(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 101546309 B
(45)授权公告日 2012.07.04

(21)申请号 200810084087.X
(22)申请日 2008.03.26
(73)专利权人 国际商业机器公司
地址 美国纽约
(72)发明人 张岭 沈羽
(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 李德山 李春晖
(51)Int.Cl.
G06F 17/30(2006.01)
(56)对比文件
CN 101097578 A,2008.01.02, 全文.
US 2006/0253462 A1,2006.11.09, 摘要、说明书第58段-第116段、附图1,5.
US 2005/0234895 A1,2005.10.20, 说明书第58-59段、附图8,9.

审查员 胡平

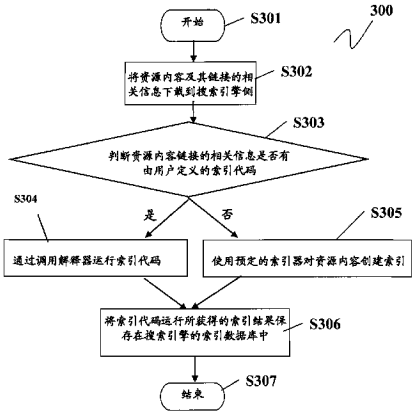
权利要求书 2 页 说明书 8 页 附图 4 页

(54)发明名称

对计算机网络中的资源内容构建索引的方法和设备

(57)摘要

本发明公开了一种对计算机网络中的资源内容构建索引的方法和设备，所述计算机网络包括作为搜索引擎的计算机和作为资源内容站点并且存储有用户的资源内容的计算机，所述方法包括以下步骤：判断所述资源内容是否链接有由所述用户定义的索引代码，所述索引代码用于处理所述资源内容以获得描述所述资源内容的信息；如果判断所述资源内容链接有由所述用户定义的索引代码，则运行所述由用户定义的索引代码，以获得描述所述资源内容的信息作为索引结果。



CN 101546309 B

1. 一种用于对计算机网络中的资源内容构建索引的方法，所述计算机网络包括作为搜索引擎的计算机和作为资源内容站点并且存储有用户的资源内容的计算机，

所述方法包括以下步骤：

判断所述资源内容是否链接有由所述用户定义的索引代码，所述索引代码用于处理所述资源内容以获得描述所述资源内容的信息；和

如果判断所述资源内容链接有由所述用户定义的索引代码，则运行所述由用户定义的索引代码，以获得描述所述资源内容的信息作为索引结果，

其中，所述用户为所述资源内容的提供者。

2. 根据权利要求 1 所述的方法，其中还包括：将所述用户定义的索引代码运行所获得的索引结果保存在所述搜索引擎的索引数据库中。

3. 根据权利要求 1 所述的方法，其中，如果判断所述资源内容未链接有所述用户定义的索引代码，则使用预定的索引器对所述资源内容构建索引。

4. 根据权利要求 1 所述的方法，其中，判断所述资源内容是否链接有所述用户定义的索引代码以及运行所述用户定义的索引代码的步骤都是在所述搜索引擎侧进行的，

所述方法还包括：在判断所述资源内容是否链接有所述用户定义的索引代码的步骤之前，将所述资源内容及其链接的相关信息下载到所述搜索引擎。

5. 根据权利要求 4 所述的方法，其中，所述用户定义的索引代码由脚本文件来实现，

运行所述用户定义的索引代码的步骤是通过调用脚本引擎实现的。

6. 根据权利要求 4 所述的方法，其中，判断所述资源内容是否链接有所述用户定义的索引代码的步骤是通过解析所述资源内容并验证所述资源内容链接的相关信息而实现的。

7. 根据权利要求 1 所述的方法，其中，判断所述资源内容是否链接有所述用户定义的索引代码以及运行所述用户定义的索引代码的步骤都是在所述资源内容站点侧进行的。

8. 根据权利要求 7 所述的方法，其中还包括：在所述资源内容站点侧判断对所述资源内容的访问是来自所述搜索引擎的搜索器的访问还是一般浏览者的访问；如果是来自所述搜索引擎的搜索器的访问，则进一步执行所述判断资源内容是否链接有所述用户定义的索引代码的步骤。

9. 根据权利要求 1 所述的方法，其中，所述用户定义的索引代码描述了所述用户对所述资源内容中的索引项的自定义权重。

10. 根据权利要求 9 所述的方法，其中，所述索引项是由所述用户选择的。

11. 根据权利要求 1 所述的方法，其中，所述用户定义的索引代码是由用户使用所述资源内容的内容和 / 或组织作为索引项并对所述索引项赋予权重值而实现的。

12. 根据权利要求 1 所述的方法，其中，所述用户定义的索引代码是基于代码模板完成的，所述代码模板对应于所述资源内容的内容模板。

13. 根据权利要求 1 所述的方法，其中，所述用户定义的索引代码由脚本文件来实现。

14. 根据权利要求 1 所述的方法，其中，所述计算机网络是被管理的网络环境。

15. 一种用于对计算机网络中的资源内容构建索引的设备，所述计算机网络包括作为搜索引擎的计算机和作为资源内容站点并且存储有用户的资源内容的计算机，所述设备设置在所述搜索引擎侧并包括：

判断装置，被配置成接收所述搜索引擎的索引器下载的所述资源内容及其链接的相关

信息，并判断所述资源内容链接的相关信息是否包含由用户定义的索引代码，所述索引代码用于处理所述资源内容以获得描述所述资源内容的信息；和

解释器，被配置成如果判断所述资源内容链接有由所述用户定义的索引代码，则运行所述用户定义的索引代码，以获得描述所述资源内容的信息作为索引结果，

其中，所述用户为所述资源内容的提供者。

16. 根据权利要求 15 所述的设备，其中还包括索引数据库，被配置成保存所述解释器的索引结果。

17. 根据权利要求 15 所述的设备，其中还包括预定的索引器，该预定的索引器被配置成：如果所述判断装置判断所述资源内容未链接有所述用户定义的索引代码，对所述资源内容构建索引。

