[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 17/00 (2006.01)
G06F 17/30 (2006.01)

[12] 发明专利申请公布说明书

[21] 申请号 200680035828.4

[43] 公开日 2008 年 9 月 24 日

[11] 公开号 CN 101273350A

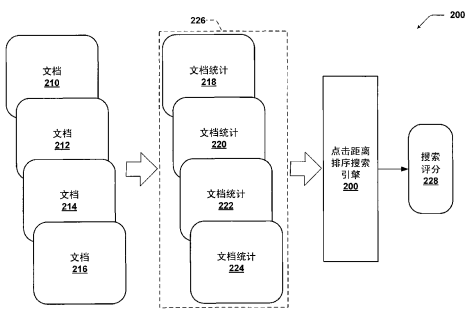
[22] 申请日 2006.9.25
[21] 申请号 200680035828.4
[30] 优先权
[32] 2005. 9.29 [33] US [31] 11/238,906
[86] 国际申请 PCT/US2006/037571 2006.9.25
[87] 国际公布 WO2007/041120 英 2007.4.12
[85] 进入国家阶段日期 2008.3.28
[71] 申请人 微软公司
地址 美国华盛顿州
[72] 发明人 M·佩特里克
[74] 专利代理机构 上海专利商标事务所有限公司
代理人 张政权
权利要求书 4 页 说明书 12 页 附图 7 页

[54] 发明名称

点击距离确定

[57] 摘要

根据本地存储的倒排索引中包括的数据为文档语料库中的每一文档进行点击距离值的高效确定。点击距离是从网络上的第一文档到另一文档的点击或用户导航次数的度量。 在本地存储的倒排索引中存储专门字词。 专门字词使源文档与一组目标文档相关。 当根据传入该专门字词之一的查询为相应的目标文档集查询倒排索引时向源文档赋一点击距离。 为文档语料库中的每一文档重复该过程。



1．一种用于确定与网络上的文档相关联的点击距离的计算机实现方法，包括：

存储文档（210）的文档和链接信息（218）；

在所述文档和链接信息（218）内，还存储与目标文档（574）相关联的专门字词（580），其中所述专门字词（580）指定对应于所述目标文档（574）的源文档（572）；

在倒排索引（510）中包括所述专门字词（580），其中所述本地存储的倒排索引（510）使所述专门字词与所述目标文档（574）的标识符相关；以及

当根据传入所述专门字词（580）的查询对所述目标文档（574）查询倒排索引时，对所述源文档（572）赋予一点击距离。

2．如权利要求 1 所述的计算机实现方法，其特征在于，所述倒排索引对应于一锚文本索引（612），所述锚文本索引被安排成存储所述锚文本（576）中包含的字词和被标识为与每一字词相关联的文档（210）的清单。

3．如权利要求 1 所述的计算机实现方法，其特征在于，存储所述文档和链接信息（218）还包括存储所述源文档（572）中包括的锚文本（576），其中存储所述锚文本（576）使得所述锚文本（576）与所述目标文档（574）相关联。

4．如权利要求 3 所述的计算机实现方法，其特征在于，还包括在锚文本表（570）中存储包括所述锚文本（576）的所述文档和链接信息（218）。

5．如权利要求 4 所述的计算机实现方法，其特征在于，当所述目标节点之一的点击距离大于当前节点的点击距离加上一变量时（718），所述目标节点之一的点击距离被设置为所述当前节点的点击距离加上所述变量（720）。

6．如权利要求 1 所述的计算机实现方法，其特征在于，存储所述专门字词（580）还包括将所述专门字词追加到存储在锚文本表（570）内的锚文本（576）。

7．如权利要求 6 所述的计算机实现方法，其特征在于，还包括构造锚图（300），它包括所述网络的文档作为所述锚图（300）的节点（310），还包

括所述锚文本表（570）的记录作为所述锚图（300）的链接（320）。

8．如权利要求 1 所述的计算机实现方法，其特征在于，在所述倒排索引（510）中包括所述专门字词（580）还包括当追加所述专门字词（580）的锚文本（576）被插入所述倒排索引（510）时也将所述专门字词（580）插入所述倒排索引（510）内。

9. 如权利要求 1 所述的计算机实现方法，其特征在于，对所述源文档（572）赋予一点击距离还包括当所述源文档对应于锚图（300）内的高权威节点（330）时将所述源文档（572）赋予一初始点击距离值。

10. 如权利要求 9 所述的计算机实现方法，其特征在于，为所述锚图（300）指定多于一个的高权威节点（330）。

11．如权利要求 9 所述的计算机实现方法，其特征在于，还包括当所述源文档（272）对应于不同于高权威节点（330）的一节点（310）时赋一初始无穷大值作为所述点击距离值（706）。

12．如权利要求 11 所述的计算机实现方法，其特征在于，还包括当所述源文档具有不同于所述初始无穷大值的点击距离值时将所述源文档插入队列中（708）。

13．如权利要求 11 所述的计算机实现方法，其特征在于，还包括从所述队列中检索所述源文档以查询所述倒排索引（714）。

14. 一种其上存储当执行时实现如权利要求 1 所述的计算机实现方法的指令的计算机可读介质。

15．一种系统，包括：

文档接口（530），被安排成提供对存储在网络上的文档的访问；

锚文本表（570），被安排成存储对应于所述网络上的所述文档的文档和链接信息，其中所述锚文本表（570）包括目标文档（574）及其相关联的锚文本（576）的记录；

专门字词（580），被追加到与每一目标文档（574）相关联的所述锚文本（576），其中所述专门字词（580）被配制成标识对应于每一目标文档（574）的源文档；

倒排索引（510），被安排成列出锚文本（576）中所包括的字词以及与每

一字词相关联的目标文档（574），使得所述专门字词（580）也随与每一专门字词（580）相关联的所述目标文档（574）一起在所述倒排索引（510）中列出；以及

客户机接口（540），被安排成实现一搜索引擎，其中所述搜索引擎通过以下步骤来确定与存储在所述网络上的每一文档相关联的点击距离：当为对应于存储在所述网络上的每一文档的目标文档（574）进行对所述倒排索引（510）的查询时，递增与存储在所述网络上的所述文档相关联的点击距离值。

16．如权利要求 15 所述的系统，其特征在于，所述倒排索引（510）对应于一锚文本索引（612），所述锚文本索引被安排成存储包含在所述锚文本（576）中的字词和被标识为与每一字词相关联的文档（210）的清单。

17．如权利要求 15 所述的系统，其特征在于，所述倒排索引（510）与所述网络上的所述文档（210）相比是本地存储的。

18．如权利要求 15 所述的系统，其特征在于，所述倒排索引（510）对应于一分区索引，其中第一分区对应于主索引，而第二分区对应于锚文本索引。

19．一种其上存储当被执行时实现如权利要求 15 所述的系统的指令的计算机可读介质。

20．一种包括用于确定点击距离的计算机可执行指令的计算机可读介质，所述指令包括：

在网络上存储文档（210）的文档和链接信息（218），使得在存储器中启动表示所述网络的网络图；

当在所述网络图中表示的每一文档（210）具有不同于第一点击距离值的点击距离值时将所述文档（210）存储在队列中；以及

当所述队列不为空时：

从所述队列检索文档（210）；

通过查询锚索引（570）来确定与所检索的文档相关联的目标文档（574），其中所述锚索引（570）包括被安排成使所述网络上的文档与其目标文档（574）相关联的专门字词（580）；

