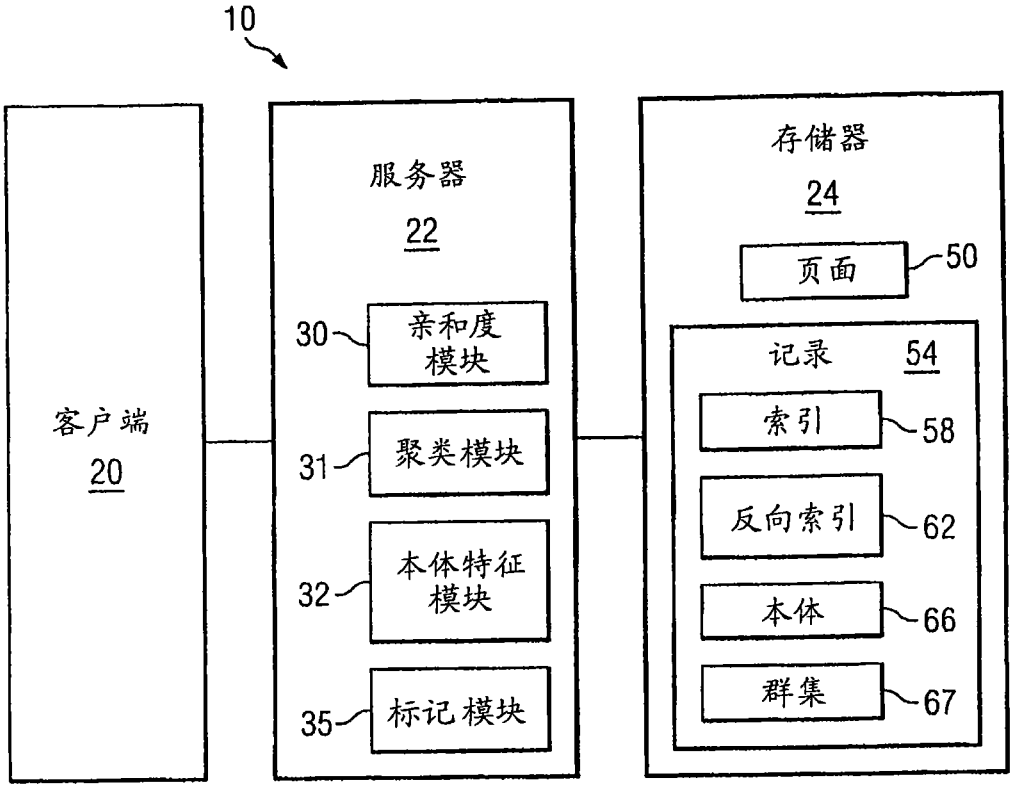

10
客户端
20
服务器
22
30
亲和度模块
31
聚类模块
32
本体特征模块
35
标记模块
存储器
24
页面
50
记录
54
索引
58
反向索引
62
本体
66
群集
67

图1

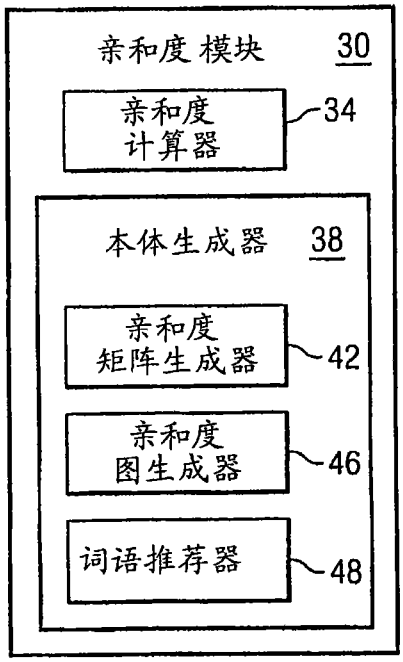

亲和度模块
30
亲和度计算器
34
本体生成器
38
亲和度矩阵生成器
42
亲和度图生成器
46
词语推荐器
48

图2

110

	w_1 狗	w_2 树林	w_3 树	w_4 图形	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 树林			0.092	0.004	0.052
w_3 树				0.042	0.056
w_4 图形					0.222