18. 根据权利要求 15 所述的设备，其中，所述用户定义的索引代码由脚本文件来实现，所述解释器是脚本引擎。

19. 根据权利要求 15 所述的设备，其中，所述判断装置还被配置成解析所述资源内容并验证所述资源内容链接的相关信息，以判断所述资源内容是否链接有所述用户定义的索引代码。

20. 根据权利要求 15 所述的设备，其中，所述用户定义的索引代码描述了所述用户对所述资源内容中的索引项的自定义权重。

21. 根据权利要求 20 所述的设备，其中，所述索引项是由所述用户选择的。

22. 根据权利要求 15 所述的设备，其中，所述用户定义的索引代码是由用户使用所述资源内容的内容和 / 或组织作为索引项并对所述索引项赋予权重值而实现的。

23. 根据权利要求 15 所述的设备，其中，所述用户定义的索引代码是基于代码模板完成的，所述代码模板对应于所述资源内容的内容模板。

24. 根据权利要求 15 所述的设备，其中，所述计算机网络是被管理的网络环境。

25. 一种用于对计算机网络中的资源内容构建索引的设备，所述计算机网络包括作为搜索引擎的计算机和作为资源内容站点并且存储有用户的资源内容的计算机，所述设备设置在所述资源内容站点侧并包括：

判断装置，被配置成判断所述资源内容链接的相关信息是否包含由用户定义的索引代码，所述索引代码用于处理所述资源内容以获得描述所述资源内容的信息；

解释器，被配置成如果判断所述资源内容链接有由所述用户定义的索引代码，则运行所述用户定义的索引代码，以获得描述所述资源内容的信息作为供所述搜索引擎下载的索引结果，

其中，所述用户为所述资源内容的提供者。

26. 根据权利要求 25 所述的设备，其中，所述判断装置还被配置成判断对所述资源内容的访问是来自所述搜索引擎的搜索器的访问还是一般浏览者的访问，并且响应于来自所述搜索引擎的搜索器的访问，执行所述判断资源内容是否链接有所述用户定义的索引代码。

对计算机网络中的资源内容构建索引的方法和设备

技术领域

[0001] 本发明涉及搜索引擎技术，尤其涉及对计算机网络中的资源内容构建索引的方法和设备。

背景技术

[0002] 随着计算机和互联网技术的发展，搜索引擎已经成为 Web 客户机（例如计算机）使用者获取信息的重要方式。传统的搜索引擎例如有 Inktomi，Excite，Lycos，Infoseek 或 FAST 等，包括在互联网和搜索器服务器之间发送和接收信息包的路由器、索引服务器和网络服务器。搜索引擎使用搜索器 (WEB 爬虫或称为蜘蛛、机器人程序）定期地访问通过 URL 定位的网页资源，提取出其中的文本信息和其它相关网页属性，并储存该信息以使得索引服务器可以处理检索到的数据。所述索引服务器解析这些文档并通过应用索引算法创建文档索引，通常是根据每个文档所包含的关键字和其它属性来创建有优先级的索引。

[0003] 网络服务器包括搜索程序，用于处理针对搜索引擎的搜索请求。一般，基于用户通过向搜索引擎提供的感兴趣的关键字，搜索程序根据用户提供的关键字通过索引器检索事先建立好的索引数据库来生成提供给用户的关键字结果页面，来帮助用户发现和访问新的“统一资源地址”(URL)。

[0004] 为建立搜索索引，搜索引擎使用了不同种类的算法来创建索引。对于现代的搜索引擎，它们使用文档内容和链接信息（例如 Google 的网页级别）二者来建立索引。当搜索引擎试图找到对用户查询最相关的文档时，则对文档索引应用搜索算法然后返回匹配的结果。

[0005] 因此基本上，搜索引擎将使用同样的算法集来为文档排序，而最重要的是，该算法是由搜索服务的提供者（例如 Google 或 Yahoo）自己设计和维护的。对于 Web 内容的所有者而言，他们所能提供的仅仅是网页，而让搜索引擎根据网页的内容决定文档索引如何建立。

[0006] WO2001027793 提出在各远程服务器设置代理程序，使用该代理程序为每个远程服务器所属的所有计算机产生搜索引擎更新信息，由此可以减小中心索引服务器的工作负荷并提高效率。根据 WO2001027793，所述代理程序仍然使用由搜索服务提供者提供的、与搜索引擎侧完全相同的索引算法。

发明内容

[0007] 考虑到现有技术的搜索引擎都是提供控制的索引，本发明的目的是提出一种不同的搜索引擎搜索引技术，旨在提供涉及用于 Web 搜索的“用户贡献的索引”。具体地，本发明利用了由用户自定义的索引算法，以期提高索引质量和搜索质量。

[0008] 根据本发明的一个方面，一种用于对计算机网络中的资源内容构建索引的方法，所述计算机网络包括作为搜索引擎的计算机和作为资源内容站点并且存储有用户的资源内容的计算机，

[0009] 所述方法包括以下步骤：

[0010] 判断所述资源内容是否链接有由所述用户定义的索引代码，所述索引代码用于处理所述资源内容以获得描述所述资源内容的信息；和

[0011] 如果判断所述资源内容链接有由所述用户定义的索引代码，则运行所述由用户定义的索引代码，以获得描述所述资源内容的信息作为索引结果。

[0012] 如果判断所述资源内容未链接有所述用户定义的索引代码，则使用预定的索引器（即由搜索服务提供者提供的索引算法）对所述资源内容构建索引，这种情况下则与传统的在所述搜索引擎侧或资源内容站点侧进行的受控索引相似。

