



(12) 发明专利

(10) 授权公告号 CN 101180624 B

(45) 授权公告日 2012. 05. 09

(21) 申请号 200580037229. 1

(51) Int. Cl.

(22) 申请日 2005. 10. 26

G06F 17/30 (2006. 01)

(30) 优先权数据

60/623, 295 2004. 10. 28 US

11/198, 471 2005. 08. 04 US

(56) 对比文件

CN 1536483 A, 2004. 10. 13, 全文.

US 6728752 B1, 2004. 04. 27, 全文.

US 20040143600 A1, 2004. 07. 22, 全文.

(85) PCT申请进入国家阶段日

2007. 04. 28

审查员 田志刚

(86) PCT申请的申请数据

PCT/US2005/038619 2005. 10. 26

(87) PCT申请的公布数据

W02006/049996 EN 2006. 05. 11

(73) 专利权人 雅虎公司

地址 美国加利福尼亚州

(72) 发明人 帕维尔·别尔欣 佐尔坦·I·真吉

简·佩德森

(74) 专利代理机构 北京东方亿思知识产权代理

有限责任公司 11258

代理人 李晓冬 南霆

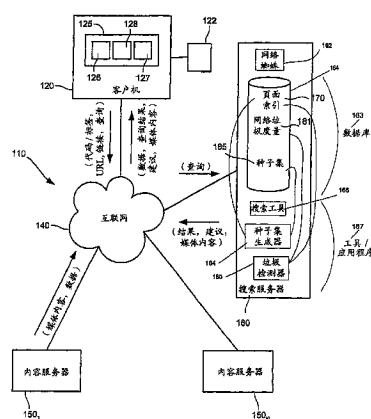
权利要求书 4 页 说明书 10 页 附图 3 页

(54) 发明名称

基于链接的垃圾检测

(57) 摘要

提供了一种用于对搜索结果集中的搜索选中项进行分级的计算机执行的方法。该方法包括接收来自用户的查询以及生成与该查询相关的选中项的列表,其中,每个选中项都具有对于查询的相关度,其中,选中项具有一个或多个指向该选中项的提升链接文档,并且其中,提升链接文档影响该选中项对查询的相关度。该方法将选中项与代表指向每个选中项的提升链接文档的数量的度量相结合。然后,该方法将度量与阈值相比较,部分地基于比较来处理选中项列表以形成修正列表,并且将修正列表传送给用户。



1. 一种对搜索结果集中的搜索选中项进行分级的计算机执行的方法,所述方法包括:
接收来自用户的查询;
生成与所述查询相关的选中项的列表,
其中,所述选中项中的每一个都具有对于所述查询的相关度,
其中,提升链接文档中的链接指向至少一个选中项,并且

其中,所述提升链接文档中的链接人为夸大了所述至少一个选中项对于所述查询的相关度;

确定对于所述至少一个选中项的第一测度,其中,所述第一测度是对于所述至少一个选中项的基于链接的流行度测度;

确定对于所述至少一个选中项的第二测度,其中,所述第二测度是对于所述至少一个选中项的可信赖度测度,其表明所述至少一个选中项是有信誉文档的可能性;

基于所述第一测度和所述第二测度之间的差别,确定所述至少一个选中项的有效质量;

其中,所述有效质量代表包含到所述至少一个选中项的链接的提升链接文档的数量,其中所述链接人为地夸大了所述至少一个选中项对于所述查询的所述相关度;

确定所述至少一个选中项的有效质量表明所述至少一个选中项可能被所述提升链接文档所人为地夸大;

因为所述至少一个选中项的有效质量表明所述至少一个选中项可能被所述提升链接文档所人为地夸大,基于所述至少一个选中项的有效质量处理所述选中项的列表,以形成修正列表,其中,所述至少一个选中项 (a) 被从所述修正列表中排除,或者 (b) 在所述修正列表中被降级;以及

将所述修正列表作为对所述查询的响应传送给所述用户。

2. 根据权利要求 1 所述的方法,其中,所述至少一个选中项中的每一个对应于一文档;并且其中,对于所述至少一个选中项中每一个所对应的文档,在所述接收查询之前执行所述确定所述有效质量。