图 3

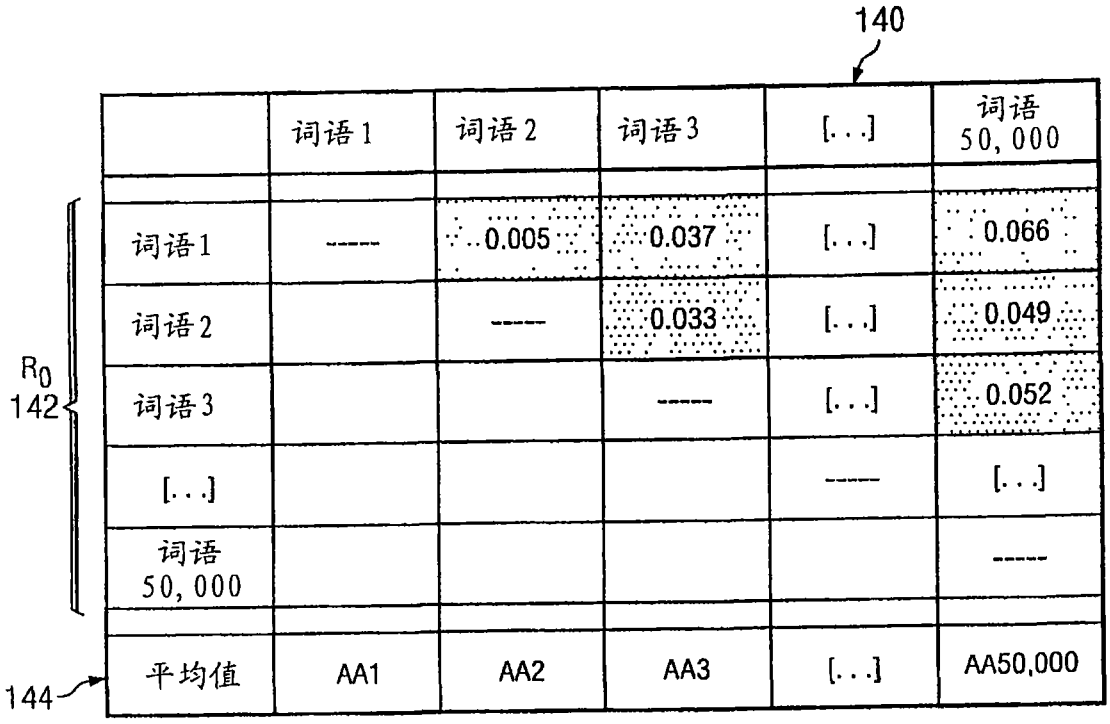
120

A 124

B 128	w_0 二进制	w_1 狗	w_2 树林	w_3 树	w_4 图形	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 树林	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 图形	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
词语1
词语2
词语3
[...]
词语 50,000
词语1

0.005
0.037
[...]
0.066
词语2

0.033
[...]
0.049
词语3

[...]
0.052
[...]

