(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 101268464 B
(45)授权公告日 2011.03.09

(21)申请号 200680034531.6
(22)申请日 2006.09.20
(30)优先权数据
11/231,955 2005.09.21 US
(85)PCT申请进入国家阶段日
2008.03.20
(86)PCT申请的申请数据
PCT/US2006/037206 2006.09.20
(87)PCT申请的公布数据
WO2007/035919 EN 2007.03.29
(73)专利权人 微软公司
地址 美国华盛顿州
(72)发明人 D·梅耶泽 H·扎拉格扎
K·佩顿纳 A·德伯鲁纳
(74)专利代理机构 上海专利商标事务所有限公司 31100
代理人 陈斌
(51)Int.Cl.
G06F 17/30(2006.01)
审查员 王亮

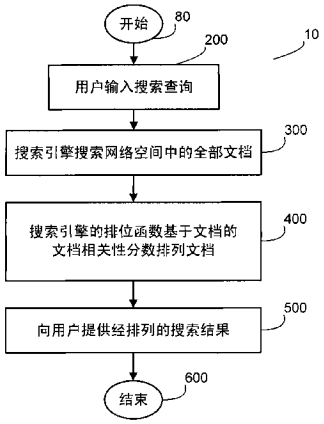
权利要求书 2 页 说明书 11 页 附图 4 页

(54)发明名称

使用文档使用统计量的排位函数

(57)摘要

揭示了向网络上文档提供文档相关性分数的方法。还揭示了具有计算机可执行指令存储在其上以执行向网络上文档提供文档相关性分数的方法的计算机可读介质。此外还揭示了包含至少一个应用程序模块的计算系统，其中该至少一个应用程序模块包括用于执行向网络上文档提供文档相关性分数的方法的应用程序代码。



CN 101268464 B

1. 一种用以确定网络上文档的文档相关性分数的方法，包括：

利用由文档的用户查询所调用的包含一或多个查询无关分量的排位函数来对网络上文档进行排位，其中至少一个查询无关分量包括考虑到网络上一或多个文档的、服务器生成的、服务器存储的使用数据的使用参数，所述使用数据由网络存储系统生成和存储，所述网络存储系统管理和向用户提供对所述文档的网络访问，所述使用数据包括对实际用户经由网络存储系统与网络上一个或多个文档交互的计量，使得文档的所述使用数据反映了所述文档被多个用户通过所述网络存储系统的查询无关交互使用，指派由所述排位函数生成的分数以排位所述网络上的文档，所述分数用于按顺序排位文档，

接受用户输入的包括搜索串的搜索询问，进行对所述网络上的文档的搜索以生成包括多个文档的搜索结果，使用所述排位函数排位所述搜索结果的多个文档以生成经排位的搜索结果，使得所述文档既根据它们各自的使用数据也根据它们与搜索串的相关性进行排位，并向所述用户提供所述经排位的搜索结果。

2. 如权利要求 1 所述的方法，其特征在于，文档的使用值包括 (i) 基于服务器维护的实际使用数据的实际使用值或者 (ii) 不是基于实际使用数据的默认使用值。

3. 如权利要求 1 所述的方法，其特征在于，所述至少一个查询无关分量由下列公式表示：

$$QID(doc) = w_u \frac{k_u U}{k_u + U}$$

其中：

U 表示实际使用值或默认使用值；以及

w_u 和 k_u 表示所述使用值的调整参数。

4. 如权利要求 1 所述的方法，其特征在于，所述至少一个查询无关分 量包括以下两者：(i) 所述使用参数以及 (ii) 点击距离或者经偏移的点击距离参数。

5. 如权利要求 1 所述的方法，其特征在于，所述至少一个查询无关分量包括所述使用参数和 URL 深度参数两者。

6. 如权利要求 1 所述的方法，其特征在于，所述每一文档的分数是使用以下公式生成的：

$$Score = \sum \frac{wtf'(k_1+1)}{k_1 + wtf'} \times \log(\frac{N}{n}) + w_{cd} \frac{k_{cd}}{k_{cd} + \frac{b_{cd}\frac{CD}{k_{ew}} + b_{ud}UD}{b_{cd} + b_{ud}}} + w_u \frac{k_u U}{k_u + U}$$

其中：

wtf′ 表示加权的项频率，

N 表示所述网络上文档的数量，

n 表示包含查询项的文档数量，

w_{cd} 表示查询无关分量的权重，

b_{cd} 表示点击距离的权重，

b_{ud} 表示 URL 深度的权重，

CD 表示文档的计算出的点击距离或者被指派的经偏移的点击距离，

k_{ew} 表示与边界权重相关的调整常数，

UD 表示 URL 深度，

U 表示实际使用值或默认使用值，

w_u 和 k_u 表示所述使用值的调整常数，以及

k_{cd} 和 k_1 是常数。

7. 如权利要求 1 所述的方法，其特征在于，还包括使管理员能够手动调整由所述排位函数生成的排位结果。

8. 一种用以确定网络上文档的文档相关性分数的方法，包括：

利用包含一或多个查询无关分量的排位函数来对网络上文档进行排位，其中至少一个查询无关分量包括考虑到网络上一或多个文档的、服务 器生成的、服务器存储的使用数据的使用参数，所述使用数据包括对实际用户与网络上一个或多个文档交互的计量，指派由所述排位函数生成的分数以排位所述网络上的文档，所述分数用于按顺序排位文档，其中，所述使用数据包括各文档的使用值，一个文档的所述使用值依赖于文档或包含文档集合的文件夹的一或多个使用相关性质，所述一或多个使用相关性质包括在给定时间段内用户的文档或文件夹查看总数、在给定时间段内每用户的文档或文件夹查看平均数、在给定时间段内在特定文档或文件夹上花费的总时间、在给定时间段内在特定文档或文件夹上花费的平均时间，其中所述给定时间段包括上个星期内、上个月、去年一年内、所述文档或文件夹的生存期内或者任何其它时间段。

使用文档使用统计量的排位函数

背景技术

[0001] 按照文档与给定搜索查询的相关性排列文档的排位函数是已知的。在本领域中仍在努力开发针对给定搜索查询提供优于由使用已知排位函数的搜索引擎产生的搜索结果的搜索结果的排位函数。

发明内容

[0002] 在此主要描述用于在网络上确定给定文档的文档相关性分数的各种技术。文档相关性分数是通过排位函数产生的，该排位函数包括一或多个查询无关分量，其中至少一个查询无关分量包括使用参数，该参数考虑在 web 服务器上维护与存储的、用于网络上一或多个文档的实际文档使用数据。排位函数可由搜索引擎使用，以基于多个文档的文档相关性分数按序（通常按降序）排列多个文档。

[0003] 提供本发明内容，它以简化的形式向读者一般地介绍在下面“具体实施方式”中描述的一或多个选择的概念。本发明内容不是要标识要求保护主题的关键和 / 或必要特征。