3. 根据权利要求 1 所述的方法,其中,确定所述第二测度包括:

形成有信誉文档的种子集,所述有信誉文档的种子集包括到其他文档的链接;

将信任值赋予所述种子集中的所述文档中的每一个;

将所述信任值传播给由所述种子集中的所述文档指向的多个文档中的每一个;以及

将按比例分配的信任值赋予由所述种子集中的所述文档指向的多个文档中的每一个。

4. 根据权利要求 3 所述的方法,其中,所述形成所述种子集包括:

分别为多个文档中的每一个确定代表包含在所述多个文档中的每一个中的外部链接的数量的外部链接度量;

使用所述外部链接度量来分级所述多个文档;

在所述多个文档中识别最高级别文档集;

估计所述最高级别文档集中每个文档的质量;

通过从所述最高级别文档集中去除被认为不适合的那些文档来形成修正的最高级别文档集;以及

使用所述修正的最高级别文档集来形成所述种子集。

5. 一种用于对搜索结果集中的搜索选中项进行分级的设备,包括:

用于接收来自用户的查询的装置;

用于生成与所述查询相关的选中项的列表的装置,

其中,所述选中项中的每一个都具有对于所述查询的相关度,

其中,提升链接文档中的链接指向至少一个选中项,并且

其中,所述提升链接文档中的链接人为夸大了所述至少一个选中项对于所述查询的相关度;

用于确定对于所述至少一个选中项的第一测度的装置,其中,所述第一测度是对于所述至少一个选中项的基于链接的流行度测度;

用于确定对于所述至少一个选中项的第二测度的装置,其中,所述第二测度是对于所述至少一个选中项的可信赖度测度,其表明了所述至少一个选中项是有信誉文档的可能性;

用于基于所述第一测度和所述第二测度之间的差别,确定所述至少一个选中项的有效质量的装置;

其中,所述有效质量代表包含到所述至少一个选中项的链接的提升链接文档的数量,其中所述链接人为地夸大了所述至少一个选中项到所述查询的相关度;

用于确定所述至少一个选中项的有效质量表明所述至少一个选中项可能被所述提升链接文档所人为地夸大的装置;

用于因为所述至少一个选中项的有效质量表明所述至少一个选中项可能被所述提升链接文档所人为地夸大,而基于所述至少一个选中项的有效质量来处理所述选中项的列表以形成修正列表的装置,其中所述至少一个选中项 (a) 被从所述修正列表中排除,或者 (b) 在所述修正列表中被降级;以及

用于将所述修正列表作为对所述查询的响应传送给所述用户的装置。

6. 根据权利要求 5 所述的设备,其中,所述至少一个选中项中的每一个对应于一文档;并且其中,对于所述至少一个选中项中每一个所对应的文档,在所述接收所述查询之前执行所述确定所述有效质量。

7. 根据权利要求 5 所述的设备,其中,所述用于确定所述第二测度的装置包括:

用于形成有信誉文档的种子集的装置,所述有信誉文档的种子集包括到其他文档的链接;

用于将信任值赋予所述种子集中的所述文档中的每一个的装置;

用于将所述信任值传播给由所述种子集中的所述文档指向的多个文档中的每一个的装置;以及

用于将按比例分配的信任值赋予由所述种子集中的所述文档指向的多个文档中的每一个的装置。

8. 根据权利要求 7 所述的设备,其中,所述用于形成所述种子集的装置包括:

用于分别为多个文档中的每一个确定代表包含在所述多个文档中的每一个中的外部链接的数量的外部链接度量的装置;

用于使用所述外部链接度量来分级所述多个文档的装置;

用于在所述多个文档中识别最高级别文档集的装置;