词语 50,000

平均值
AA1
AA2
AA3
[...]
AA50,000
R0
142
144

图5

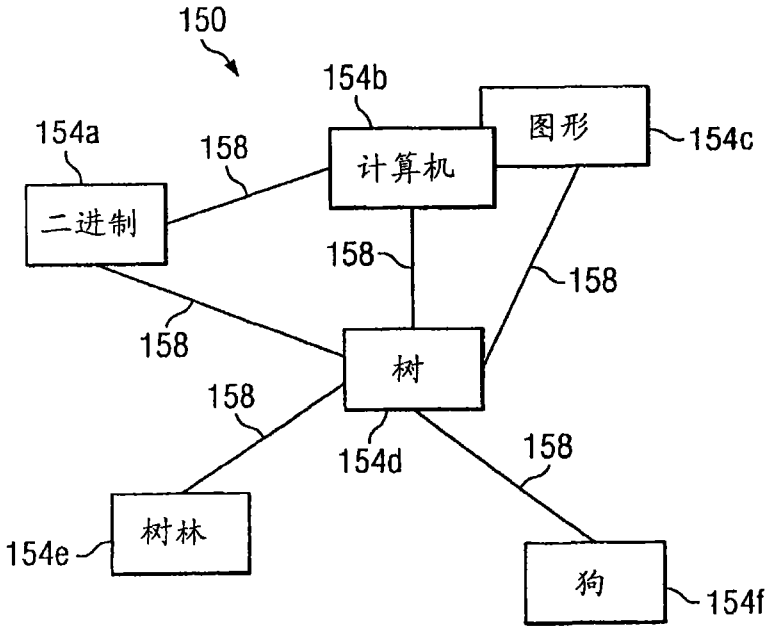

150
154b
图形
154c
154a
158
计算机
二进制
158
158
158
树
158
154d
158
树林
154e
狗
154f

图6

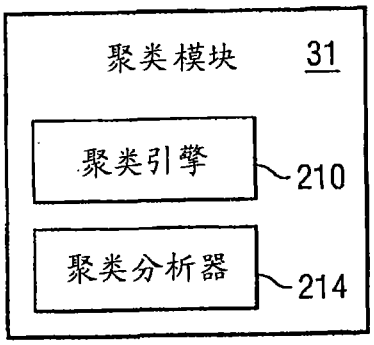

聚类模块 31
聚类引擎 210
聚类分析器 214

图7

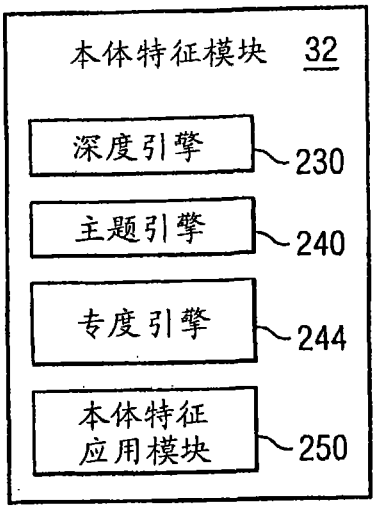

本体特征模块 32
深度引擎 230
主题引擎 240
专度引擎 244
本体特征应用模块 250

图8

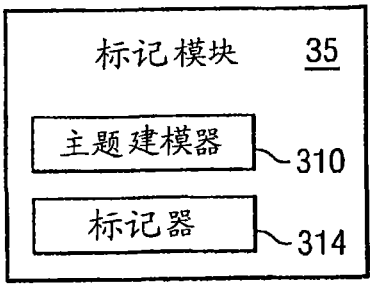

标记模块 35
主题建模器 310
标记器 314

图9

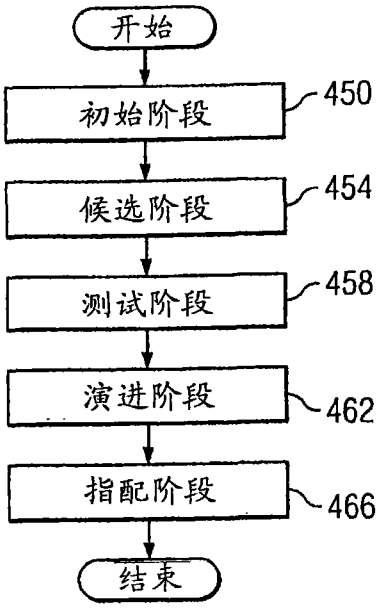