为与所检索文档相关联的每一目标文档赋予一点击距离（572），其中当每一目标文档（574）的点击距离大于与所移除文档相关联的点击距

离加上一变量时，用不同于所述第一点击距离值的新点击距离值更新每一目标文档；和

将更新后的每一目标文档（574）添加到所述队列。

点击距离确定

背景

在文本文档搜索中，用户一般将查询输入搜索引擎。搜索引擎针对经索引文档的数据库评估该查询，并返回最佳满足该查询的排序的文档列表。由搜索引擎用算法生成表示文档有多满足该查询的度量的评分。通常使用的评分算法依赖于将查询分成搜索项，以及使用关于各个项在要搜索的文本文档的正文中出现的统计信息。文档以根据其相应评分的排序顺序列出，使得用户可在搜索结果列表的顶部看到最佳匹配搜索结果。

某些搜索引擎可能采用以改善结果质量的另一评估是通过选定的排序函数来修改结果的排序。一个示例性的排序函数确定当一页面链接至另一页面时，它实际上为该另一页面投了票。对一页面投的票越多，该页面就越重要。排序函数也可考虑是谁投的票。页面越重要，它们的投票也越重要。这些投票被累积，并用作网络上页面排序的组成部分。

使用排序函数来改进排序的质量。然而，排序函数的有效性可能受到网络拓扑的影响。例如，使用上述投票的排序函数在内联网设置中可能较不奏效。内联网是使用与因特网相同的某些协议但仅可由用户的子集，诸如一公司的雇员访问的网络。内联网的页面没有与因特网精确相同地结构化或连接，因此其与因特网设置相比，排序函数产生的结果的相关性可能不会被降低。

概述

本发明的各方面涉及提供快速点击距离确定以便根据点击距离排序搜索结果。点击距离以相对于其它点击距离确定方法的相对较短时间来确定。点击距离是测量到达网站的给定页面所需的“点击”次数的查询无关相关性度量。网络上的文档通常被组织成树结构，具有一根节点以及从该根延伸至其它节点的后续分支。内联网的根节点通常指的是其主页。

在树结构中，点击距离按自根节点在路径上遍历的分支数目表示。一旦为

页面确定了点击距离，点击距离可被纳入该页面的评分中。纳入点击距离的页面评分确定该页面在搜索结果内的其它页面中的排序。

根据一个方面，首先“爬寻”网络来生成与网络的链接和页面相关联的特性的表。“爬寻”指的是将若干文档（或信息的任何类似离散单位）自动收集到被称为索引的数据库中。爬寻通过跟随某些文档内的文档引用链接并在随后处理找到的每一文档来遍历网络上的多个文档。通过标识文档中的关键词（keyword）或一般文本来处理文档以创建索引。

示例性索引可以是具有一字词（word）列和一指示可在其中找到这些字词的文档的列的倒排列表。当用户输入一个或多个搜索项时，获取结果并应用包括点击距离函数的排序算法。点击距离函数基于页面的所确定点击距离，或正或负地影响某些页面的评分，改进向用户返回的结果。

此处所述的点击距离确定通过使点击距离确定对于排序搜索结果的排序引擎而言是本地的来减少确定点击距离所需的时间。标识源文档的特殊字词被包括在倒排索引中，并使得源文档与共享该源文档的目标文档列表相关联。遍历该倒排索引允许通过检查这些专门字词和它们所涉及的文档列表来确定点击距离。无需反复和昂贵地引用其它数据表或文档本身的语料库来检查倒排索引。

提供该概述以用简化形式介绍将在以下详细描述中进一步描述的一些概念。该概述不旨在标识要求保护的主题的关键特征或必要特征，也不旨在用于帮助确定所要求保护的主题的范围。

附图简述

参考以下附图描述了本发明的非限定性且非详尽的实施例，其中除非另有指定，否则相同的参考标号指代各个视图中相同的部分。

图 1 示出了可在一个示例性实施例中使用的示例性计算设备；

图 2 示出了可包括点击距离快速确定功能性的用于排序搜索结果的系统；

图 3 示出了示例性网络图；

图 4 示出了示例性分层结构网络图；

图 5 示出了用于索引文档的示例性系统的功能框图；

图 6 示出了用于索引的示例性结构的功能框图；

图 7 示出了根据本发明用于确定点击距离的示例性过程的逻辑流程图。

详细描述

以下参考附图更全面描述本发明的实施例，附图构成了本发明的一部分且示出了用于实现本发明的具体示例性实施例。然而，实施例可用众多不同形式实现，且不应被解释为限于此处所述的实施例；相反，提供这些实施例使得本公开将是彻底且完整的，且将向本领域的技术人员充分传达本发明的范围。本发明的实施例可被实现为方法、系统或设备。从而，本发明的实施例可采取完全硬件实现、完全软件实现或组合软件和硬件方面实现的形式。因此以下详细描述不用作限定意义。

本发明的各个实施例的逻辑操作被实现为（1）计算系统上运行的计算机实现步骤的序列和/或（2）计算系统内的互连机器模块。实现取决于对实现本发明的计算系统性能要求的选择。从而，组成此处所述的本发明的实施例的逻辑操作可被替换地称为操作、步骤或模块。

说明性操作环境

参考图 1，用于实现本发明的一个示例性系统包括诸如计算设备 100 的计算设备。计算设备 100 可被配置成客户机、服务器、移动设备或任何其它计算设备。在非常基本的配置中，计算设备 100 一般包括至少一个处理单元 102 和系统存储器 104。取决于计算设备的确切配置和类型，系统存储器 104 可以是易失性的（诸如 RAM）、非易失性的（诸如 ROM、闪存等）或这两者的某种组合。系统存储器 104 通常包括操作系统 105、一个或多个应用程序 106 并且可以包括程序数据 107。在一个实施例中，应用程序 106 包括用于实现本发明的功能性的点击距离确定应用程序 120。在图 1 用虚线 108 内的那些组件示出了该基本配置。

计算设备 100 可以具有其他的特征和功能性。例如，计算设备 100 也可以包括诸如磁盘、光盘或磁带的其他数据存储设备（可移动和/或不可移动）。这种其他存储在图 1 中由可移动存储 109 和不可移动存储 110 示出。计算机存储

介质可包括以用于储存诸如计算机可读指令、数据结构、程序模块或其它数据等信息的任一方法或技术实现的易失性和非易失性，可移动和不可移动介质。系统存储器 104、可移动存储 109 和不可移动存储 110 都是计算机存储介质的示例。计算机存储介质包括但不限于，RAM、ROM、EEPROM、闪存或其他存储器技术、CD-ROM、数字通用盘（DVD）或其它光盘存储、磁盒、磁带、磁盘存储或其它磁存储设备、或可以用来储存所期望的信息并可由计算机 100 访问的任何其它介质。任何这样的计算机存储介质可以是设备 100 的一部分。计算设备 100 也可以具有输入设备 112，诸如键盘、鼠标、笔、语音输入设备、触摸输入设备等。也可以包括诸如显示器、扬声器、打印机等输出设备 114。

计算设备 100 也可以包括使设备能够诸如通过网络来与其他计算设备 118 通信的通信连接 116。通信连接 116 是通信介质的一个示例。通信介质通常具体化为诸如载波或其它传输机制等已调制数据信号中的计算机可读指令、数据结构、程序模块或其它数据，并包括任何信息传送介质。术语“已调制数据信号”指以对信号中的信息进行编码的方式设置或改变其一个或多个特征的信号。作为示例而非限制，通信介质包括有线介质，如有线网络或直接连线连接，以及无线介质，如声学、RF、红外线和其它无线介质。如此处所使用的术语计算机可读介质包括存储介质和通信介质两者。