用于估计所述最高级别文档集中的每个文档的质量的装置；

用于通过从所述最高级别文档集中去除被认为不适合的那些文档来形成修正的最高级别文档集的装置；以及

用于使用所述修正的最高级别文档集来形成所述种子集的装置。

9. 根据权利要求 1 所述的方法，其中：

对所述选中项列表中的每个选中项执行确定第一测度、确定第二测度、以及确定有效质量的步骤；

所述方法还包括以下步骤：

在为所述选中项列表中的每个选中项确定所述有效质量之后，按照为所述选中项列表中的所述选中项生成的所述有效质量的降序，对所述列表排序，以生成有序的列表；

识别所述有序的列表的顶部，其中，与在所述顶部的选中项相关联的有效质量高于与不在所述顶部的选中项相关联的有效质量；以及

对于所述有序的列表的顶部的每个选中项，基于所述有效质量与所述第一测度的比率确定是否将所述选中项分类为垃圾。

10. 根据权利要求 5 所述的设备，其中：

所述设备还包括用于对所述选中项列表中的每个选中项确定第一测度的装置、确定第二测度的装置、以及确定有效质量的装置；

所述设备还包括：

用于在为所述选中项列表中的每个选中项确定所述有效质量之后，按照为所述选中项列表中的所述选中项生成的所述有效质量的降序，对所述列表排序，以产生有序的列表的装置；

用于识别所述有序的列表的顶部的装置，其中，与在所述顶部的选中项相关联的有效质量高于与不在所述顶部的选中项相关联的有效质量；以及

用于对于所述有序的列表的顶部的每个选中项，基于所述有效质量与所述第一测度的比率确定是否将所述选中项分类为垃圾的装置。

11. 根据权利要求 1 所述的方法，其中：

确定所述至少一个选中项的有效质量表明所述至少一个选中项可能被所述提升链接文档所人为地夸大的步骤还包括：

计算所述有效质量与所述第一测度的比率；以及

确定所述比率不小于指定的阈值；并且

所述处理选中项的列表以形成修正列表的步骤还包括：

响应于确定所述比率不小于所述指定阈值，执行以下中的一个：

从所述修正列表中排除所述至少一个选中项，以及

将所述修正列表中的所述至少一个选中项降级。

12. 根据权利要求 5 所述的设备，其中：

用于确定所述至少一个选中项的有效质量表明所述至少一个选中项可能被所述提升链接文档所人为地夸大的装置还包括：

用于计算所述有效质量与所述第一测度的比率的装置；以及

用于确定所述比率不小于指定阈值的装置；并且

用于处理所述选中项的列表以形成修正列表的装置还包括：
用于响应于确定所述比率不小于所述指定阈值，执行以下中的一个的装置：
从所述修正列表中排除所述至少一个选中项，以及
将所述修正列表中的所述至少一个选中项降级。

基于链接的垃圾检测

技术领域

[0001] 本发明总体上涉及搜索系统,更具体地,涉及对结果集中的搜索选中项(hit)进行分级的搜索系统。

背景技术

[0002] 在整个语料库(corpus)不能被吸收以及不存在或不可能存在指向期望的项目的精确指针的情况下,搜索是有用的。通常,搜索是指以下过程:设计(formulate,制定)或接受搜索查询,从文档的语料库确定匹配的文档集,并且返回该集合,或者如果该集合太大则返回该集合的某些子集。在具体的实例(本公开并不限于该实例)中,考虑搜索称为“Web”的超链接文档的集合。语料库包含许多可搜索的项目,在此称为页面,或者更一般地,称为文档。搜索引擎通常使用在接收到搜索查询之前所生成的索引从语料库中识别与搜索查询匹配的文档。“匹配”可以意指许多情况,并且搜索查询可以具有多种形式。通常,搜索查询是包含一个或多个字或术语的字符串,并且当文档包含搜索查询字符串中的一个或多个字或术语(或者所有的字或术语)时产生匹配。每个匹配文档都被称为一个选中项(hit),并且选中项的集合被称为结果集或搜索结果。语料库可以是数据库或者其他数据结构或者非结构化数据。文档通常是Web页面。