[0013] 当本发明的方法在搜索引擎侧进行的情况下，所述方法还包括：在判断所述资源内容是否链接有所述用户定义的索引代码的步骤之前，将所述资源内容及其链接的相关信息下载到所述搜索引擎。

[0014] 优选地，所述用户定义的索引代码一般由脚本文件来实现，相应地运行所述用户定义的索引代码的步骤是通过调用脚本引擎实现的。

[0015] 判断所述资源内容是否链接有所述用户定义的索引代码的步骤是通过解析所述资源内容并验证所述资源内容链接的相关信息而实现的。

[0016] 另外，当本发明的方法在资源内容站点侧进行的情况下，需要预先在资源内容站点侧判断对所述资源内容的访问是来自所述搜索引擎的搜索器的访问还是一般浏览者的访问；如果是来自所述搜索引擎的搜索器的访问，则进一步执行所述判断资源内容是否链接有所述用户定义的索引代码的步骤。

[0017] 所述用户定义的索引代码描述了所述用户对所述资源内容中的索引项的自定义权重，并且所述索引项也可由所述用户选择。优选地，所述用户定义的索引代码是由用户使用所述资源内容的内容和/或组织作为索引项并对所述索引项赋予权重值而实现的。

[0018] 为方便用户完成自定义的索引代码，可允许用户基于代码模板来完成索引代码。所述代码模板对应于所述资源内容的内容模板。

[0019] 根据本发明的另一个方面，一种用于对计算机网络中的资源内容构建索引的设备，所述计算机网络包括作为搜索引擎的计算机和作为资源内容站点并且存储有用户的资源内容的计算机，所述设备设置在所述搜索引擎侧并包括：

[0020] 判断装置，被配置成接收所述搜索引擎的索引器下载的所述资源内容及其链接的相关信息，并判断所述资源内容链接的相关信息是否包含由用户定义的索引代码，所述索引代码用于处理所述资源内容以获得描述所述资源内容的信息；和

[0021] 解释器，被配置成运行所述用户定义的索引代码，以获得描述所述资源内容的信息作为索引结果。

[0022] 所述设备还包括被配置成保存所述解释器的索引结果的索引数据库。

[0023] 所述设备还包括预定的索引器，该预定的索引器被配置成：如果所述判断装置判断所述资源内容未链接有所述用户定义的索引代码，对所述资源内容构建索引。

[0024] 优选地，所述用户定义的索引代码由脚本文件来实现，相应地所述解释器是脚本引擎。

[0025] 根据本发明的再一个方面，用于对计算机网络中的资源内容构建索引的设备设置在所述资源内容站点侧，相应地其判断装置被配置成判断所述资源内容链接的相关信息是

否包含由用户定义的索引代码，而其解释器被配置成运行所述用户定义的索引代码，以获得描述所述资源内容的信息作为供所述搜索引擎下载的索引结果。

[0026] 所述判断装置还被配置成判断对所述资源内容的访问是来自所述搜索引擎的搜索器的访问还是一般浏览者的访问，这样仅仅响应于来自所述搜索引擎的搜索器的访问，所述判断装置执行所述判断资源内容是否链接有所述用户定义的索引代码，减小了资源内容站点侧的开销。

[0027] 本发明可以应用于因特网和被管理的网络环境。由于用户或内容所有者比其他任何人更好地理解文档（内容，版面设计，组织等各索引项），本发明的“用户贡献的索引”允许每个用户或内容所有者基于感兴趣的索引项来提供最佳描述文档的索引代码（即索引算法），因此本发明的“用户贡献的索引”可更有效地有助于提高索引质量，由此也提高了搜索质量。尤其在被管理的网络环境中，由于内容所有者提供的索引代码更值得信赖，因此本发明更优选地应用于被管理的网络环境（如内网）。

附图说明

[0028] 参照下面结合附图对本发明实施例的说明，会更加容易地理解本发明的以上和其它目的、特点和优点。在附图中，相同的或对应的技术特征或部件将采用相同或对应的附图标记来表示。

[0029] 图 1 是示出可实现本发明的分布式数据处理系统的框图。

[0030] 图 2 是应用了本发明第一实施例的索引构建设备的系统的框图。

[0031] 图 3 示出了根据本发明第一实施例的索引构建方法的流程图。

[0032] 图 4 示出了根据本发明第三实施例的索引构建方法的流程图。

具体实施方式

[0033] 下面参照附图来说明本发明的实施例。应当注意，为了清楚的目的，附图和说明中省略了与本发明无关的、本领域普通技术人员已知的部件和处理的表示和描述。

[0034] 系统体系

[0035] 现在参考附图，特别是图 1，描述了可实现本发明的分布式数据处理系统的框图。分布式数据处理系统 100 是可实现本发明的计算机网络。分布式数据处理系统 100 包含网络 102，网络 102 是用于在不同的设备和分布式数据处理系统 100 内连接到一起的计算机之间提供通信链接的媒介。

[0036] 在所描述的例子中，服务器 104 与存储器 106 一起连接到网络 102。此外，例如工作站、个人计算机、手机、PDA 等的客户端 108、110 和 112 也被连接到网络 102。在所描述的例子中，服务器 104 向客户端 108、110 和 112 提供如引导文件的数据、操作系统以及应用程序。分布式数据处理系统 100 可包括另外的服务器、客户端以及其它未显示的设备。在所描述的例子中，分布式数据处理系统 100 是因特网，网络 102 表示对使用 TCP/IP 协议套件来彼此通信的网络以及网关的集合。当然，分布式数据处理系统 100 还可被实现为不同类型的网络。