附图说明

[0004] 图 1 表示示出一方法的示例性步骤的示例性逻辑流程图，该方法响应于用户输入的搜索查询产生经排列的搜索结果；

[0005] 图 2 是一些用于实现在此揭示的方法和过程的示例性操作环境的主要组件的框图；

[0006] 图 3 表示示出一示例性方法的示例性步骤的逻辑流程图，该方法用于确定网络上文档的文档相关性分数；以及

[0007] 图 4 表示示出一方法的示例性步骤的逻辑流程图，该方法使用包含文档使用参数的排位函数排列生成的搜索结果。

具体实施方式

[0008] 为加强对在此揭示的方法和过程的原理的理解，使用下面的特定实施例的描述和特定语言来描述这些特定实施例。然而将会理解，使用特定语言不是要限制所揭示方法和过程的范围。对所讨论的揭示方法和过程的原理的改变、进一步修改以及这类进一步应用，对于被揭示方法和过程所属领域的技术人员而言，都是在正常的预期范围内的。

[0009] 揭示了确定网络上文档的文档相关性分数的方法。每一文档相关性分数是使用排位函数计算的，希望该排位函数包含一或多个查询无关分量（例如，不依赖于给定搜索查询或搜索查询项的函数分量），一或多个查询相关分量（例如，依赖于给定搜索查询或搜索查询项的细节的函数分量），或者两者的组合。由排位函数确定的文档相关性分数可用于按照每一文档相关性分数排列网络空间（例如公司内联网空间）内的文档。可使用揭示方法的示例性搜索过程示出作为在图 1 中的示例性过程 10。

[0010] 图 1 描绘示例性搜索过程 10，它从过程步骤 80 开始，其中用户输入搜索查询。从

步骤 80，示例性搜索过程 10 进行至步骤 200，其中搜索引擎为一或多个搜索查询项搜索网络空间内的所有文档。从步骤 200，示例性搜索过程 10 进行至步骤 300，其中搜索引擎的排位函数基于每一文档的相关性分数排列网络空间内的文档，而文档相关性分数则基于一或多个查询无关分量、一或多个查询相关分量或两者的组合。从步骤 300，示例性搜索过程 10 进行至步骤 400，其中向用户呈现经排列的搜索结果，这一呈现通常按照降序以标识网络空间内与搜索查询最相关的文档。

[0011] 如下面更详细地讨论的，在确定文档相关性分数的一些示例性方法中，用于确定文档相关性分数的排位函数的至少一个查询无关分量考虑与网络空间内一或多个用户对一或多个文档实际使用相关的“文档使用数据”或“文档使用统计量”。文档使用数据和 / 或统计量是由独立于给定搜索引擎的 web 服务器上的应用程序代码生成和存储的。例如，文档使用数据可由网站维护，使得每当用户请求 URL 时，服务器就更新使用计数。使用计数器可以维护在给定时间间隔获得的文档相关数据，这一给定时间间隔诸如可以是上个星期内、上个月、去年一年内、或者给定文档或文档集合的生存期内。应用程序代码可用于通过 (i) 特殊的应用编程接口 (API)，(ii)web 服务请求，或者 (iii) 请求返回网站上每一 URL 的使用数据的管理网页来从网站获得使用数据。

[0012] 特定的网站可用于生成和维护网络空间内的使用数据，并且在本地或远程存储系统中存储使用数据。用于生成、维护和存储网络空间内文档的使用数据的合适网站包括但不限于 WINDOWS ® SHAREPOINT ® Services（服务）站点。

[0013] 所揭示的用以确定文档相关性分数的方法还可使用包括一或多个附加查询无关分量的排位函数。合适的附加查询无关分量包括但不限于在 2004 年 8 月 30 日提交的题为“SYSTEM AND METHOD FOR RANKINGSEARCH RESULTS USING CLICK DISTANCE（使用点击距离排位搜索结果的系统和方法）”的美国专利申请序列号 10/955,983 中描述的、考虑网络空间内每一文档的点击距离的查询无关分量，在 2005 年 8 月 15 日提交的题为“RANKING FUNCTIONS USING A BIASED CLICK DISTANCE OF ADOCUMENT ON A NETWORK（使用网络上文档的经偏移的点击距离的排位函数）”的美国专利申请序列号 11/206,286 中描述的、考虑网络空间内每一文档的经偏移的点击距离的查询无关分量，以及在 2004 年 8 月 30 日提交的标题为“SYSTEM AND METHOD FOR RANKING SEARCH RESULTSUSING CLICK DISTANCE（使用点击距离排位搜索结果的系统和方法）”的美国专利申请序列号 10/955,983 中描述的、考虑网络空间内每一文档的 URL 的查询无关分量。上述美国专利申请的每一主题都已转让给本发明专利申请的受让人，通过引用将它们完整地包括于此。

[0014] 在又一示例性实施例中，所揭示的用以确定文档相关性分数的方法使用包括至少一个查询无关分量的排位函数，该至少一个查询无关分量既包括上述文档使用参数，也包括一或多个上述附加查询无关分量。

[0015] 文档相关性分数可用于排列网络空间内的文档。例如，一种排列网络上文档的方法包括以下步骤：使用上述方法确定网络上每一文档的文档相关性分数；以及基于每一文档的文档相关性分数按所需顺序（一般按降序）排列这些文档。

[0016] 文档相关性分数还可用于排列搜索查询的搜索结果。例如，一种排列搜索查询的搜索结果的方法可包括以下步骤：使用上述方法确定搜索查询的搜索结果中每一文档的文档相关性分数，以及基于每一文档的文档相关性分数按所需顺序（一般按降序）排列这些

文档。

[0017] 使用在此揭示方法的应用程序可在包括各种硬件组件的计算机系统上加载并执行。下面描述用于实践在此揭示的方法的示例性计算机系统和示例性操作环境。

[0018] 示例件操作环境

[0019] 图 2 示出了可在其上实现在此公开各方法的合适的计算系统环境 100 的示例。计算系统环境 100 仅为合适的计算环境的一个示例，并非对在此公开各方法的使用范围或功能性提出任何局限。计算环境 100 也不应解释成对于在示例性操作环境 100 中所示出的任一组件或其组合有任何依赖或要求。

[0020] 在此公开的方法可运行于多种其它通用或专用计算系统环境或配置。适合在此处公开的方法中使用的公知的计算系统、环境和 / 或配置的示例包括，但不限于，个人计算机、服务器计算机、手持式或膝上型设备、多处理器系统、基于微处理器的系统、机顶盒、可编程消费者电子产品、网络 PC、小型机、大型机、包括上述系统或设备中的任一个的分布式计算机环境等。

[0021] 在此公开的方法和过程可在诸如由计算机执行的程序模块等计算机可执行指令的通用上下文中描述。一般而言，程序模块包括例程、程序、对象、组件、数据结构等，它们执行特定任务或实现特定抽象数据类型。在此公开的方法和过程也可以在分布式计算环境中实现，其中任务由通过通信网络连接的远程处理设备来执行。在分布式计算环境中，程序模块可以位于包括存储器存储设备在内的本地和远程计算机存储介质中。