[0003] 典型的Web页面索引包含数十亿个条目,所以,一般的搜索可能具有包括数百万个页面的结果集。显然,在这种情况下,搜索引擎可能需要进一步限制结果集,以使返回给查询者(其一般为人计算机用户,但不必然为这种情形)的结果集具有合理大小。限制该集合的一种方法是基于用户将仅读取或使用少数在有序的搜索结果中出现靠前的选中项的假设,按顺序显示搜索结果。

[0004] 由于这种假设,许多Web页面制作者期望他们的页面在有序的搜索结果中靠前出现。搜索引擎根据相关页面的各种特征来只选择和返回最高质量的页面。由于在查询结果列表中的顶部位置(高级别)可能带来商业利益,因此某些Web页面的制作者企图恶意地提高他们页面的级别。这种具有人为地提高的级别的页面被称为“网络垃圾”页面,并且通称为“网络垃圾”。

[0005] 存在许多与网络垃圾相关的技术。一种是人为地使Web页面适于被多种查询所选择。这可以通过增加具有大量与实质内容无关的以及用小的或不可见的字体表示的术语的页面来实现。这种增加使得页面更加暴露(即,可能与更多查询相关),但是不能真正地提高其对于任何特殊查询的相关度(relevance)。在这一点上,垃圾的制作者使用了另一种技术:基于更频繁地被其他页面引用的页面通常被搜索引擎认为是更好的(具有更高的相关度)的观察结果,他们向页面加入了許多引入(超)链接,也称为内部链接(inlink)。难以区分由于其较高价值而被许多其他页面引用的真正高质量页面和具有许多内部链接的网络垃圾。

[0006] 网络垃圾页面的识别以及其在搜索结果列表中的后续降级对于维持或提高由搜索引擎生成的答复的质量是重要的。因而,网络垃圾检测对于搜索引擎而言是有用的工作。

经常采用人工编辑以通过校验存在于搜索引擎索引中的大量页面来识别网络垃圾,但这通常是不切实际的。

[0007] 因此,需要一种改进的搜索处理,其能够克服网络垃圾并能提供更符合用户需要而不是符合文档制作者的操纵的搜索结果。

发明内容

[0008] 本发明的实施例提供了一种用于处理搜索请求的系统和方法,包括对构成搜索结果集的选中项进行分级。可以利用选中项的有效质量(其是指向特定页面的垃圾场(spam farm)的大小的测度(measure))以及其他参数来对选中项进行分级。

[0009] 在一个实施例中,本发明提供了一种对搜索结果集中的搜索选中项分级的计算机执行的方法。该计算机执行的方法包括接收来自用户的查询并生成与查询相关的选中项的列表,其中,每个选中项对于查询都具有相关度,其中,选中项具有指向该选中项的一个或多个提升链接文档(boosting linked document),并且其中,提升链接文档影响选中项对于查询的相关度。然后,该方法将度量(metric)与选中项的至少一个子集中的每个相结合,该度量代表指向该选中项的至少一个子集中的每个并且人为地夸大了选中项的相关度的提升链接文档的数目。接着,该方法将表示指向该选中项的垃圾场的大小的度量与阈值进行比较,部分地基于该比较来处理选中项的列表以形成修正列表,并且向用户传送修正列表。

[0010] 一方面,度量是第一测度和第二测度的组合。选中项的第一测度代表选中项的链接流行度(link popularity),第二测度是选中项为有信誉文档(reputable document)的可能性的测度。

[0011] 另一方面,通过以下处理来生成第二测度:形成有信誉文档的种子集(seed set),有信誉文档的种子集是链接文档;为种子集中的每个文档赋予信任值;将信任值传播给由链接文档所指向的每个被链接文档;以及为每个被链接文档赋予按比例分配的信任值。

[0012] 另一方面,通过以下处理来形成有信誉文档的种子集:确定多个文档中的每个的外部链接(outlink)度量,其中,外部链接度量表示每个文档的外部链接的数目;使用外部链接度量对多个文档分级;识别最高级别的文档的集合;估计最高级别文档的质量;通过从最高级别文档中排除被认为不合适的那些文档来形成修正文档集合;以及使用修正的保留集来形成种子集。