[0037] 企图将图 1 作为例子，而不是作为本发明所述过程的结构限制。在不偏离本发明精神和范围的条件下，可对图 1 所示系统作出许多更改。

[0038] 本发明可实现为如图 1 所示的服务器 104 的数据处理系统。该数据处理系统可以是包括连接到系统总线的多个处理器的对称对处理器 (SMP) 系统。亦可使用单处理器系统。本发明还可实现为图 1 中客户端计算机的数据处理系统。

[0039] 构建索引的方法和设备

[0040] 根据本发明公开了一种对计算机网络中的资源内容构建索引的方法，所述计算机网络包括作为搜索引擎的计算机和作为资源内容站点并且存储有用户的资源内容的计算机，所述方法包括以下步骤：判断所述资源内容是否链接有由所述用户定义的索引代码，所述索引代码用于处理所述资源内容以获得描述所述资源内容的信息；如果判断所述资源内容链接有由所述用户定义的索引代码，则运行所述由用户定义的索引代码，以获得描述所述资源内容的信息作为索引结果。本发明的方法可以在搜索引擎侧或在资源内容站点一侧。

[0041] 相应地，本发明的用于对计算机网络中的资源内容构建索引的设备，包括：判断装置，被配置成判断所述资源内容链接的相关信息是否包含由用户定义的索引代码；和解释器，被配置成运行所述用户定义的索引代码，以获得描述所述资源内容的信息作为索引结果。

[0042] 第一实施例

[0043] 图 2 是应用了根据本发明第一实施例的索引构建设备的系统框图，包括资源内容站点 210、搜索器 220、计算机网络 230、索引构建设备 240、索引数据库 250 和检索器 260。在第一实施例中，索引构建设备 240 设置在搜索引擎一侧，即与搜索器 220、索引数据库 250 和检索器 260 一起构成搜索引擎，该索引构建设备可以设置在搜索引擎服务器中或者搜索引擎的索引服务器中。

[0044] 至少一个资源内容站点 210 存储有资源内容，该资源内容可以是 HTML、XML、Newsgroup 文章、FTP 文件、字处理文档、多媒体信息等各种信息，在本实施例中以网页文件为例。在该网页文件中嵌入了各种链接的相关信息，在本实施例中以脚本 (Script) 文件为例，该脚本文件可以包含用户定义的用于实现索引算法的索引代码以及其它脚本。该索引代码可以由 JavaScript、VBScript 或搜索引擎服务器侧上的 Script 引擎所支持的任何其他 Script 语言编程，在本实施例以 JavaScript 为例。

[0045] 网页文件所链接的索引代码实现了针对包含该索引代码的网页文件的索引算法。例如 wiki 页的网页可以通过使用如下的调用代码来嵌入由用户定义的 JavaScript 索引代码：

[0046] <script type =”text/javascript”

[0047] src =”/wiki/pages/indexer/wiki-indexer.js”></script>

[0048] 在以上的代码中 script type =”text/javascript”表示采用 javascript 脚本语言，src =”/wiki/pages/indexer/wiki-indexer.js”></script> 中列出了脚本文件名 (wiki-indexer.js) 及地址 (服务器上的目录 /wiki/pages/indexer/)，表示调用了该 wiki-indexer.js 脚本文件。

[0049] 用户定义的索引代码的功能与传统索引器的相同，都是用于处理网页文件以获得描述网页文件的信息，特别是用于在执行时解析由搜索器搜索到的网页信息，从中抽取出索引项，生成表示文档以及生成文档库的索引表。所不同的是，由于用户或者可以说是资源

内容所有者比其他任何人更好地理解资源内容（网页文件的内容，版面设计，组织），因此根据本发明，用户定义的索引代码允许用户自己选择对资源内容的索引项并且自己定义对所选择索引项的权重。这样本发明的“用户定义的索引”可更有效地提高索引质量，由此也提高了搜索质量。

[0050] 例如，通常使用的索引项有客观索引项和内容索引项两种：客观项与文档的语意内容无关，如作者名、URL、更新时间、编码、长度、链接流行度等；内容索引项是用来反映文档内容的，如关键词及其权重、短语、单字等。用户定义的索引代码允许用户优选地基于内容和/或组织方面的内容索引项设计索引算法，最佳地描述文档。用户可以通过对其中的某些索引项赋予较高或较低权重来完成其索引算法。

[0051] 在一个最优实施例中，用户可以选择例如段落、重点、章节等内容组织作为索引项，通过在脚本文件的索引代码中加大某个段落、某个重点、某个章节的权重，表示该索引项对文档的区分度同时有助于在执行时计算查询结果的相关度，从而最佳地描述文档。这样，在随后执行该脚本文件时，其中的索引代码将解析该网页源文件，识别经预定义的章节，并赋予那些区域的文本以较高或较低的索引值。

[0052] 为方便用户或者内容所有者设计出以上所述的基于 JavaScript 的索引代码，用户或者内容所有者可以首先基于内容模板创建代码模板，或者也可以为用户或内容所有者提供各种类型的代码模板，由用户或内容所有者对感兴趣索引项的权重赋值或者仅对特定索引项的权重赋值。每个代码模板处理不同的网页（例如 HTML）版面设计和内容。可以对同类型模板的网页使用单一的索引代码，也可以根据对象的内容个性化设计索引代码，本领域普通技术人员基于以上描述完全可以设计出个性化的索引代码或者针对各类型内容模板的代码模板。

[0053] 搜索引擎的搜索器（也称为网络爬虫）220 访问所述资源内容站点 210 的网页文件时，搜索器 220 依据网页上的 URL 链接下载网页文件及其链接的脚本文件，并发送回搜索引擎侧。

[0054] 搜索引擎包括搜索器 220、索引构建设备 240、索引数据库 250 和检索器 260。可替换地，也可以将索引数据库 250 并入索引构建设备 240 中。