快速点击距离确定的说明性实施例

包括计算设备 100 的系统所执行的实施例允许确定网络上包括的特定文档的点击距离。如此处所述确定的点击距离随后可用于改进搜索引擎所产生的搜索结果的排序。

除了允许此处的公开和权利要求涵盖更广泛的各种替换实施例的更宽泛的含义以外，如此处所使用的且在权利要求书全文中使用的以下术语一般如下定义：

“锚文本”一般被定义为与源文档中包括的引导至目标文档的导航的链接相关联的文本。为搜索查询的目的，包括在源文档中的锚文本提高了目标文档的排序。例如，当查询匹配某一锚文本中的字词时，该锚的目标文档在相关性排序中得以提升。

“点击距离”一般指的是在两个文档（例如，*文档 1* 和*文档 2*）之间导航所需的点击或链接的最少数目。如果*文档 1* 是通常认为的网页，具有高度可靠性（“权威性”），则如果这两个文档之间的点击距离较小，*文档 2* 的可靠性就得以提升。页面的可靠性（或权威性）程度是可在搜索引擎的排序算法中使用的查询无关相关性度量。

“文档”一般指的是可作为网络搜索查询或爬寻的结果返回的任何可能的资源，诸如网络文档、文件、文件夹、网页或其它资源。

本发明的实施例一般涉及产生点击距离确定，使得点击距离确定可用于改进搜索引擎的排序函数。搜索引擎的质量一般由根据排序函数所分配的排序的文档相关性来确定。排序函数可基于多个特征。某些特征可依赖于查询，而其它特征被认为是查询无关的。点击距离是从主页或权威 URL 到给定页面的查询无关相关性度量。在 web 图（见以下的图 3 和 4）上，点击距离可被表示为权威页面与给定页面之间的最短路径。在之前的实施例中，一算法执行了广度优先遍历，并计算了给定节点到图中所有其它节点的距离。遍历取 N 次迭代来计算，其中 N 是图的直径（最短距离中的最大值）。变量 N 可比图上的节点总数小。

图 2 示出了用于排序搜索结果的系统，它可包括根据本发明的用于快速确定点击距离的功能性。搜索引擎 200 接收包含多个查询项的查询。每一查询项可包括多个分量项，诸如当查询项是短语（例如，短语“文档管理系统”可被认为是单个查询项）时。此外，查询可包括一个或多个运算符，诸如布尔运算符、约束条件等，它们通常为已知搜索引擎所支持。

分布式网络上的多个文档，由文档 210、212、214 和 216 表示，可供搜索使用。实际上，搜索引擎可搜索任何数目的文档，且一般搜索包含大量（例如，数百万）文档的集合。从因特网设置到内联网设置，文档的数量会有所减少，但是减少通常是从数十亿到数百万，使得文档的相对数目仍是相当大的。索引模块（未示出）为每一文档生成独立文档统计（例如，218、220、222 和 224）。文档统计被存储在索引 226 中。

搜索引擎 200 查阅索引 226 来基于查询和相应的文档统计为每一文档确定搜索评分 228。在本发明中，所包括的文档统计之一是文档的点击距离。在另

一实施例中，所包括的另一文档统计是与文档相关联的 URL 深度。点击距离和 URL 深度然后与查询相关统计组合来形成文档的最终评分。一般，文档评分 228 然后以降序排序以向用户给出搜索算法认为与查询最相关的文档列表。

在所示系统中，搜索引擎 200 表示点击距离排序搜索引擎，它在确定文档的搜索评分时考虑文档的点击距离。在一个实例中，自主页的点击距离是页面重要性的度量，其中分层结构中较接近权威的页面被认为比分层结构中较低的页面重要。然而，可能存在其中相反情况成立的其它场景，即在其中分层结构中较低的文档被认为比分层结构中较高的那些页面重要。由此点击距离被认为是查询无关相关性度量，因为它在总体上而非根据查询（例如，查询相关排序函数可对搜索项出现在文档中的次数进行计数）来评估文档的重要性。

图 3 示出了根据本发明的示例性网络图。网络图由节点（例如，310）和边或链接（例如，320）组成。节点（例如，310）表示页面和位于网络上可作为搜索查询的结果返回的其它资源。链接（例如，320）通过使用页面上列出的导航链接将这些页面中的每一个连接在一起。可为每一页面收集链接信息集，它们可用于计算特定页面的点击距离。

在一个实施例中，节点 330 表示一组文档中的最高权威页面即网络上的根节点。网络中其余页面的点击距离可从节点 330 计算。例如，节点 340 具有自节点 330 的两个“点击”的点击距离。如上所述，“点击”指的是在自最高权威节点的最短路径上遍历的分支的数目。可选择自节点 330 的其它路径来达到节点 340，但点击距离与最短路径有关。

以不遵循特定次序的节点示出了网络图 300，在这一方面上类似于因特网。因为缺乏次序，点击距离对排序页面的适用性可能难以概念化。然而，通常页面和资源的网络的确遵循应用的次序，如以下图 4 中所示。

图 4 示出了根据本发明的示例性分层结构网络图。分层结构网络图 400 就它也包括节点（例如，410）和链接（例如，420）这一点类似于图 3 中所示的网络图 300。然而，分层结构网络图 400 是基于结构化的站点或内联网的固有分层结构的。因此分层结构网络图 400 可被概念化为具有从根节点延伸的分支的树结构。

对分层结构网络图 400，点击距离的适用性和计算更可认识。例如，节点

330 对应于树的最高权威节点即根节点。节点 340 因而具有相关联的点击距离 3，离根节点 3 次点击或用户导航远。换言之，由于要求用户遍历树的 3 个分支来从节点 330 导航至节点 340，因此点击距离也为 3。

图 3 和 4 中所表示的网络图是在索引文档用于计算点击距离期间在存储器中构造的图的示例。在索引期间构造图使得点击距离能够被包括在索引中所存储的文档统计中并用于排序页面。

图 5 示出了根据本发明用于索引文档的示例性系统的功能框图。系统 500 包括索引 510、流水线 520、文档接口 530、客户机接口 540、锚文本插件 550、索引插件 560 和锚文本表 70。

索引 510 被结构化为包括单独的索引分区，包括主分区和用于锚文本的另一分区。在另一实施例中，除由索引 510 表示的倒排索引以外，提供单独的锚文本索引。以下在图 6 的讨论中提供对索引 510 的结构的更详细描述。使用这些索引的记录来对客户机查询提供结果。在一个实施例中，索引 510 对应于共同提供索引记录的存储的多个数据库。

流水线 520 是用于获取文档或文档记录以便索引的收集机制的说明性表示。流水线 520 使得对应于数据的记录被输入到索引 510 之前能够由各个插件（例如，锚文本插件 550）对数据进行过滤。

文档接口 530 提供协议、网络接入点和数据库接入点以跨多个数据库和网络位置检索文档。例如，文档接口 530 可提供对因特网的访问，同时也提供对本地服务器的数据库的访问和对当前计算设备上的数据库的访问。其它实施例可使用各种协议来访问其它文档位置，而不背离本发明的精神或范围。

客户机接口 540 提供由客户机进行的访问以定义和启动搜索。搜索可根据关键词、索引键和/或“范围键”来定义。范围键指的是用于进一步缩小搜索查询的范围的字词。例如，范围键可与特定的文件类型相关联。使用该范围键作为搜索项的搜索将搜索结果的范围限于对应于该文件类型的文档。采用范围键，搜索范围可根据诸如文件类型的属性、诸如某些数据库或 URL 的位置或按照减少要搜索的文档数目的其它准则而缩小。