[0022] 参考图 2，用于实现在此公开的方法和过程的一个示例性系统包括计算机 110 形式的通用计算设备。计算机 110 的组件可以包括但不限于，处理单元 120、系统存储器 130 和将包括系统存储器 130 在内的各种系统组件耦合至处理单元 120 的系统总线 121。系统总线 121 可以是若干类型的总线结构中的任一种，包括存储器总线或存储器控制器、外围总线和使用各种总线体系结构中的任一种的局部总线。作为示例，而非限制，这样的体系结构包括工业标准体系结构 (ISA) 总线、微通道体系结构 (MCA) 总线、增强型 ISA(EISA) 总线、视频电子技术标准协会 (VESA) 局部总线和外围部件互连 (PCI) 总线 (也称为 Mezzanine 总线)。

[0023] 计算机 110 通常包括各种计算机可读介质。计算机可读介质可以是能由计算机 110 问的任何可用介质，而且包含易失性 / 非易失性介质以及可移动 / 不可移动介质。作为示例，而非限制，计算机可读介质可以包括计算机存储介质和通信介质。计算机存储介质包括易失性和非易失性、可移动和不可移动介质，它们以用于存储诸如计算机可读指令、数据结构、程序模块或其它数据这样的信息的任意方法或技术来实现。计算机存储介质包括，但不限于，RAM、ROM、EEPROM、闪存或其它存储器技术、CD-ROM、数字多功能盘 (DVD) 或其它光盘存储、磁带盒、磁带、磁盘存储或其它磁性存储设备、或能用于存储所需信息且可以由计算机 100 访问的任何其它介质。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，且包含任何信息传递介质。术语已调制数据信号摂指的其一个或多个特征以在信号中编码信息的方式被设定或更改的信号。作为示例，而非限制，通信介质包括有线介质，诸如有线网络或直接线连接，以及无线介质，诸如声学、RF、红外线和其它无线介质。上述中的任意组合也应包括在此处使用的计算机可读介质的范围之内。

[0024] 系统存储器 130 包括计算机存储介质，其形式为易失性和 / 或非易失性存储器，譬如只读存储器 (ROM)131 和随机存取存储器 (RAM)132。基本输入 / 输出系统 133(BIOS) 包含有助于诸如启动时在计算机 110 中元件之间传递信息的基本例程，它通常存储在 ROM 131 中。RAM 132 通常包含处理单元 120 可以立即访问和 / 或目前正在操作的数据和 / 或程序模块。作为示例而非局限，图 2 示出了操作系统 134、应用程序 135、其它程序模块 136 和程序数据 137。

[0025] 计算机 110 也可包括其它可移动 / 不可移动、易失性 / 非易失性计算机存储介质。仅作为示例，图 2 示出了从不可移动、非易失性磁介质中读取或向其写入的硬盘驱动器 140，从可移动、非易失性磁盘 152 中读取或向其写入的磁盘驱动器 151，以及从诸如 CD ROM 或其它光学介质等可移动、非易失性光盘 156 中读取或向其写入的光盘驱动器 155。可以在示例性操作环境中使用的其它可移动 / 不可移动、易失性 / 非易失性计算机存储介质包括但不限于，磁带盒、闪存卡、数字多功能盘、数字录像带、固态 RAM、固态 ROM 等等。硬盘驱动器 141 通常由不可移动存储器接口，诸如接口 140 连接至系统总线 121，磁盘驱动器 151 和光盘驱动器 155 通常由可移动存储器接口，诸如接口 150 连接至系统总线 121。

[0026] 以上讨论并在图 2 中示出的驱动器及其相关联的计算机存储介质为计算机 110 提供了对计算机可读指令、数据结构、程序模块和其它数据的存储。例如，在图 2 中，硬盘驱动器 141 被示为存储操作系统 144、应用程序 145、其它程序模块 146 和程序数据 147。注意，这些组件可以与操作系统 134、应用程序 135、其它程序模块 136 和程序数据 137 相同或不同。操作系统 144、应用程序 145、其它程序模块 146 和程序数据 147 在这里被标注了不同的标号是为了说明至少它们是不同的副本。

[0027] 用户可以通过输入设备如键盘 162 和定点设备 161(通常指鼠标、跟踪球或触摸板) 向计算机 110 输入命令和信息。其它输入设备 (未示出) 可以包括麦克风、操纵杆、游戏垫、圆盘式卫星天线、扫描仪等。这些和其它输入设备通常由耦合至系统总线的用户输入接口 160 连接至处理单元 120，但也可以由其它接口或总线结构，诸如并行端口、游戏端口或通用串行总线 (USB) 连接。监示器 191 或其它类型的显示设备也经由接口如视频接口 190 连接到系统总线 121。除监视器 191 以外，计算机 110 也可以包括其它外围输出设备，诸如扬声器 197 和打印机 196，它们可以通过输出外围接口 195 连接。

[0028] 计算机 110 可在使用至一个或多个远程计算机如远程计算机 180 的逻辑连接的网络化环境下操作。远程计算机 180 可以是个人计算机、服务器、路由器、网络 PC、对等设备或其它常见网络节点，且通常包括上文相对于计算机 110 描述的许多或所有元件，尽管在图 2 中只示出存储器存储设备 181。图 2 中所示的逻辑连接包括局域网 (LAN)171 和广域网 (WAN)173，但也可以包括其它网络。这样的网络环境常见于办公室、企业范围计算机网络、内联网和因特网。

[0029] 当在 LAN 网络环境中使用时，计算机 110 通过网络接口或适配器 170 连接至局域网 171。当在 WAN 网络环境中使用时，计算机 110 通常包括调制解调器 172，或用于通过 WAN173，如因特网建立通信的其它装置。调制解调器 172 可以是内置或外置的，它可以通过用户输入接口 160 或其它合适的机制连接至系统总线 121。在网络化环境中，相对于计算机 110 所描述的程序模块或其部分可以存储在远程存储器存储设备中。作为示例而非局限，图 2 示出远程应用程序 185 驻留在存储器设备 181 上。将会理解 :所示的这些网络连接起

示例性的作用，也可以使用在计算机之间建立通信链路的其他手段。

[0030] 在此揭示的方法和过程可使用一或多个应用程序来实现，这包括但不限于，服务器系统软件应用程序（例如，WINDOWS SERVER SYSTEM™ 软件应用程序），搜索排位应用程序，以及用于生成、维护和存储网络空间内的文档的使用数据的应用程序（例如，WINDOWS® SHAREPOINT® Services 应用程序），被指定为示例性系统 100 中的应用程序 135、应用程序 145 和远程应用程序 185 的众多应用程序之一的任一个应用程序。

[0031] 如上所述，本领域技术人员将了解，所揭示的为给定文档生成文档相关性分数的方法可在其它计算机系统配置中实现，这包括手持设备、多处理器系统、基于微处理器或可编程电子消费品、联网的个人计算机、小型机、大型机等等。所揭示的为给定文档生成文档相关性分数的方法也可在分布式计算环境中实践，其中任务由通过通信网络链接的远程处理设备来完成。在分布式计算环境中，程序模块可位于本地和远程两者的存储器存储设备中。

[0032] 示例性实施例的实现

[0033] 如上所述，提供确定网络上文档的文档相关性分数的方法。所揭示的方法可使用考虑网络上每一文档的文档使用值的排位函数来排列网络上文档。

[0034] 所揭示的确定网络上文档的文档相关性分数的方法可包括多个步骤。在一个示例性实施例中，确定网络上文档的文档相关性分数的方法包括以下步骤：向包括 N 个文档的网络上的一或多个文档指派实际使用值 (U_A)，其中该实际使用值 (U_A) 基于在服务器上维护和存储的实际使用数据；如果少于 N 个文档被指派了实际使用值 (U_A)，则向没有与其相关联的实际使用数据的文档指派默认使用值 (U_D)；以及使用每一文档的使用值（即 U_A 或 U_D）来确定网络上给定文档的文档相关性分数。