[0055] 索引构建设备 240 包括判断装置 241、解释器 242 以及预定索引器 243。以下结合涉及本发明第一实施例的索引构建方法的图 3 对第一实施例的索引构建设备 240 作详细说明。

[0056] 步骤 S301 开始，在步骤 S302 搜索引擎的搜索器 220 下载了资源内容（在本实施例中为网页文件）和链接的相关信息（在本实施例中为脚本文件）之后，由索引构建设备 24 的判断装置 241 解析所下载的网页文件，并判断网页文件是否链接有由所述用户定义的索引代码（即本实施例中的 JavaScript 代码）（步骤 S303）。如果在步骤 S303 判断装置 241 判断网页文件链接有由所述用户定义的索引代码，则由解释器 242 运行由用户定义的索引代码（步骤 S304），以获得描述所述资源内容的信息作为索引结果，将其保存在搜索引擎的索引数据库 250 中（步骤 S306）并结束流程处理（步骤 S307）。

[0057] 具体地，为了判断嵌入的脚本文件是否包含了索引代码，索引构建设备 240 的判断装置 241 被配置成用于验证脚本文件中的代码，例如通过要求用户定义的索引代码在设计之后满足统一的规范（如命名的规范），则判断装置 241 验证下载网页中脚本文件的代

码，识别其中的索引代码，并调用解释器 242 来运行识别到的索引代码。

[0058] 解释器 242（在本实施例中针对 JavaScript 形式的索引代码是脚本引擎），用于解释执行 JavaScript 代码中的索引代码，将其编译成计算机能执行的机器代码，例如可以是 Mozilla Rhino 解释器。解释器 242 被调用后解释执行索引代码，打开该索引代码相关的网页文件并进行解析，识别网页文件的各部分并按照用户定义的权重抽取出索引项，产生表示文档以及生成文档库的索引表。这样脚本文件中索引代码的输出结果就是针对该网页文件的文档索引。该文档索引将被储存在索引数据库 250 中。

[0059] 如果在步骤 S303 判断装置 241 在所下载的网页文件中未查找到索引代码或者甚至在所下载的网页文件中都未查找到任何嵌入的 JavaScript 代码，则索引构建设备 240 将使用默认的索引算法（预定的索引器 243）来索引该网页文件（步骤 S305），并将索引结果保存在索引数据库 250 中（步骤 S306）。

[0060] 在此预定的索引器 243 提供控制的索引，优选地设置在索引构建设备 240 中，但是应当理解预定索引器 243 并非必须的，因为本发明可以设计成不对未链接有用户定义索引代码的资源内容构建索引；另外预定的索引器还可以不设置在索引构建设备 240 中而单独地设置在搜索引擎服务器中或者搜索引擎的索引服务器中。

[0061] 针对搜索引擎的搜索请求。搜索引擎服务器的搜索器 260 将根据用户提供的关键字检索事先建立好的索引数据库 250，进行文档与查询的相关度评价，对将要输出的结果进行排序，生成提供给用户的关键字结果页面。

[0062] 以上介绍了本发明的索引构建方法应用于因特网的情况，用于替代受控的搜索引擎索引技术，本发明的索引构建方法充分利用了用户或内容所有者对资源内容的了解，按照用户定义的权重抽取出索引项，提高了索引质量和搜索质量。

[0063] 第二实施例

[0064] 以上第一实施例介绍了本发明的索引构建方法应用于因特网的情况，本发明的索引构建方法还可应用于被管理的网络环境（例如内网）中。在被管理的网络环境（例如内网）中的处理步骤与第一实施例中的相同。

[0065] 在内网中，资源内容（例如 Web 文档）通常不像因特网那样被链接和引用，其中大部分文档为部门、业务单位所有并且是自包含的：少数指向其他部门文档的向内链接和少数来自其他部门文档的向外链接。因此与因特网相比，Web 链接信息在内网搜索中帮助较小，而 Web 内容在搜索排序中起最重要的作用。并且由于缺乏链接信息，传统的内网搜索常常无法提供与基于因特网的 Web 搜索一样的搜索质量。

[0066] 因此本发明的索引构建方法应用于内网索引，替换目前的内网搜索引擎索引器，可以取得更好的效果。

[0067] 进一步，考虑到因特网中存在试图调节排序结果的排名作弊的情况，在被管理的内网中用户或内容所有者提供的索引代码要比因特网中一般用户提供的索引代码更值得信赖，因此本发明的“用户贡献的索引”在被管理的网络环境中可更有效地有助于提高索引质量和搜索质量。

[0068] 另外，一般内网的内容均经过良好的组织，因此本发明的“用户贡献的索引”可通过提供规范的索引模板实现索引，来有利地提高索引和搜索质量。

[0069] 第三实施例

[0070] 以上第一和第二实施例公开了在搜索引擎侧运行用户定义的索引代码产生索引结果的技术方案，本发明的实现并不限于此。本发明还可在用户侧即资源内容站点 210 运行包含用户定义的索引代码的脚本文件。图 4 示出了根据本发明第三实施例的索引构建方法的流程图。

[0071] 从步骤 S401 开始，在资源内容站点 210 准备资源内容链接的脚本文件之后，资源内容站点 210 需要通过判断搜索器的请求、其报头或者其它识别搜索器的机制来判断未知的访问是否是搜索器的访问、还是一般浏览者的访问（步骤 S402）。

[0072] 如果是搜索器的访问则首先判断资源内容链接的相关信息（如脚本文件）中是否有由用户定义的索引代码（步骤 S403），如果有则通过调用资源内容站点 210 的脚本引擎来运行用户定义的索引代码（步骤 S404），将索引代码运行的输出结果作为索引结果由搜索器下载到所述搜索引擎的索引数据库 250 中（步骤 S406），并结束流程处理（步骤 S407）。由此，可以减轻搜索引擎侧中心的工作负荷并提高效率。