锚文本插件 550 是若干收集器流水线插件之一。锚文本插件 550 标识包括在文档中的锚文本及其相关特性。锚特性由锚文本插件 550 在爬寻通过文档接

口 530 提供的文档时收集。在一个实施例中，锚文本插件 550 的功能性实际上被包括特性插件中而非作为单独插件提供。特性插件标识文档的所有域及其相关联的特性，包括锚特性。在一个实施例中，由于锚文本与目标文档相关联，因而将目标文档与锚文本相关联推迟到爬寻完成。例如，当索引文档 A，且文档 A 具有指向文档 B 的锚文本时，该锚文本被应用于文档 B。但由于此时文档 A 正被索引，该过程被推迟。而且，可能存在要应用于文档 B 的多个锚，要求在正确索引文档 B 之前发现它们。将目标文档的索引推迟到爬寻完成之后改善了索引结果的正确性，但不是可用的唯一方法。

索引插件 560 是连接至流水线 520 的另一插件。索引插件提供用于生成、分区和更新索引 510 的机制。在一个实施例中，索引插件 560 提供在将关键词和从所爬寻的文档生成的锚文本键刷新到索引 510 之前临时高速缓存这些结果的字词列表。从包括在这些字词列表中的爬寻结果填充索引 510 的记录。

锚文本表 570 包括已由锚文本插件 550 收集的锚特性。作为文档中的锚文本的实例，锚文本表 570 包括锚文本和与该锚文本相关联的特性的记录。在所示示例中，锚文本表 570 中的记录可在各个域中包括标识当前文档的源 ID 572、标识链接的目标文档的目标 ID 574、锚文本条目 576 以及链接 578。在其它实施例中，可在锚文本表 570 中包括其它域。

为了实现快速点击距离确定，将专门的字词（例如，580）或范围键追加到锚文本条目 576 中的锚文本。该专门字词（例如，580）提供对锚文本表 570 的记录中所包括的目标文档的源文档指定。检查锚文本表 570 的第一记录，专门字词为指定锚文本的源为文档 A 的“文档 A”。在一个实施例中，每一文档由文档 ID 标识。因此该专门字词是提供独特字词以添加到将目标文档与源文档相关联的锚文本索引的文档 ID 的变型。

采用被添加到锚文本表 570 的特性，从爬寻收集的锚和链接特性可用于生成网络或锚图的表示，其节点对应于文档而分支对应于链接（见图 4）。该锚图然后可被加载到存储器中，用于解决快速点击距离确定。

尽管在系统 500 中示出了功能块之间的单向和双向通信，但这些通信类型中的任何一种可被改变成另一类型，而不背离本发明的精神或范围（例如，所有的通信可具有要求双向通信而非单向通信的确认消息）。

图 6 示出了根据本发明的索引的示例性结构的功能框图。索引 600 包括主索引 610 和锚文本索引 620。在一个实施例中，索引 600 被认为相对于与搜索引擎查询过程相关联的其它数据结构是“本地存储”的。在该实施例中，文档的语料库驻留在网络上，数据收集模块、插件（见图 5）和插件数据结构驻留在服务器上，仅索引 600 驻留在本地存储位置中。这使得对索引 600 的查询比对语料库或其它数据结构的查询要高效得多。

主索引 610 包括对应于关键词和对应于文档爬寻而返回的其它索引键的记录。主索引 610 也包括涉及文档其它特性的其它索引分区。对此的记录对应于被转向并被输入到锚文本索引 620 中的锚文本。

一般，锚文本索引 620 包括对应于网络上文档中所包括的锚文本的目标文档的记录。这些目标文档被组织成与包括在锚文本或关联于目标文档的 URL 中的字词相关联列出目标文档 ID 的倒排索引。在爬寻完成之后，从锚文本表中生成锚文本索引 620。对应于每一目标文档的锚文本级联在一起以便于对各项评估每一目标文档并将目标文档输入到锚文本索引 620 中。包括用于锚文本的单独索引分区允许在将锚文本作为文档的评分函数中的因子并入之前基于该锚文本来进行相关性计算。将在以下图 6 的讨论中更全面地描述将锚文本并入用于排序文档的评分函数中。

为快速点击距离确定的目的将追加的专门字词包括在锚文本索引 620 中，锚文本索引 620 的记录也包括对应于该专门字词的记录。采用图 5 中所示的示例锚文本表 570，可在锚文本索引 620 中包括以下记录：

字词

B、C、D、E、F、G

文档 A

B、C、D

文档 B

E、F

文档 F

G

该示例示出了如何对用于将源文档与目标文档链接的锚文本“字词”列出

目标 ID，也对专门字词提供目标 ID。专门字词的这一列表在锚文本索引 620 自身内将源文档与目标文档相关联。由于关联是建立在锚文本索引 620 内的，因而无需反复访问锚文本表来确定每一文档的点击距离。相反，可对锚文本索引 620 执行广度优先遍历以更快确定点击距离。以下关于图 7 更详细描述了用于确定索引中每一文档的点击距离的示例性过程。为了合乎比例地安置此处所述的实施例的速度和效率的增加，在一个示例中，确定点击距离的先前实现花费长达五小时来完成。采用当前实施例，同一确定花费约三十秒。这种速度和效率的惊人增加使得点击距离成为排序搜索引擎查询结果的非常有用的因素。

图 7 示出了根据本发明用于确定点击距离（CD）的示例性过程的逻辑流程图。过程 700 从框 702 开始，其中分布式网络上的文档已被索引，且锚图的生成已被启动。生成锚图的过程被称为锚爬寻。在一个实施例中，锚爬寻对由其中收集链接和锚文本信息并将其置于如以上图 5 中所述的锚文本表中的过程收集的数据进行爬寻。处理在框 704 处继续。

在框 704，将初始锚图加载到存储器中。完成后的锚图对应于从网络收集的文档标识（例如，文档 ID）和链接信息的结构化表示。可对应于锚图的网络图的示例在图 3 和 4 中示出。锚图包括对应于网络的文档的节点和对应于文档之间的锚或链接的边。处理在框 706 处继续。

在框 706 处，在初始化后锚图中的父节点的点击距离（CD）值也被初始化。这些父或最高权威节点被称为已赋值节点。这些节点被赋予点击距离值 0（零）。可以对单个锚图指定多于一个的高权威节点。例如，管理员可手动对一组一百个节点进行排序，并将它们全部指定为高权威节点。另外，高权威节点的点击距离不必为 0（零），管理员可赋予任何数字。改变高权威节点的点击距离不会更改其余算法，而仅提供手动指定节点的重要性的方法。例如，管理员可提高某些节点的点击距离评分。在其它情况中，管理员可降低点击距离评分（通过使点击距离高于算法默认计算出的值）。每一未赋值节点的点击距离被初始化为最大值。在一个实施例中，最大值实质上将点击距离值设为无穷大。对节点赋予无穷大值使得它可容易地被识别为还未计算其点击距离的节点。当已赋值节点的点击距离值的初始化完成时，处理移动至框 708。

在框 708，将具有不同于最大值的相关联点击距离的节点插入到队列中。

在一个示例中，该步骤仅在第一次迭代中发生。插入到队列中的节点对应于最高权威节点，因为它们的点击距离被置为 0（零），一个不同于最大值的值。一旦具有不同于最大值的点击距离值的节点被添加到队列中，处理在判定框 710 处继续。

在判定框 710 处，作出队列是否为空的判断。空队列表明没有其它节点需要计算其目标节点的点击距离。如果队列为空，则处理移动至框 712，在那里过程 700 结束。然而，如果队列不为空，则处理在框 714 处继续。