[0035] 如在此使用的，术语“实际使用数据”表示与一或多个用户对文档的“使用”相关联的一或多种类型的数据。给定文档或文档集合的实际使用数据类型可包括但不限于，在给定时间段内所有用户的文档查看数量、在给定时间段内每用户文档查看的平均数量、在给定时间段内在特定文档上花费的总时间、在给定时间段内在特定文档上花费的平均时间等等。给定时间段可以是例如上个星期内、上个月、去年一年内、文档的生存期或者任何其它所需的时间段。

[0036] 生成、维护和存储网络空间内文档的文档使用数据或统计量的步骤可由通常存在于计算系统上的应用程序代码来执行。文档使用数据是独立于给定搜索查询或搜索引擎来生成、维护和存储的，并且通常由维护文档（或页面）并使得文档（或页面）对用户可用的服务器上的应用程序代码来生成、维护和存储的。用于生成、维护和存储文档使用数据或统计量的合适应用程序包括但不限于 WINDOWS® SHAREPOINT® Services 和其它类似的应用程序。

[0037] 在这些服务站点以及执行类似功能的其它网站上存储和维护的文档使用数据，并如上所述可由应用程序代码来访问。例如，文档使用数据可通过 (i) 特殊的应用编程接口 (API)，(ii)web 服务请求，或 (iii) 请求返回网站上每一 URL 的使用数据的管理网页而从给定的网站（例如，WINDOWS® SHAREPOINT® Services 站点）访问。

[0038] 所揭示的确定网络上文档的文档相关性分数的方法可包括多个附加步骤，包括但

不限于：监控网络空间内一或多个文档的实际文档使用；在本地或远程数据存储文件中存储一或多个文档的实际文档使用数据；基于文档或包含该文档的文件夹的实际使用数据，计算文档的实际使用值（U_A）；在本地或远程数据存储文件中存储实际使用值（U_A）；向本地或远程数据存储文件请求存储的文档使用数据或实际使用值（U_A）（例如，在用户的特定搜索查询之后向搜索引擎请求这类数据）；从本地或远程数据存储文件检索一或多个文档的实际文档使用数据或实际使用值（U_A）；以及可选地，将文档使用值（即，实际或默认的）与一或多个附加文档属性合并以确定文档的文档相关性分数。

[0039] 图 3 表示示出示例性方法的示例性步骤的逻辑流程图，该方法提供网络上文档的实际或默认使用值，之后跟随着由系统管理员进行的可任选降低 / 提升过程。如图 3 所示，示例性方法 401 开始于框 402 并进行至步骤 403。在步骤 403，爬行（crawl）网络上的第一文档以获得实际使用数据。

[0040] 爬行第一文档以获得实际使用数据的步骤（步骤 403）可使用爬行应用程序来执行，该应用程序能够确定第一文档是否具有与其相关联的任何实际使用数据，并且如果第一文档具有与其相关联的实际使用数据，则检索该实际使用数据。适于在所揭示的提供网络上文档的实际或默认使用值的方法中使用的爬行应用程序包括但不限于在美国专利号 6,463,455 和 6,631,369 中描述的爬行应用程序，通过引用将这两个主题整体包括在此。

[0041] 如上所述，实际使用数据可从存储网络上一或多个文档的实际使用数据的一或多个文件获得。实际使用数据可作为文档分量与文档存储在一起，或者可与实际文档分开地存储在数据存储文件中。合适的远程存储系统包括但不限于，可从微软公司（华盛顿，雷德蒙德）购得的 WINDOWS® SHAREPOINT® Services(WSS) 产品，以及任何其它类似的远程存储系统。例如，WSS 远程存储系统记录包括例如所有用户对给定网络上每一文档的请求数量的实际使用数据，并且生成上个星期内、上个月、去年一年内、或文档的整个生存期或者任何其它时间段内每文档的点击数量统计。而且，如上所述，应当理解，在此揭示的方法不限于 WSS 远程存储系统，而可在所揭示方法中使用 WSS 远程存储系统或任何其它类似的文档数据系统。

[0042] 一旦爬行了文档，示例性方法 401 进行到判定框 404。在判定框 404，应用程序代码作出文档是否具有与其相关联的实际使用数据的判定。如果作出文档具有与其相关联的实际使用数据的判定，则示例性方法 401 进行到步骤 405，其中将基于实际使用的使用值（U_A）指派给文档。实际使用值（U_A）可使用关联于文档的实际使用数据的一或多个分量来确定。例如在一些实施例中，指派给文档的实际使用值（U_A）可仅与查看文档的用户数量相关。在其它实施例中，指派给文档的实际使用值（U_A）可与以下各项相关：在给定时间段内所有用户的文档查看数量、在给定时间段内每用户的文档查看平均数量、在给定时间段内在特定文档上花费的总时间，在给定时间段内在特定文档上花费的平均时间、或者任何上述标准的组合，其中给定时间段包括上个星期内、上个月、去年一年内、文档的生存期或者任何其它所需的时间段。

[0043] 在某些情况下，关联于给定文档的实际使用数据表示在给定时间段内未使用或未查看的文档。在这一情况下，文档可被指派使用值（U_A）等于零以表示在该时间段内没有使用；然而，通常是为基于实际使用或者未实际使用的使用值（U_A）指派一个非零的数。

[0044] 而且，在某些情况下，实际使用数据可与文档集合而非单个文档相关联。例如，文

件夹可包含文档集合，并且相关联的服务器可仅跟踪与访问（即，使用）该文件夹相关的使用数据，而不跟踪该文件夹内的单个文档。在此实施例中，如果存在关联于文件夹的实际使用数据，则可基于文件夹的实际使用数据为文件夹内的每一文档提供使用值（U_A）。通常，每一使用值（U_A）对于文件夹内的每一文档是相同的；然而若有需要，则可向文件夹内不同文档指派不同的使用值（U_A）。

[0045] 从步骤 405，示例性方法 401 进行至下述判定框 406。

[0046] 现在返回到判定框 404，如果作出文档不具有与其相关联的实际使用数据的判定，则示例性方法 401 进行至步骤 407，其中将默认使用值（U_D）指派给文档。例如，可将默认使用值（U_D）指派给作为不维护文档使用数据的网站一部分的文档。指派给文档的默认使用值（U_D）可用于提供文档相对于具有实际使用数据的文档的初始重要性。例如，如果给定文档的较高使用值指示网络内文档的相对重要性，那么向文档指派较低默认使用值（U_D）就会使文档的重要性相对于网络上其它文档而有所降低。