[0013] 下面的详细描述与附图一起,将为本发明的性质和优点提供更好的理解。

附图说明

[0014] 图1是可以用于实现本发明实施例的信息检索和通信网络的示例性框图;

[0015] 图2是根据本发明一个实施例的信息检索和通信网络的示例性框图;

[0016] 图3A-B是简单的垃圾场的示例性示意图。

具体实施方式

[0017] 定义

[0018] 除非另外定义,本文中所使用的所有技术和科学术语具有本发明所属领域的技术

人员通常理解的意义。对如本文中所使用的,如下定义下列术语。

[0019] 网页级别 (PageRank) 是一族用于将数值权重赋予由搜索引擎编入索引的超链接文档 (或网络页面或网站) 的著名算法。网页级别使用链接信息来为 Web 上的文档赋予全局重要性分数 (global importance score)。网页级别方法已经被授予专利权并且在美国专利第 6, 285, 999 号中被描述。文档的网页级别是 Web 上的文档的基于链接的流行度的测度。

[0020] 信任级别是与网页级别相关的链接分析技术。信任级别是用于把 Web 上的信誉好的页面、好的页面与网络垃圾分开的方法。信任级别是基于 Web 上的好的文档很少链接到垃圾的假设。信任级别包括两个步骤, 一个步骤是种子选择, 另一个步骤是分数传播。文档的信任级别是文档为有信誉 (即, 非垃圾) 文档的可能性的测度。

[0021] 链接或超链接是指通常通向另一个页面、另一个站点、或同一页面的另一部分的网络页面上的可点击内容。因此, 可点击内容据说链接到其他页面 / 站点 / 同一页面的其他部分。当网络蜘蛛 (Spider) 索引网站时, 其利用链接从一个页面爬行到另一个页面。

[0022] 导入链接 (inbound link) 或内部链接 / 导出链接或外部链接。当站点 A 链接到站点 B 时, 站点 A 具有导出链接, 而站点 B 具有导入链接。对导入链接进行计数, 以确定链接流行度。

[0023] Web, 或者万维网 (“WWW”, 或简称为 Web) 是一种信息空间, 其中称为资源的感兴趣项目由称为统一资源标识符 (URI) 的全局标识符来标识。术语 Web 通常用作互联网的同义词; 然而, Web 实际上是通过互联网运行的业务。

[0024] 网络页面或网页是指万维网的页面或文件, 通常为 HTML/XHTML 格式 (文件扩展名通常是 htm 或 html) 并且具有能够从一个页面或部分导航到另一个页面或部分的超文本链接。网页经常使用相关的图形文件以提供说明, 并且这些图形文件也可以是可点击的链接。使用网络浏览器来显示网页, 并且可以设计网页, 以利用经常提供运动、图形、交互、以及声音的小型应用程序 (在页面内运行的子程序)。

[0025] 网站是指存储在网络服务器的单个文件夹或在相关子文件夹中的网页的集合。网站通常包括一般命名为 index.htm 或者 index.html 的首页。

[0026] 虚拟主机 (web host) 负责为由没有自己的网络服务器的个人或公司控制的网站提供服务器空间、网络服务、和文件维护。许多互联网服务提供商 (ISP) 会分配 (allow) 给用户少量的服务器空间以存放个人网络页面。

[0027] 垃圾指的是大量散发的通常具有商业性质的不需要的文档或电子邮件。

[0028] 网络垃圾指的是网络上的垃圾页面。制造网络垃圾的行为被称为网络作弊 (web spamming)。网络作弊指的是想要误导搜索引擎以给某些文档比它们的应得级别更高级别的行为。网络上的垃圾页面是某些形式的作弊的产物。作弊的一种形式是链接作弊。

[0029] 垃圾页面是接收其级别分数的重大非法提升的网络文档, 因此, 垃圾页面很可能出现在顶部的搜索结果中并有意误导搜索引擎。

[0030] 链接作弊 (link spamming) 是指制造经常相互连接并形成称为垃圾场的组的垃圾文档, 建立该垃圾场使得大量的提升的文档将增加一个或几个目标页面的基于链接的重要性级别。