在框 714 处，从队列中检索一节点，并确定作为锚的目标节点的一组节点的判断。此处所述的实施例使得该判断可被高效且快速地处理。代替对锚文本表进行迭代查询，可对锚文本索引进行简单的查询。锚索引非常高效地解析某一类型的查询。这种类型的查询可被描述为要求“返回相关联锚文本包含*字词 X* 的所有文档”的查询，其中*字词 X* 表示单个字词。对点击距离确定，执行同一类型的查询。然而，在对锚文本索引的点击距离查询中，该字词被称为专门字词，它已被追加到锚文本。专门词语对应于该锚文本的源文档。在锚文本索引中，每一专门字词包括与该专门字词相关联的目标文档的清单。例如，参考以上在图 6 的讨论中描述的示例，表述“返回相关联锚文本包含*文档 A* 的所有文档”的查询返回表示源文档 A 的目标文档的目标 ID B、C 和 D 的列表。再一次，利用本地存储的锚文本索引允许有效率得多的点击距离确定。由于锚文本是本地存储的，用于跨网络通信的通信过程是不必要的。此外，该过程不是迭代的。与处理由锚文本表提供的结构不同，仅需要一次对锚文本索引的查询来返回与从队列中检索出的节点相关联的所有目标节点（即，目标文档）。一旦该节点被检索出，且目标节点被确定，处理移动至框 716。

在框 716，取回下一目标节点。下一目标节点指的是由初始文档所链接的文档中的下一文档。一旦取回下一目标节点，处理继续至判定框 718。

在判定框 718，作出与目标节点相关联的点击距离是否大于当前页面的点击距离加一（CD + 1）的判断。在一个实施例中，仅当目标节点具有无穷大的点击距离（假定高权威节点被置为零，且管理员未手动设置点击距离）时才满足框 718 中的条件。例如，如当前点击距离为 1，则 CD + 1＝2。点击距离 2 小于无穷大，该条件满足。判断目标点击距离是否大于点击距离加一能够避免

改变具有较小点击距离的目标文档。使用之前的示例，如果目标节点的点击距离为 1，且当前点击距离也为 1，则目标点击距离不大于 CD + 1＝2。在这种情况下，已经记录了至目标节点的较短路径，从而不需要被更新。相应地，当目标点击距离不大于当前点击距离加一时，处理前进至判定框 722。然而，如果目标点击距离大于当前点击距离加一，则处理移动至框 720。

在框 720，更新目标节点的点击距离值，且将该目标节点添加至队列作为需要进行其目标的点击距离计算的节点。以新的点击距离值更新该目标节点以移除无穷大值，并将节点设置为计算出的点击距离值。在一个实施例中，节点的点击距离值被置为当前点击距离加一（CD + 1）。处理在判定框 722 处继续。

在判定框 722 处，作出是否已对从队列中检索出的当前节点取回了所有目标节点的判断。如果有要对当前节点取回的剩下的目标节点，则处理返回至框 716，其中取回下一目标节点。然而，如果已经取回了对应于当前节点的所有目标节点，则处理返回至判定框 710 来重新检查队列现在是否为空。再一次，一旦队列为空，则处理移动至框 712，在那里过程 700 结束。

可以理解，在过程 700 中描述的操作框可按需重复以对网络上的每一文档赋予点击距离值。有可能不是网络上的所有节点都通过任何其它节点被连接至初始高权威节点。相应地，在本发明的另一实施例中，未以任何方式连接至高权威节点的节点被假定具有低重要性，且被赋予低于锚图的平均的点击距离。

采用根据此处所述的快速确定过程确定的每一文档的点击距离，点击距离则可按需在用于改进响应于查询的网络上文档的排序结果的任何评分或排序函数中使用。当执行评分函数且计算文档的相关性评分时，评分现在部分反映文档的点击距离值。

尽管用结构特征和/或方法步骤专用的语言描述了本发明，但可以理解，所附权利要求书中定义的本发明不必限于所述的特定特征或步骤。相反，特定特征和步骤被公开为实现所要求保护的本发明的各形式。由于可在不背离本发明的精神和范围的情况下作出本发明的众多实施例，因此本发明驻留在所附的权利要求书中。

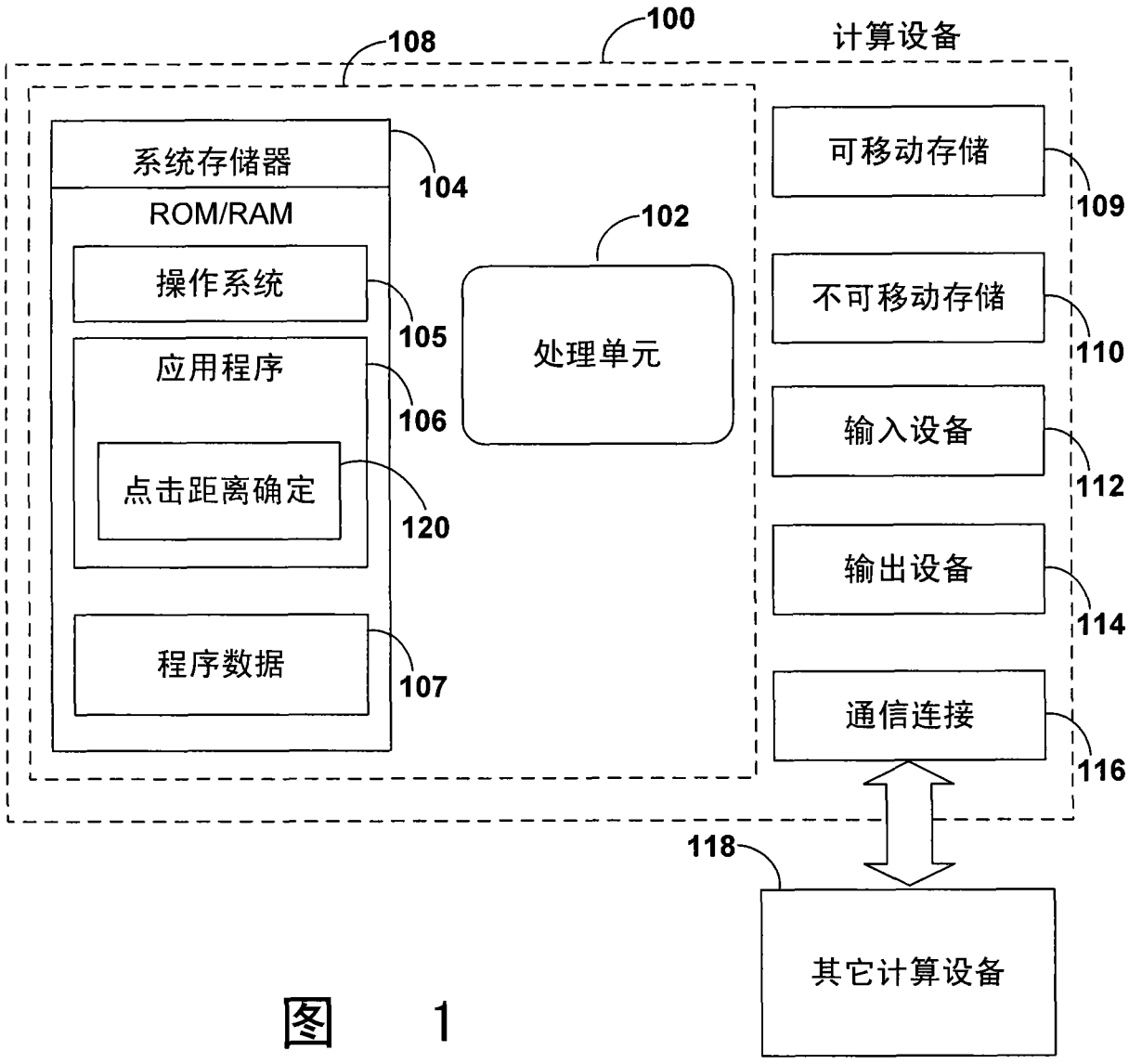

100
计算设备
108
系统存储器
104
ROM/RAM
操作系统
105
应用程序
106
点击距离确定
120
程序数据
107
102
处理单元
可移动存储
109
不可移动存储
110
输入设备
112
输出设备
114
通信连接
116
118
其它计算设备