[0047] 在一个其中给定文档的较高使用值指示网络内文档的相对重要性的示例性实施例中，指派给文档的默认使用值（U_D）可与指派给网络上其它文档的实际使用数据（U_A）相关。例如，为了降低文档的相对重要性，可向文档指派默认使用值（U_D），其中该默认使用值（U_D）小于如上所述指派给网络上其它文档的任何实际使用值（U_A）。如果需要增加文档的相对重要性，可向文档指派一个默认使用值（U_D），其中该默认使用值（U_D）大于指派给网络上其它文档的任何实际使用值（U_A）或者大于指派给网络上部分其它文档的部分实际使用值（U_A）。

[0048] 在其它实施例中，可向没有实际使用数据的文档指派默认使用值（U_D），使得向该文档给出与具有已指派的实际使用值（U_A）的文档相比为平均的相对重要性。例如，在此实施例中，没有实际使用数据的文档的默认使用值（U_D）范围可从最小已指派实际使用值（U_{Amin}）至最大已指派实际使用值（U_{Amax}），或者在最小已指派实际使用值（U_{Amin}）与最大已指派实际使用值（U_{Amax}）之间的特定范围内。在此实施例中，为没有实际使用数据的文档提供平均的相对重要性，表示与具有与其相关联的实际使用数据的文档相比为中度的使用。

[0049] 从步骤 407，示例性方法 401 进行至判定框 406。在判定框 406，应用程序代码作出网络上的所有文档是否具有实际（U_A）或默认（U_D）使用值的判定。如果作出网络上的所有文档不具有实际（U_A）或默认（U_D）使用值的判断，则示例性方法 401 进行至步骤 408，其中爬行下一文档以获得实际使用数据。从步骤 408，示例性方法 401 返回至判定框 404 并且如上所述地进行。

[0050] 返回至判定框 406，如果应用程序代码作出网络上的所有文档具有实际（U_A）或默认（U_D）使用值的判定，则示例性方法 401 进行至判定框 409。在判定框 409，系统管理员作出是否要降低任何实际（U_A）或默认（U_D）使用值以便更确切地表示网络空间内给定文档的重要性的判定。如果作出降低一或多个实际（U_A）或默认（U_D）使用值以便更确切地表示网络空间内一或多个文档的重要性的判定，则示例性方法 401 进行至步骤 410，其中或负或正地调整一或多个文档（或 URL）的实际（U_A）或默认（U_D）使用值。从步骤 410，示例性方法 401 进行至下述步骤 411。

[0051] 返回至判定框 409，如果作出了不降低（或不提升）一或多个实际（U_A）或默认（U_D）使用值的判定，则示例性方法 401 直接进行至步骤 411。在步骤 411，在排位函数中使用实

际 (U_A) 和默认 (U_D) 使用值来确定网络空间内每一文档的总文档相关性分数。从步骤 411，示例性方法 401 进行至结束框 412。

[0052] 一旦所有实际 (U_A) 和默认 (U_D) 使用值已经确定并且可任选地被降低（或可选地被提升），则若有需要，就可使用每一文档的实际 (U_A) 或默认 (U_D) 使用值作为排位函数中的参数以提供每一文档的文档相关性分数。这一文档相关性分数可用于排列搜索查询的搜索结果。使用包含默认使用值参数的排位函数排列生成的搜索结果的示例性方法在图 4 中示出。

[0053] 图 4 提供示出示例性方法 20 的示例性步骤的逻辑流程图，其中示例性方法 20 包括使用包含使用值参数的排位函数排列生成的搜索结果的方法。如图 4 所示，示例性方法 20 开始于框 201 并且进行至步骤 202。在步骤 202，用户通过输入搜索查询来请求搜索。在步骤 202 之前，网络上每一文档的实际或默认使用值已经在先前计算过。从步骤 202，示例性方法 20 进行至步骤 203。

[0054] 在步骤 203，将网络上每一文档的实际或默认使用值与存储在索引中的每一文档的任何其它文档统计量（例如，其它查询无关统计量）合并。将实际或默认使用值与其它文档统计量合并，这允许较快的查询响应时间，因为与排位相关的所有信息被群集在一起。因此，在索引内列出的每一文档在合并之后具有相关联的实际或默认使用值。一旦合并完成，示例性方法 20 就进行至步骤 204。

[0055] 在步骤 204，为给定文档提供包括使用参数的查询无关文档统计量作为排位函数的一分量。也为给定文档提供查询相关数据，一般作为排位函数的独立分量。排位函数的查询相关数据或者内容相关部分依赖于实际的搜索项和给定文档的内容。

[0056] 在一个实施例中，排位函数包括至少一个查询无关 (QID) 分量，该分量包括使用参数。在一个实施例，该查询无关 (QID) 分量可由下式表示：

[0057] $$QID(doc) = w_u \frac{k_u U}{k_u + U} \quad (1)$$

[0058] 其中：

[0059] U 表示表示实际使用值或默认使用值；以及

[0060] w_u 和 k_u 表示使用值的调整参数。在另一实施例中，查询无关 (QID) 分量可由下式表示：

[0061] $QID(doc) = w_u U + k_u \quad (2)$

[0062] 其中：

[0063] U 表示实际使用值或默认使用值；以及

[0064] w_u 和 k_u 表示使用值的调整参数。在又一实施例中，查询无关 (QID) 分量可由下式表示：

[0065] $QID(doc) = w_u[1+\exp(-k_u U - B)] + C \quad (3)$

[0066] 其中：

[0067] U 表示实际使用值或默认使用值；以及

[0068] w_u、k_u、B 和 C 表示使用值的调整参数（即，比例常数）。

[0069] 在另一实施例中，排位函数包括上述查询无关 (QID) 分量与至少一个查询相关 (QD) 分量之和，诸如

[0070] Score = QD(doc, query)+QID(doc).

[0071] QD 分量可以是任何文档记分函数。在一个实施例中，QD 分量对应于在 2004 年 3 月 18 日提交的题为“FIELD WEIGHTING IN TEXT DOCUMENTSEARCHING（在文本文档搜索中的域加权）”的美国专利申请序列号 10/804,326 中描述的域加权记分函数，其主题通过引用整体包含在此。如在美国专利申请序列号 10/824,326 中提供的，可用作域加权记分函数的表示的一个公式如下：

[0072] $$QD(doc,query)=\sum\frac{wtf'(k_1+1)}{k_1+wtf'}\times\log(\frac{N}{n})$$

[0073] 其中：

[0074] wtf′ 表示加权的项频率或者在搜索查询中的给定项乘以在所有域（例如，文档的标题、主体等）上的权重的项频率之和，并且按照每一域的长度和相应的平均长度来归一化，

[0075] N 表示网络上文档的数量，

[0076] n 表示包含查询项的文档数量，以及

[0077] k_1 是可调的常数。

[0078] 上述各项和公式在美国专利申请序列号 10/804,326 中进一步详细地描述，其主题通过引用整体包含于此。

[0079] 在一些实施例中，排位函数还可包括一 QID 分量，该 QID 分量考虑 (i) 由在 2004 年 8 月 30 日提交的标题为“SYSTEM AND METHOD FORRANKING SEARCH RESULTS USING CLICK DISTANCE”的美国专利申请序列号 10/955,983 中揭示的方法确定的点击距离值，(ii) 由在 2005 年 8 月 15 日提交的标题为“RANKING FUNCTIONS USING A BIASED CLICKDISTANCE OF A DOCUMENT ON A NETWORK”的美国专利申请序列号 11/206,286 所揭示的方法确定的经偏移的点击距离值（上述两者的主题通过引用整体包括于此），(iii) 文档的 URL 深度，或者 (iv) 上述 (i) 或 (ii) 和 (iii) 的组合。例如，这一可任选的附加 QID 分量可包括如下的函数：