[0073] 如果在步骤 S403 的判断结果为“否”，则可以在搜索引擎侧或者甚至在资源内容站点 210 使用预定的索引器对资源内容创建索引（步骤 S405），在资源内容站点使用预定的索引器可参考 WO2001027793，在此不作赘述。

[0074] 为减小用户侧的资源开销，亦可以在资源内容站点 210 侧第一次独立地运行用户定义的索引代码之后或者在第一次响应搜索器的访问运行用户定义的索引代码之后，将索引结果记录在资源内容站点 210 的存储器中，以便针对之后的搜索器访问提供相同的索引结果。如果资源内容进行了更新，则相应地执行更新后的资源内容的索引代码并更新索引结果。

[0075] 相应地，实现第三实施例的索引构建设备也设置在所述资源内容站点侧，并类似地包括：判断装置，被配置成判断来自搜索引擎搜索器的访问并且判断资源内容链接的相关信息是否包含了由用户定义的索引代码；以及解释器，被配置成运行用户定义的索引代码，以获得描述资源内容的信息作为供所述搜索引擎的搜索器下载的索引结果。

[0076] 在第三实施例中由于在用户侧即资源内容站点 210 运行用户定义的索引代码而用户侧通常配置有脚本引擎，因此，在搜索引擎侧无须设置解释器 242 及相应的接口组件，就此而言整个系统的开销会变小。

[0077] 其它实施例

[0078] 对本领域的普通技术人员而言，能够理解本发明的方法和装置的全部或者任何步骤或者部件，可以在任何计算设备（包括处理器、存储介质等）或者计算设备的网络中，以硬件、固件、软件或者它们的组合加以实现，这是本领域普通技术人员在阅读了本发明的说明的情况下运用他们的基本编程技能就能实现的，因此在这里省略了详细说明。

[0079] 因此，基于上述理解，本发明的目的还可以通过在任何信息处理设备上运行一个程序或者一组程序来实现。所述信息处理设备可以是公知的通用设备。因此，本发明的目的也可以仅仅通过提供包含实现所述方法或者设备的程序代码的程序产品来实现。也就是说，这样的程序产品也构成本发明，并且存储有这样的程序产品的存储介质也构成本发明。显然，所述存储介质可以是任何公知的存储介质或者将来所开发出来的任何存储介质，因此也没有必要在此对各种存储介质一一列举。

[0080] 在本发明的设备和方法中，显然，各部件或各步骤是可以分解和 / 或重新组合的。

这些分解和 / 或重新组合应视为本发明的等效方案。并且，执行上述系列处理的步骤可以自然地按照说明的顺序按时间顺序执行，但是并不需要一定按照时间顺序执行。某些步骤可以并行或彼此独立地执行。

[0081] 以上描述了本发明的优选实施方式。本领域的普通技术人员知道，本发明的保护范围不限于这里所公开的具体细节，而可以具有在本发明的精神实质范围内的各种变化和等效方案。