[0031] 垃圾场是指一组互连的垃圾网页, 它们被生成以提升特定目标页面的基于链接的

重要性分数（例如，网页级别分数）。

[0032] 概述

[0033] 本发明的实施例旨在提供用于基于链接的垃圾的检测方法和系统。对响应于搜索查询所产生的搜索结果进行处理，以确定选中项的有效质量。选中项的有效质量是已被建立以指向选中项从而人为地提升选中项的相对重要性的垃圾场的大小的测度。根据本发明实施例的方法和系统利用选中项的有效质量，并使那些其有效质量表明它们可能是由基于链接的垃圾人为地提升的选中项降级。对于给定的网络文档的有效质量的确定依赖于部分地估计给定的网络文档的基于链接的流行度（例如，网页级别）与可信度（trustworthiness）（例如，信任级别）之间的差异的技术的结合。接下来进一步详细描述用于确定给定的网络文档的有效质量的技术。

[0034] 网络实现

[0035] 图 1 示出了包括一个或多个可用于实现本发明实施例的客户机系统 20_{1-N} 的信息检索和通信网络 10 的总的概况。在计算机网络 10 中，客户机系统 20_{1-N} 通过互联网 40 或其他通信网络（例如，通过任何局域网（LAN）或广域网（WAN）连接）连接至任意数量的服务器系统 $50_1 \sim 50_N$ 。如将在本文中所描述的，根据本发明来配置客户机系统 20_{1-N} ，以与服务器系统 $50_1 \sim 50_N$ 中任意一个进行通信，例如，来存取、接收、检索、以及显示媒体内容和诸如网络页面的其他信息。

[0036] 图 1 所示系统中的几个组件包括这里无需再详细解释的常规的、众所周知的组件。例如，客户机系统 20 可以包括台式个人计算机、工作站、便携式电脑、个人数字助理（PDA）、移动电话、或任何 WAP 使能的装置或者能够直接或间接连接到互联网的任何其他计算装置。客户机系统 20 通常运行浏览程序，例如，在移动电话、PDA、或其他无线装置等的情况下，运行 Microsoft 的 Internet Explorer™ 浏览器、Netscape Navigator™ 浏览器、Mozilla™ 浏览器、Opera™ 浏览器、Apple 的 Safari™ 或 WAP 使能浏览器，以允许客户机系统 20_{1-N} 的用户通过互联网 40 存取、处理、以及观看可从服务器系统 $50_1 \sim 50_N$ 得到的信息和页面。同样，客户机系统 20 通常包括一个或多个用户接口装置 22（例如键盘、鼠标、触摸屏、笔等），用于与由浏览器在显示器（例如，监视屏、LCD 显示器等）上提供的图形用户界面（GUI）以及与由服务器系统 $50_1 \sim 50_N$ 或其他服务器提供的页面、视窗、以及其他信息进行交互。本发明适于供互联网使用，该互联网是指特定的全球相关的网络集。然而，应该理解，可以使用其他网络来代替互联网或除了互联网以外还可以使用其他网络，例如内联网、外联网、虚拟专用网（VPN）、非基于 TCP/IP 的网络、任何 LAN 或 WAN 等。

[0037] 根据一个实施例，客户机系统 20 及其所有组件是操作者可以使用包括计算机软件的应用程序进行配置的，其中，该计算机软件利用中央处理单元（例如，Intel Pentium™ 处理器、AMD Athlon™ 处理器、Apple 的 Power PC 等或者多处理器）来运行。优选地，此处所描述的用于操作和配置客户机系统 20 以传送、处理、以及显示数据和媒体内容的计算机软件被下载并存储在硬盘上，但是众所周知，也可以将全部程序代码或部分程序代码存储在任何其他易失性或非易失性存储介质或装置（例如 ROM 或 RAM）中，或者设置在能够存储程序代码的任何介质（例如光盘（CD）介质、数字多功能盘（DVD）介质、软盘等）上。另外，可以通过互联网从软件源（例如，从服务器系统 $50_1 \sim 50_N$ 中的一个）将全部程序代码、或者部分程序代码传送和下载到客户机系统 20，或者使用任何通信介质和协议（例如，TCP/