[0080] $$QID(doc)=w_{cd}\frac{k_{cd}}{k_{cd}+\frac{b_{cd}\frac{CD}{k_{ew}}+b_{ud}UD}{b_{cd}+b_{ud}}}$$

[0081] 其中：

[0082] w_{cd} 表示查询无关分量诸如包含点击距离或经偏移的点击距离参数的分量的权重，

[0083] b_{cd} 表示点击距离或经偏移的点击距离相对于 URL 深度的权重，

[0084] b_{ud} 表示 URL 深度的权重，

[0085] CD 表示文档的经计算或指派的点击距离或经偏移的点击距离，

[0086] k_{ew} 表示通过优化排位函数的精度确定的调整常数，它与其它调整参数相类似（即，k_{ew} 可在所有边界具有相同的边界加权值时表示边界加权值，或者 k_{ew} 可在边界加权值互相不同时表示平均或中间边界值），

[0087] UD 表示 URL 深度，以及

[0088] k_{cd} 是点击距离饱和常数。

[0089] 经加权的项（w_{cd}，b_{cd} 和 b_{ud}）协助定义其相关各项中每一项（即，分别为包含点击距离或经偏移的点击距离参数的分量、给定文档的点击距离或经偏移的点击距离值、以及给定文档的 URL 深度）的重要性以及记分函数的最后结果。

[0090] URL 深度 (UD) 可任选地附加到以上引用的查询无关分量，以平滑点击距离或经偏移的点击距离值在记分函数上具有的影响。例如，在某些情况下，不是很重要的文档（即，具有较大 URL 深度）可能具有短点击距离或经偏移的点击距离值。URL 深度由文档的 URL 中斜杠的数量来表示。例如，www.example.com\d1\d2\d3\d4.htm 包括四个斜杠，因此具有为 4 的 URL 深度。然而，该文档可具有从主页 www.example.com 的直接链接，这给它相对较短的点击距离或经偏移的点击距离。在上述引用的函数中包括 URL 深度项并且针对点击距离或经偏移的点击距离值加权 URL 深度项，这补偿了相对较长的点击距离或经偏移的点击距离以更准确地反映文档在网络中的重要性。依赖于网络，URL 深度为 3 或更大被视为深链接。

[0091] 在一个实施例中，用于确定给定文档的文档相关性分数的排位函数包括如下的函数：

[0092]
$$Score = \sum \frac{wtf'(k_1+1)}{k_1+wtf'} \times \log(\frac{N}{n}) + w_{cd} \frac{k_{cd}}{k_{cd} + \frac{b_{cd}\frac{CD}{k_{ew}} + b_{ud}UD}{b_{cd}+b_{ud}}} + w_u \frac{k_u U}{k_u + U}$$

[0093] 其中各项如上所述。

[0094] 在其它实施例中，URL 深度可从排位函数中移除或者可添加其它分量到排位函数以提高查询相关分量、查询无关分量或两者的精度。而且，上述包含使用参数的查询无关分量可结合到其它排位函数（未示出）以改善对搜索结果的排位。

[0095] 一旦在步骤 204 将给定文档的文档统计量提供给排位函数，示例性方法 20 就进行至步骤 205。在步骤 205，针对给定文档确定文档相关性分数，将其存储在存储器中，并且与给定文档相关联。从步骤 205，示例性方法 20 进行至判定框 206。

[0096] 在判定框 206，应用程序代码作出是否已经为网络内每一文档计算了文档相关性分数的判定。如果作出尚未为网络内每一文档计算文档相关性分数的判定，则示例性方法 20 返回至步骤 204 并且如上所述地继续。如果作出已经为网络内每一文档计算了文档相关性分数的判定，则示例性方法 20 进行至步骤 207。

[0097] 在步骤 207，查询的搜索结果包括按照众多文档的文档相关性分数排列这些文档。所得的文档相关性分数考虑网络内每一文档的实际或默认使用值。一旦排列了搜索结果，示例性方法 20 就进行至步骤 208，其中向用户显示经排列的结果。从步骤 208，示例性方法 20 进行至步骤 209，其中由用户选择和查看具有最高的排位结果。从步骤 209，示例性方法 20 进行至步骤 210 并在此示例性方法 20 结束。

[0098] 除了生成网络内文档的文档相关性分数并且使用文档相关性分数来排列搜索查询的搜索结果的上述方法之外，在此还揭示了具有存储在其上的用于执行上述方法的计算机可执行指令的计算机可读介质。

[0099] 在此还揭示了计算系统。示例性计算系统包含至少一个能在计算系统上使用的应用程序模块，其中该至少一个应用程序模块包括在该计算系统上加载的应用程序代码，其

中该应用程序代码执行生成网络内文档的文档相关性分数的方法。应用程序代码可使用任何上述计算机可读介质来加载到计算系统上，其中存储在计算系统上的上述计算机可读介质具有用于如上所述地生成网络内文档的文档相关性分数并且使用文档相关性分数来排列搜索查询的搜索结果的计算机可执行指令。

[0100] 尽管已经详细地参考本说明书的特定实施例描述了本说明书，但是应该认识到，本领域的技术人员在理解了上述内容之后，可以容易地想到这些实施例的改变、变体或等价方案。因此，所揭示方法、计算机可读介质以及计算系统的范围应当由所附权利要求书及其任何等价方案来确定。

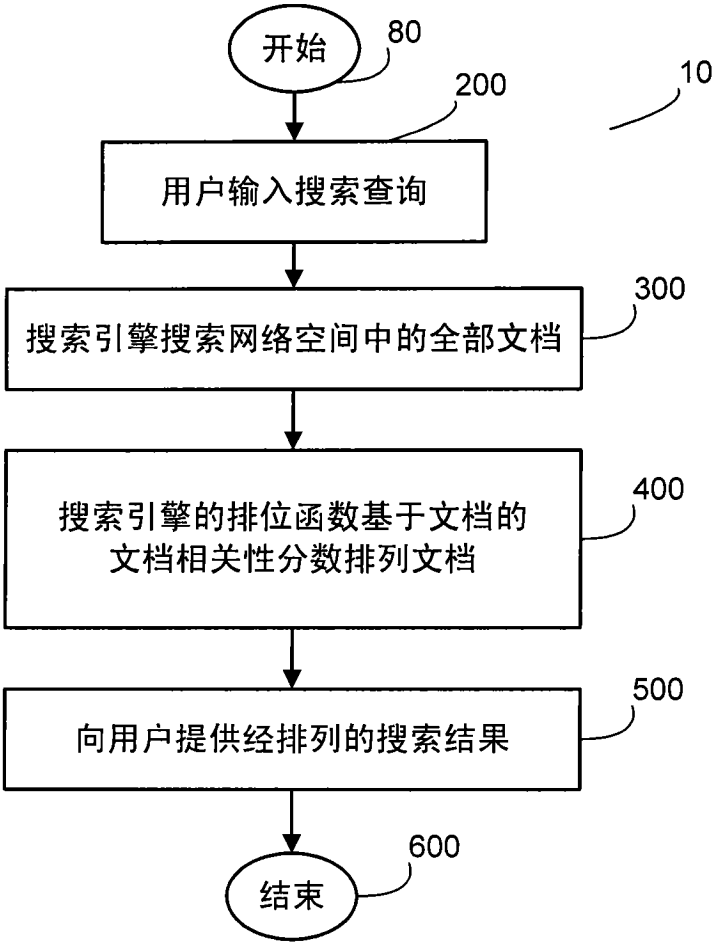

开始
80
200
10
用户输入搜索查询
300
搜索引擎搜索网络空间中的全部文档
400
搜索引擎的排位函数基于文档的
文档相关性分数排列文档
500
向用户提供经排列的搜索结果
600
结束