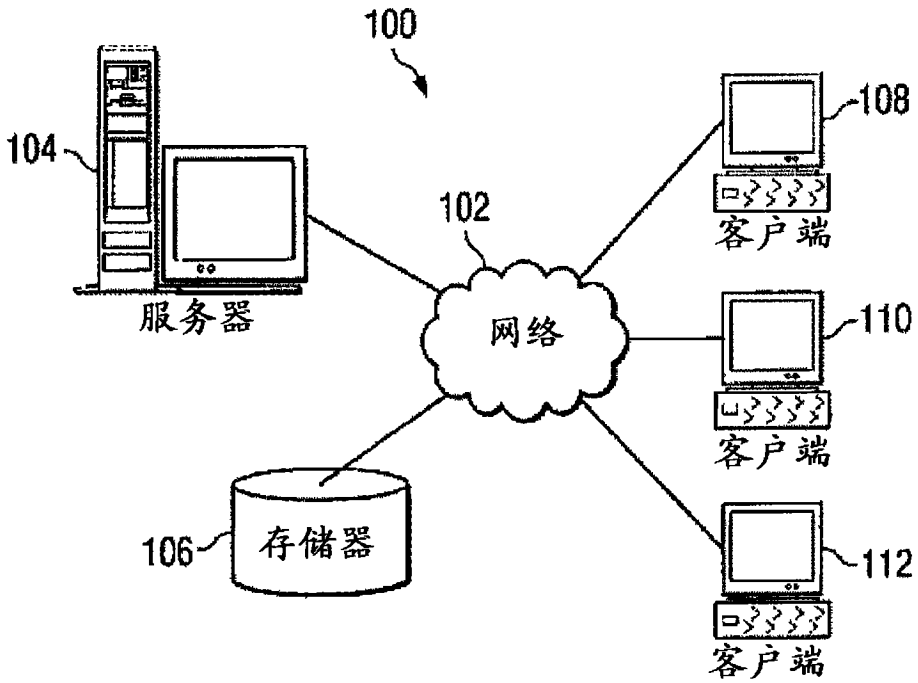


图 1

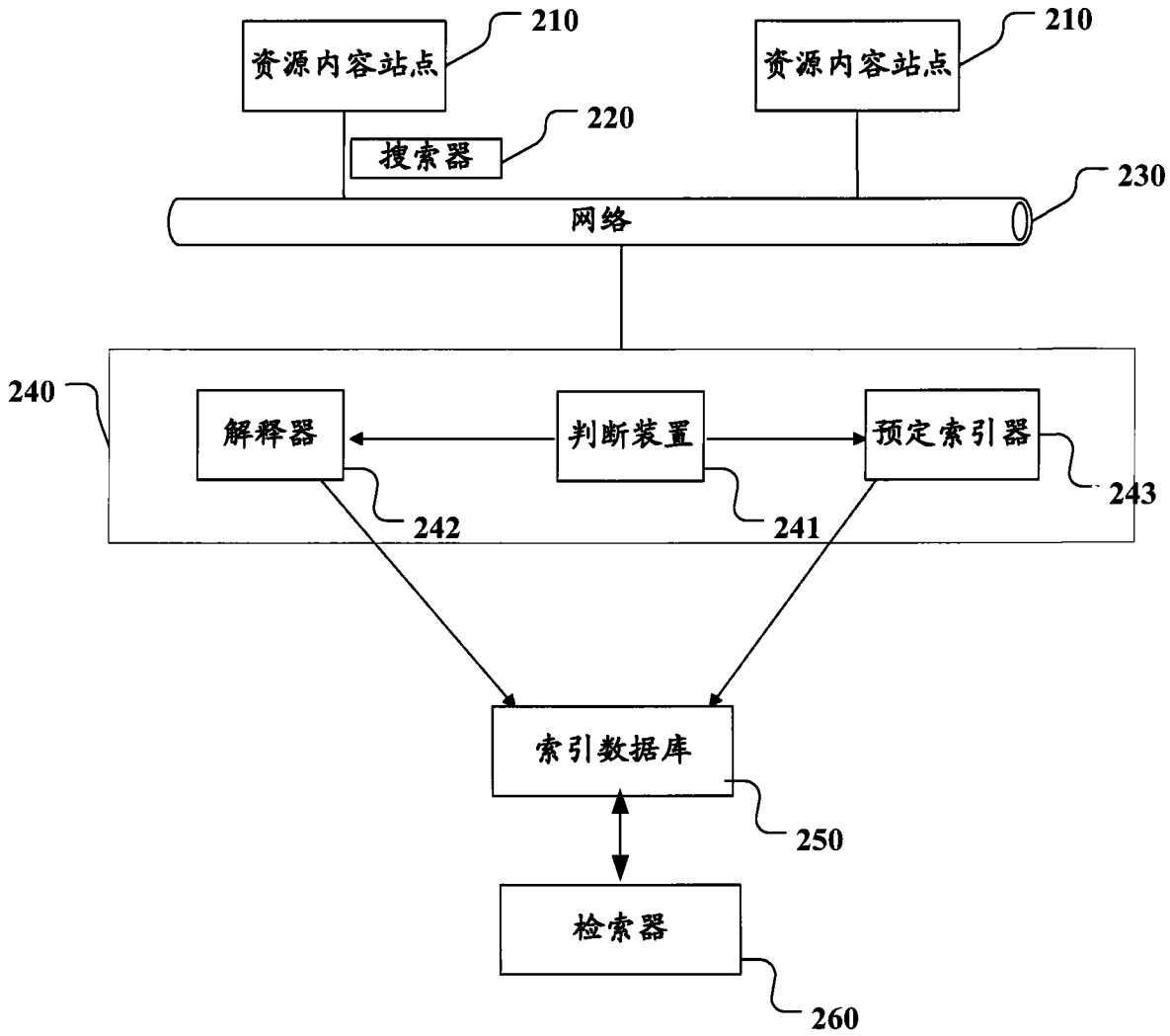

资源内容站点
210
资源内容站点
210
搜索器
220
网络
230
240
解释器
判断装置
预定索引器
243
242
241
索引数据库
250
检索器
260