开始
初始阶段 450
候选阶段 454
测试阶段 458
演进阶段 462
指配阶段 466
结束

图12

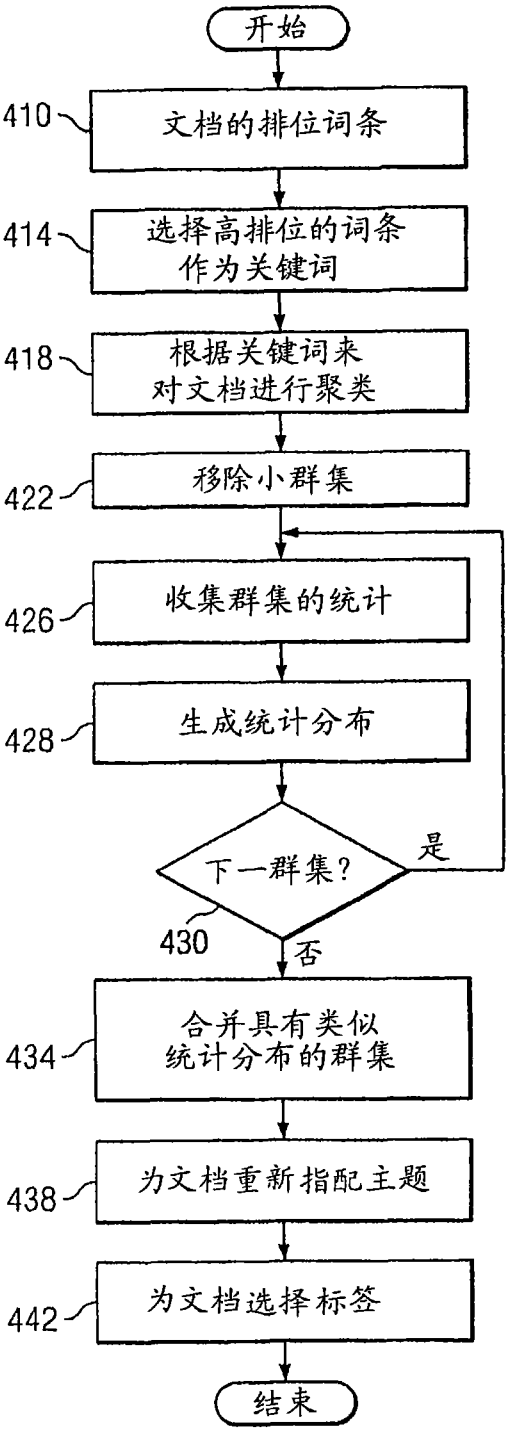

开始
410
文档的排位词条
414
选择高排位的词条
作为关键词
418
根据关键词来
对文档进行聚类
422
移除小群集
426
收集群集的统计
428
生成统计分布
下一群集?
是
430
否
434
合并具有类似
统计分布的群集
438
为文档重新指配主题
442
为文档选择标签
结束

图10

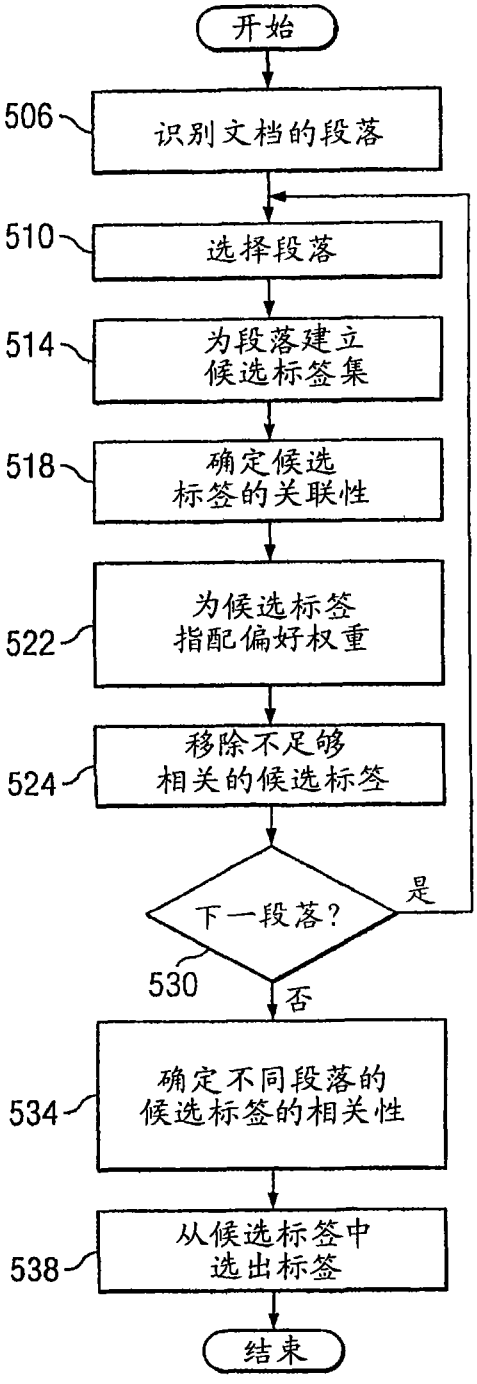

开始
506
识别文档的段落
510
选择段落
514
为段落建立
候选标签集
518
确定候选
标签的关联性
522
为候选标签
指配偏好权重
524
移除不足够
相关的候选标签
下一段落?
是
530
否
534
确定不同段落的
候选标签的相关性
538
从候选标签中
选出标签
结束

图11