图 1

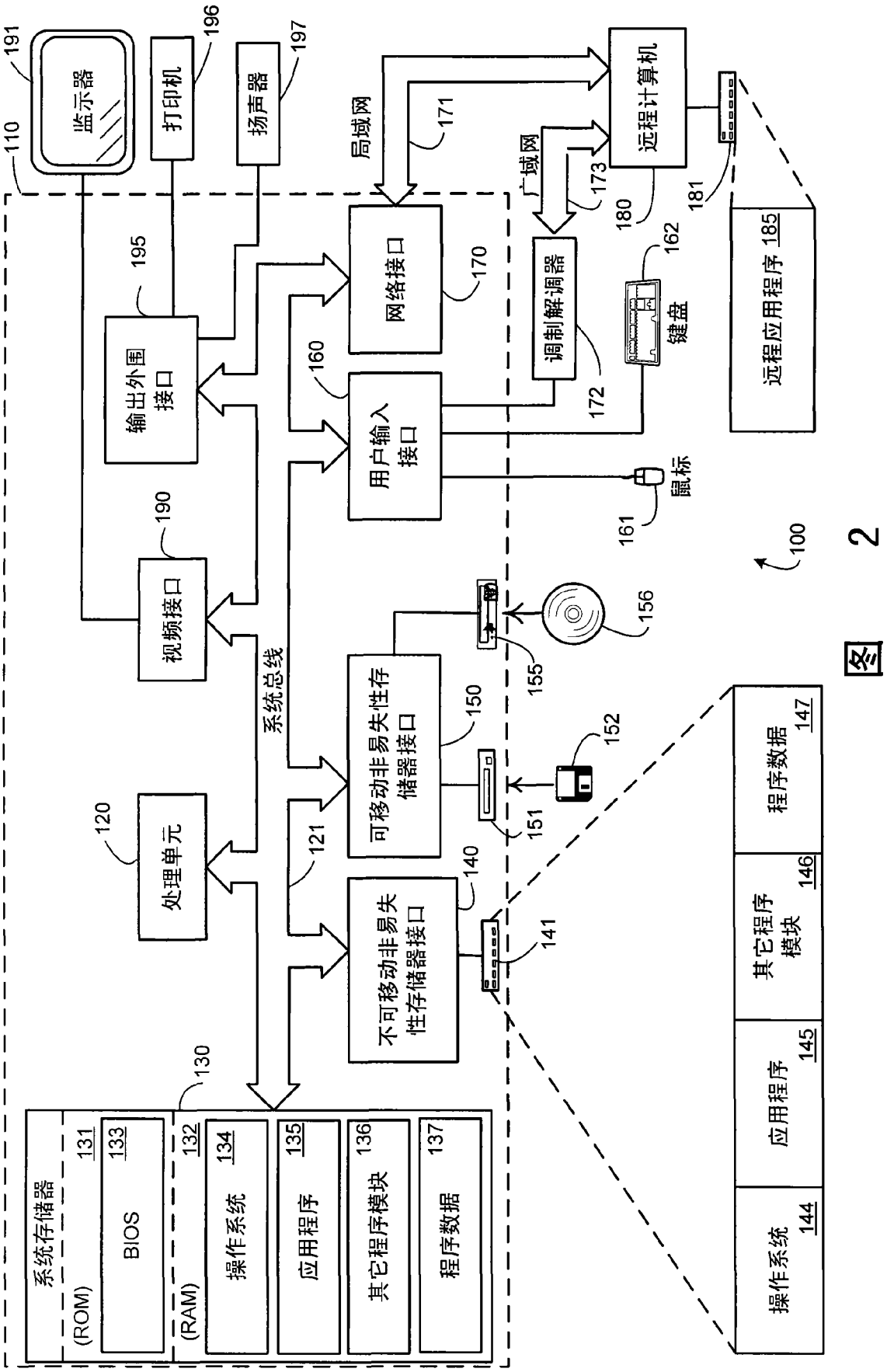

监示器
191
110
打印机
196
扬声器
197
局域网
171
远程计算机
广域网
173
180
181
远程应用程序 185
网络接口
170
输出外围接口
195
调制解调器
172
162
键盘
160
用户输入接口
鼠标
161
190
视频接口
100
156
155
系统总线
可移动非易失性存储器接口
150
152
151
120
处理单元
121
140
不可移动非易失性存储器接口
141
130
系统存储器
(ROM)
131
BIOS
133
(RAM)
132
操作系统
134
应用程序
135
其它程序模块
136
程序数据
137
程序数据
147
其它程序模块
146
应用程序
145
操作系统
144