图 2

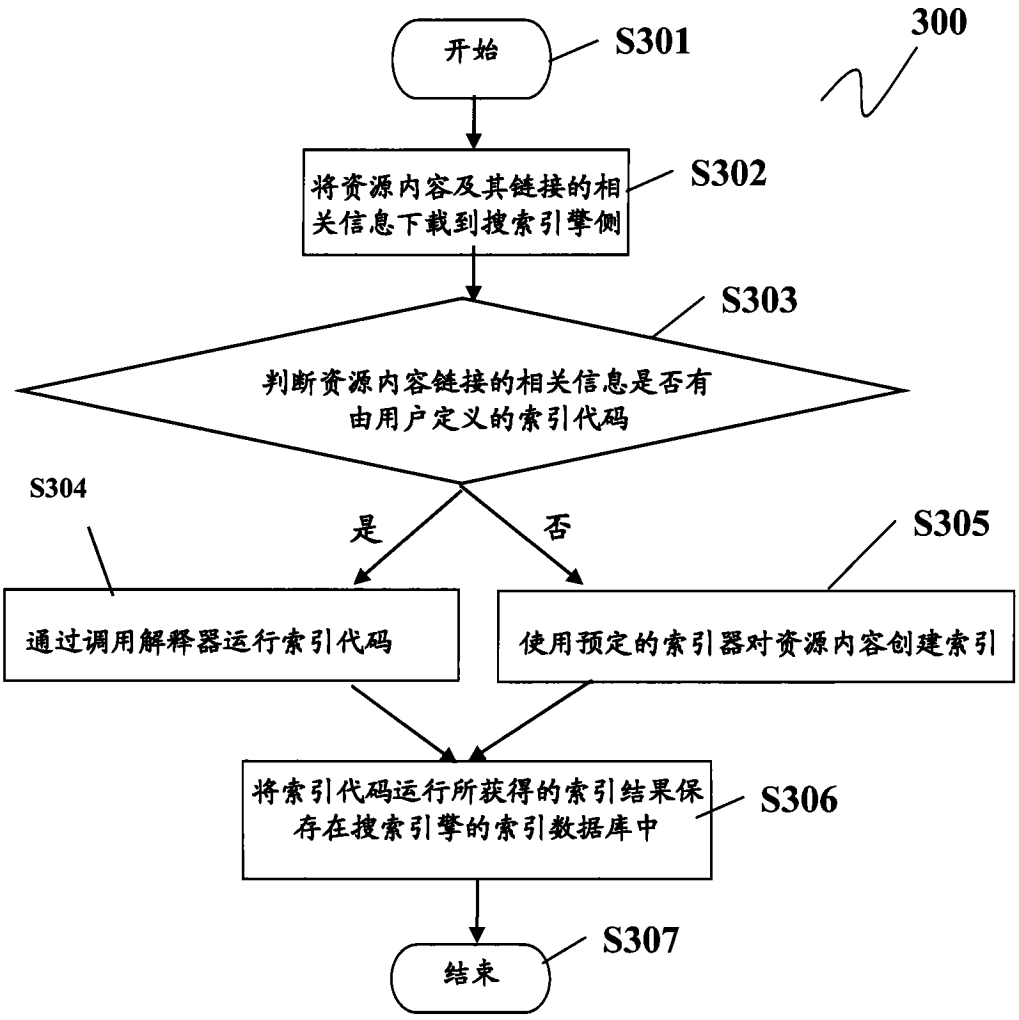

300
开始
S301
将资源内容及其链接的相关信息下载到搜索引擎侧
S302
S303
判断资源内容链接的相关信息是否有由用户定义的索引代码
S304
是
否
S305
通过调用解释器运行索引代码
使用预定的索引器对资源内容创建索引
将索引代码运行所获得的索引结果保存在搜索引擎的索引数据库中
S306
结束
S307

图 3

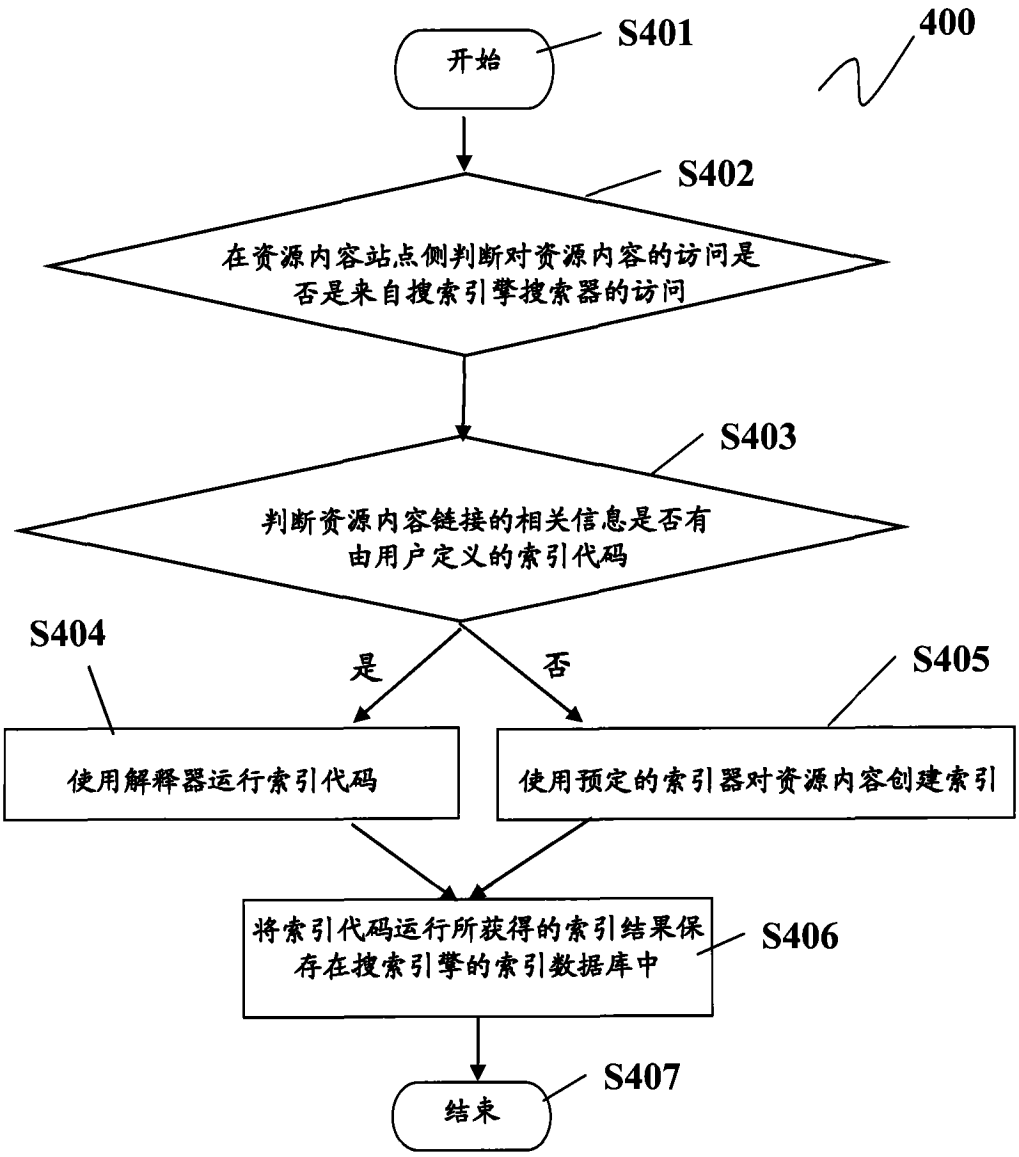

400
开始
S401
S402
在资源内容站点侧判断对资源内容的访问是否是来自搜索引擎搜索器的访问
S403
判断资源内容链接的相关信息是否有由用户定义的索引代码
S404
是
否
S405
使用解释器运行索引代码
使用预定的索引器对资源内容创建索引
将索引代码运行所获得的索引结果保存在搜索引擎的索引数据库中
S406
结束
S407

图 4