图 1

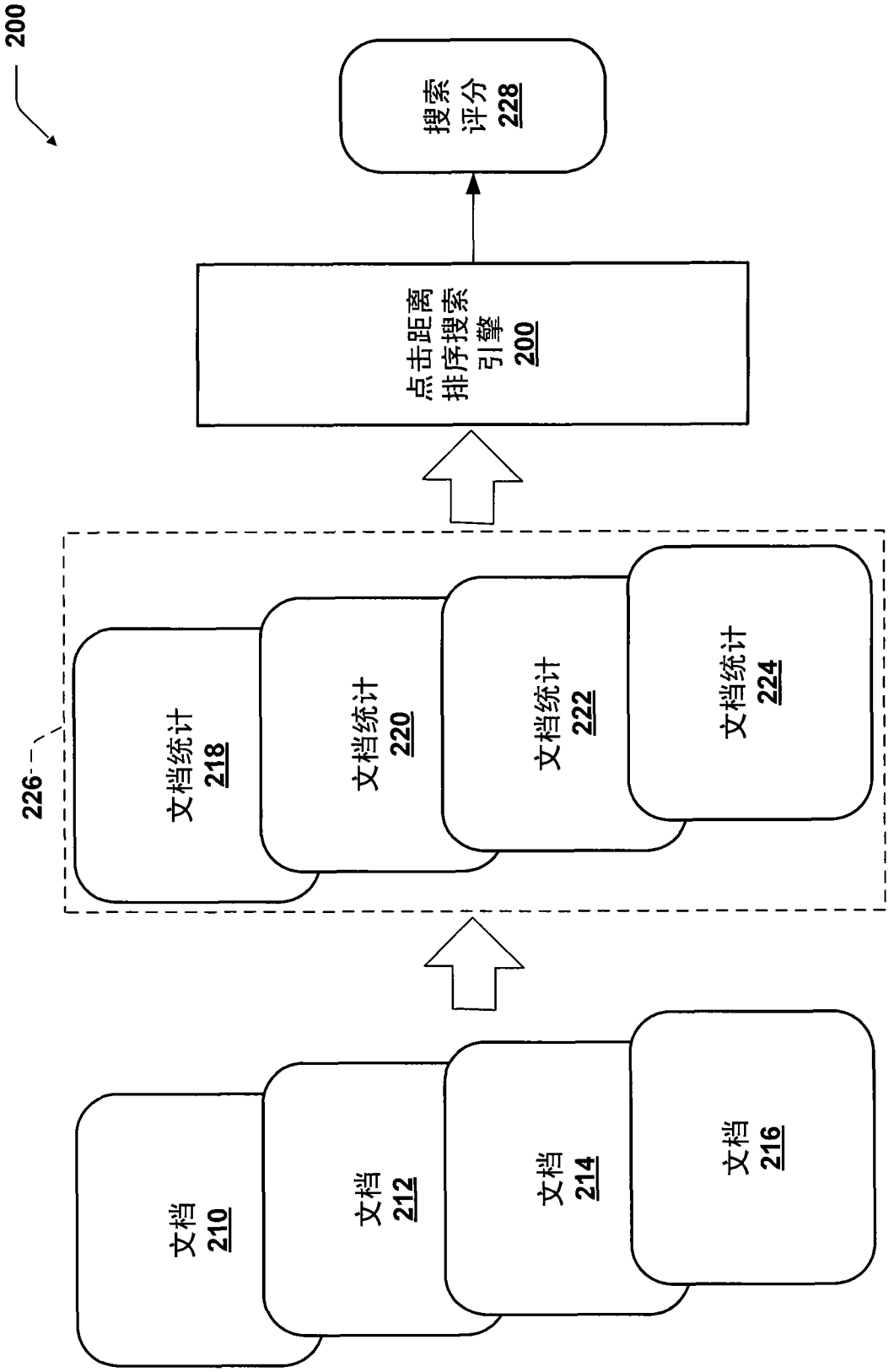

200
搜索评分 228
点击距离排序搜索引擎 200
226
文档统计 218
文档统计 220
文档统计 222
文档统计 224
文档 210
文档 212
文档 214
文档 216