图 2

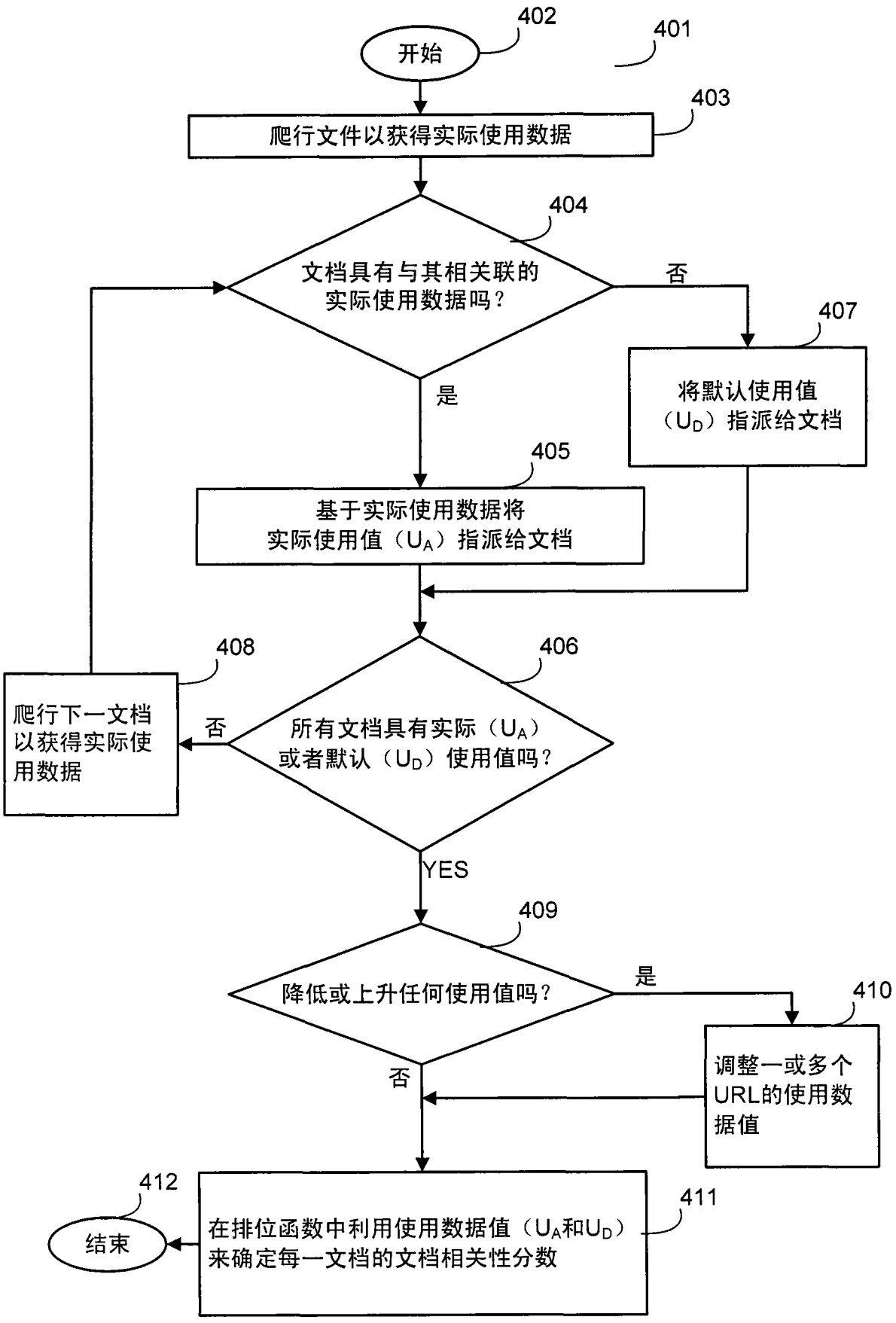

402
401
开始
403
爬行文件以获得实际使用数据
404
文档具有与其相关联的实际使用数据吗？
否
407
将默认使用值（U_D）指派给文档
是
405
基于实际使用数据将实际使用值（U_A）指派给文档
408
406
爬行下一文档以获得实际使用数据
否
所有文档具有实际（U_A）或者默认（U_D）使用值吗？
YES
409
降低或上升任何使用值吗？
是
410
调整一或多个URL的使用数据值
否
412
411
结束
在排位函数中利用使用数据值（U_A和U_D）来确定每一文档的文档相关性分数

图 3

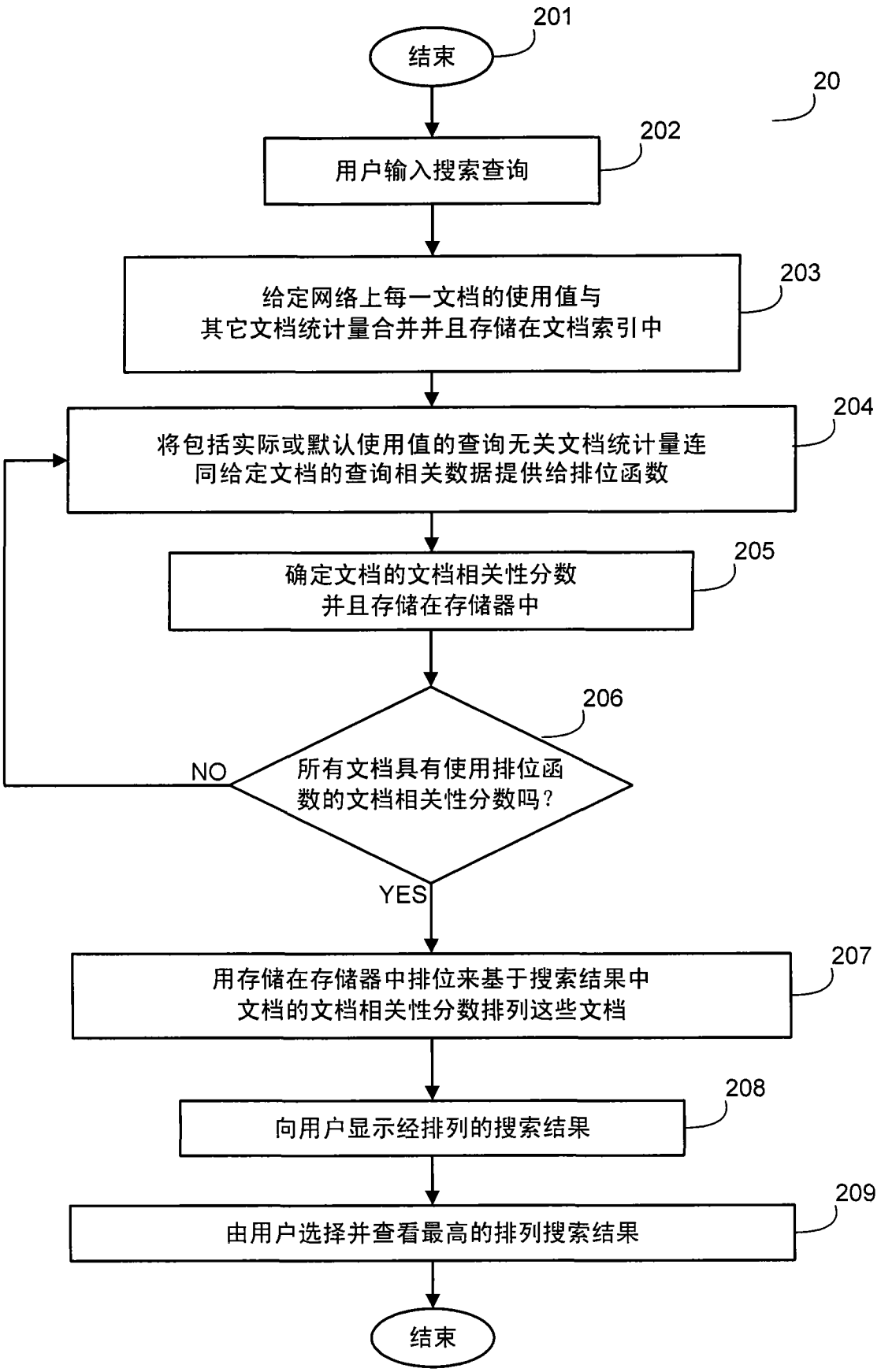

201
结束
20
202
用户输入搜索查询
203
给定网络上每一文档的使用值与
其它文档统计量合并并且存储在文档索引中
204
将包括实际或默认使用值的查询无关文档统计量连
同给定文档的查询相关数据提供给排位函数
205
确定文档的文档相关性分数
并且存储在存储器中
206
NO
所有文档具有使用排位函
数的文档相关性分数吗?
YES
207
用存储在存储器中排位来基于搜索结果中
文档的文档相关性分数排列这些文档
208
向用户显示经排列的搜索结果
209
由用户选择并查看最高的排列搜索结果
结束

图 4