图 2

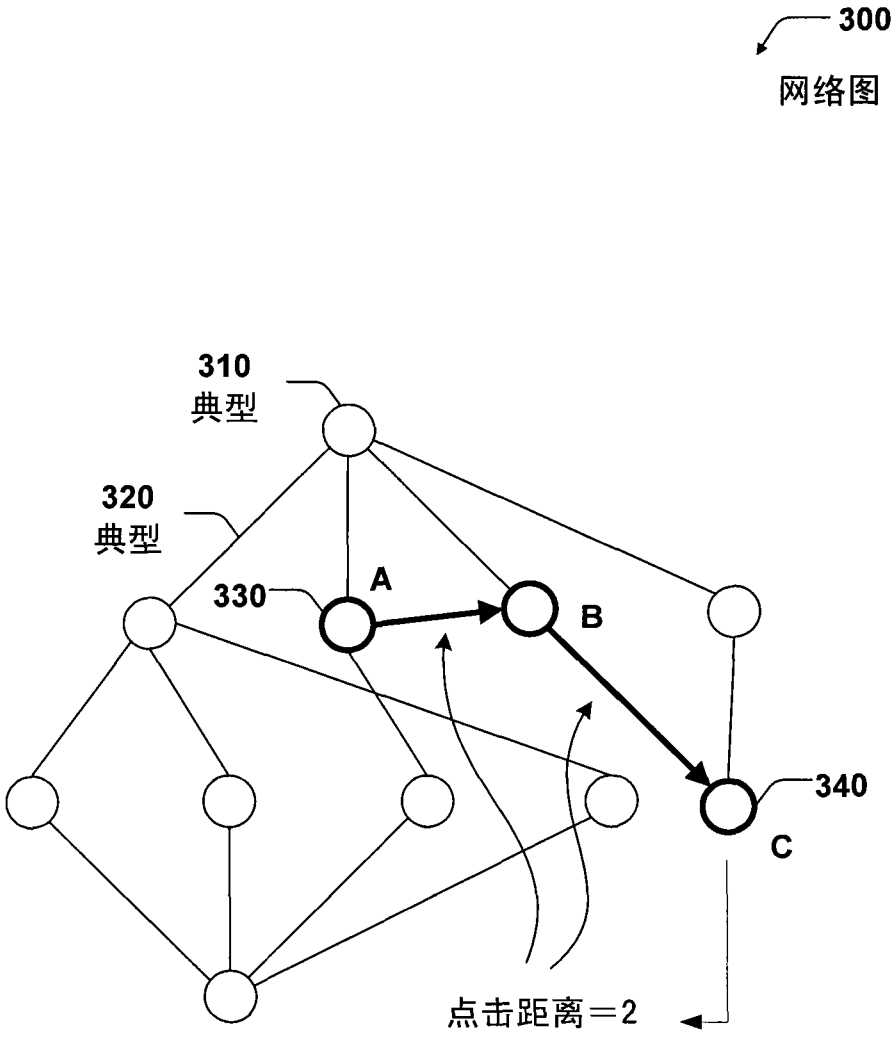


图 3

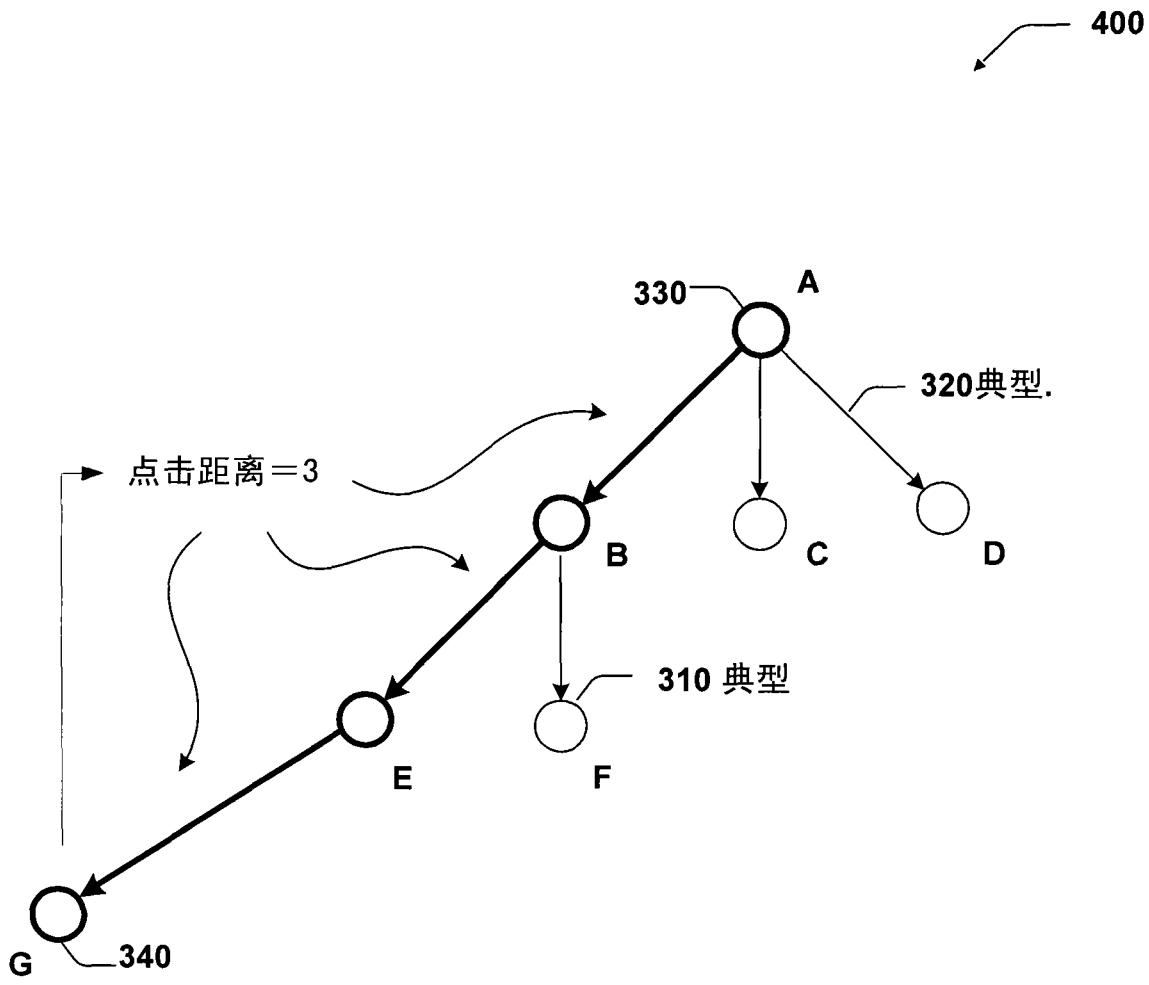
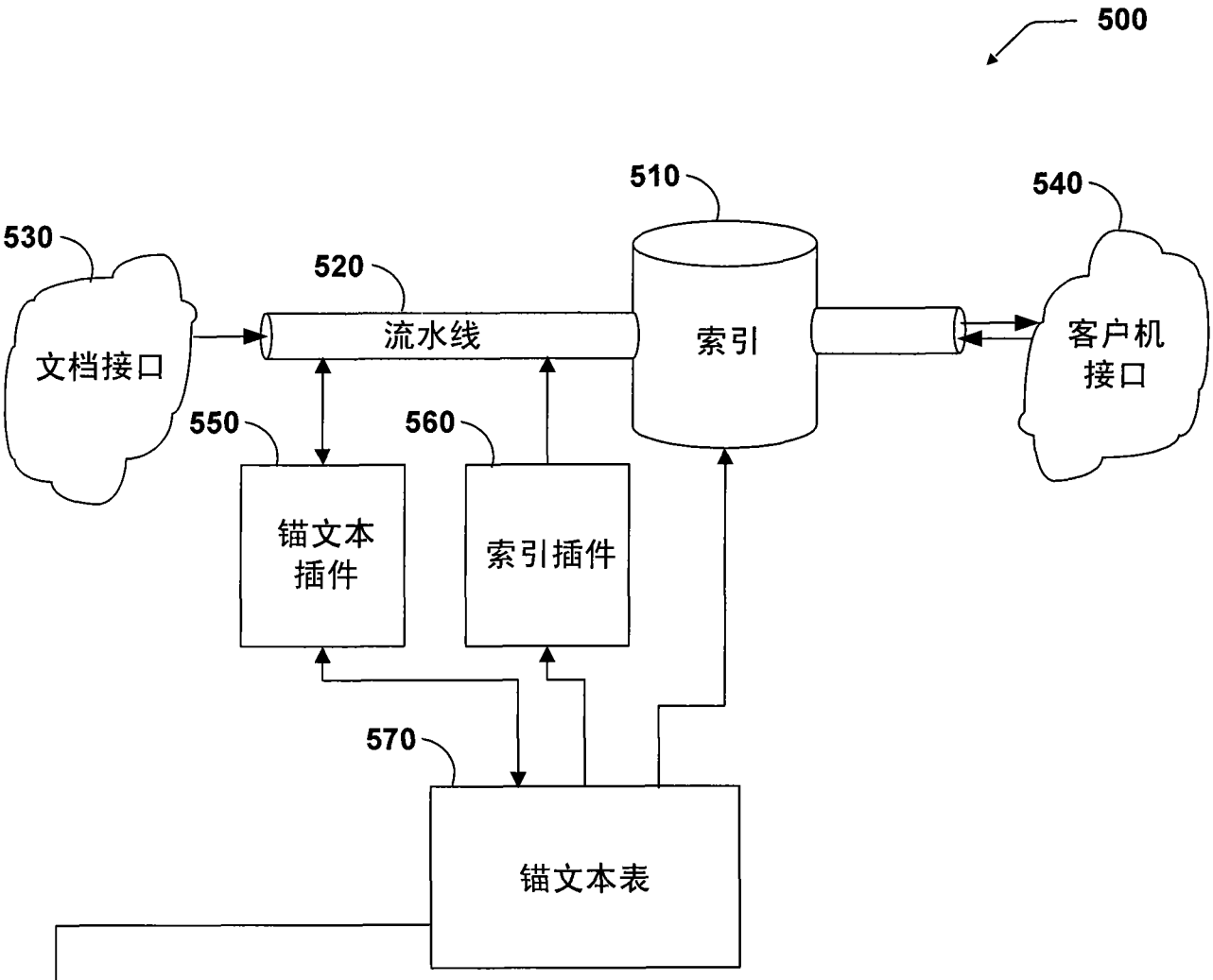


图 4



572 源ID	574 目标ID	576 锚文本 (580 typ.)
A	B	WORD 文档A
A	C	WORD 文档A
A	D	WORD 文档A
B	E	WORD 文档B
B	F	WORD 文档B
E	G	WORD 文档E

图 5

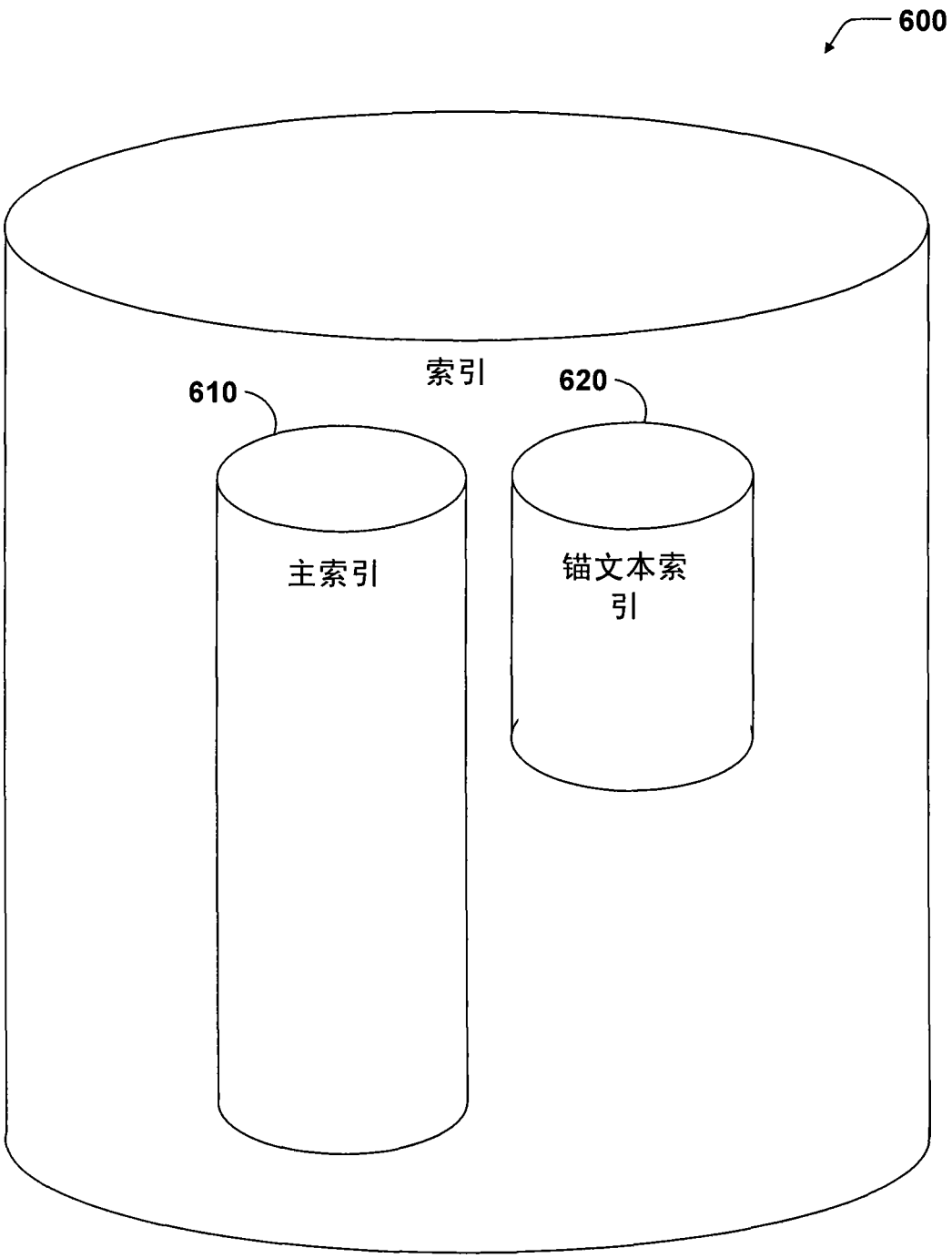


图 6

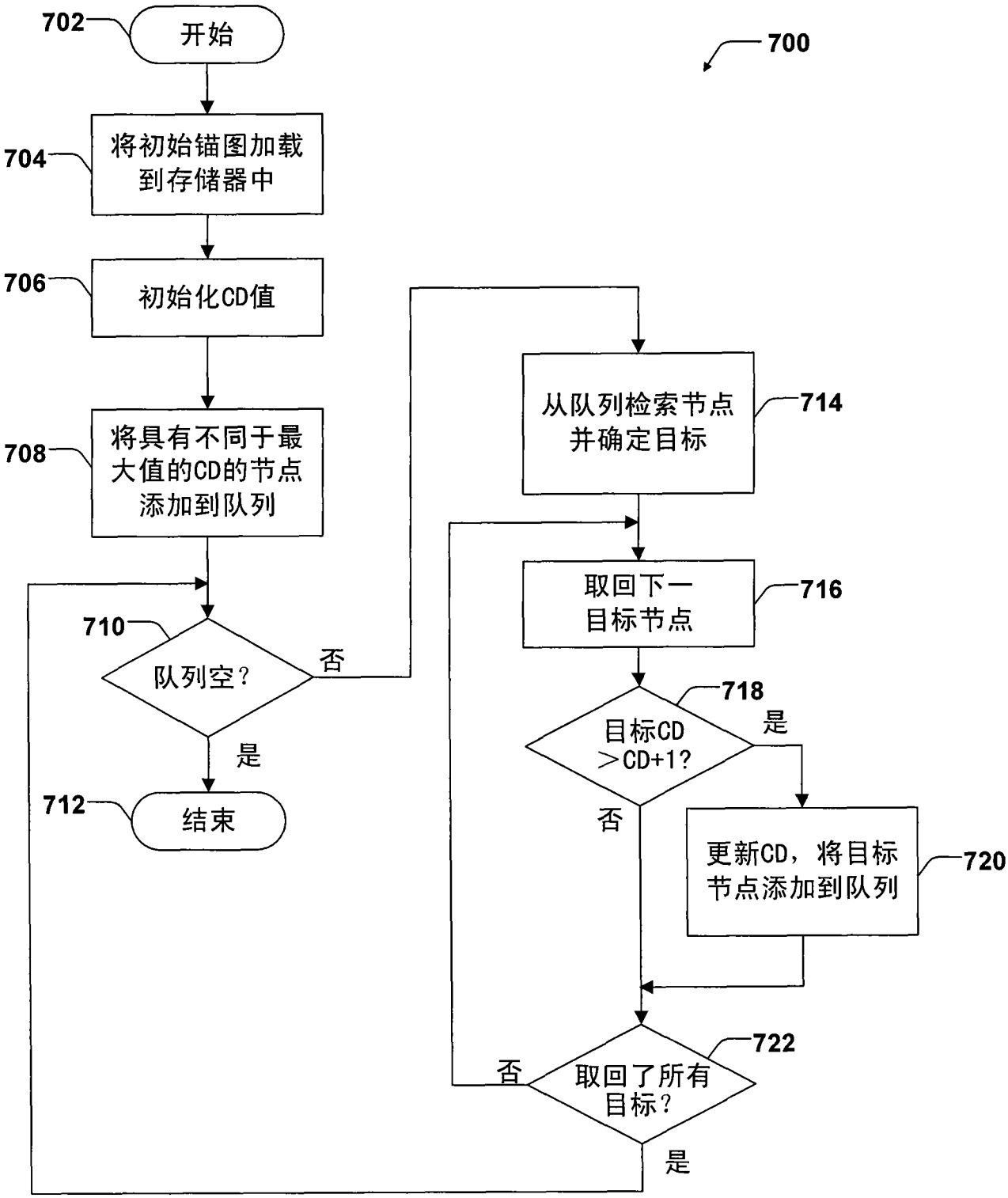

702
开始
700
704
将初始锚图加载到存储器中
706
初始化CD值
708
将具有不同于最大值的CD的节点添加到队列
710
队列空？
否
是
712
结束
714
从队列检索节点并确定目标
716
取回下一目标节点
718
目标CD＞CD+1?
是
否
720
更新CD，将目标节点添加到队列
722
取回了所有目标？
否
是

图 7