IP、HTTP、HTTPS、以太网、或其他常规介质和协议) 通过任何其他网络连接(例如,外联网、VPN、LAN、或其他传统网络)来传送。

[0038] 应该理解,用于实现本发明的各方面的计算机代码可以是 C、C++、HTML、XML、Java、JavaScript 等代码,或者任何其他合适的脚本语言(例如,VBScript),或者能在客户机系统 20 上执行或被编译以在客户机系统 20 上执行或在系统 20_{1-N} 上执行的任何其他合适的编程语言。在某些实施例中,没有将代码下载到客户机系统 20,并且由服务器来执行所需要的代码,或者执行已存在于客户机系统 20 的代码。

[0039] 搜索系统

[0040] 图 2 示出了根据本发明的一个实施例的用于传送媒体内容的另一个信息检索和通信网络 110。如图所示,网络 110 包括客户机系统 120、一个或多个内容服务器系统 150、以及搜索服务器系统 160。在网络 110 中,客户机系统 120 通过互联网 140 或其他通信网络可通信地连接至服务器系统 150 和 160。如上所述,客户机系统 120 及其组件被配置为通过互联网 140 或其他通信网络与服务器系统 150 和 160 以及其他服务器系统进行通信。

[0041] 1. 客户机系统

[0042] 根据一个实施例,在客户机系统 120 上执行的客户机应用程序(表示为模块 125)包括用于控制客户机系统 120 及其组件与服务器系统 150 和 160 进行通信以及处理和显示从服务器系统 150 和 160 接收的数据内容的指令。如上所述,虽然客户机应用程序模块 125 可以设置在诸如软盘、CD、DVD 等任何软件存储介质上,但是,优选地,将客户机应用程序 125 从诸如远程服务器系统(例如,服务器系统 150、服务器系统 160、或其他远程服务器系统)的软件源传送和下载到客户机系统 120。例如,一方面,客户机应用程序模块 125 可以以 HTML 包装(wrapper)的形式通过互联网 140 提供给客户机系统 120,其中,HTML 包装包括诸如嵌入式 JavaScript 或者 ActiveX 控件的各种控件,用于处理各种对象、画面、以及窗口中的数据并且表现(render)数据。

[0043] 另外,客户机应用程序模块 125 包括用于处理数据和媒体内容的各种软件模块,例如用于处理搜索请求和搜索结果数据的搜索模块 126,用于以文本、数据帧、以及活动窗口(例如,浏览器窗口和对话框)的形式表现数据和媒体内容的用户界面模块 127,以及用于与在客户机 120 上执行的各种应用程序进行交互和通信的应用程序接口模块 128。在客户机系统 120(优选地,应用程序接口模块 128 被配置为与其交互)上执行的各种应用程序的实例包括各种电子邮件应用程序、即时通信(IM)应用程序、浏览器应用程序、文档管理应用程序、以及其他应用程序。此外,界面模块 127 可以包括浏览器,例如配置在客户机系统 120 上的默认浏览器或不同的浏览器。

[0044] 2. 搜索服务器系统

[0045] 根据一个实施例,搜索服务器系统 160 被配置为将搜索结果数据和媒体内容提供给客户机系统 120。内容服务器系统 150 被配置为,例如,响应于在由搜索服务器系统 160 提供的搜索结果页面中所选择的链接,将数据和媒体内容(例如网络页面)提供给客户机系统 120。在某些变化中,搜索服务器系统 160 返回内容、以及到内容的链接和/或其他引用,或者返回到内容的链接和/或其他参考而不是返回内容。

[0046] 一个实施例中的搜索服务器系统 160 涉及(reference)由例如页面、页面的链接、表示被索引的页面的内容的数据等构成的各种页面索引 170。页面索引既可以通过包括自

动网络爬虫、网络蜘蛛等的各种收集技术来生成,也可以通过用于在分级结构中对网络页面进行分类和分级的人工或半自动分类算法和接口程序(interface)来生成。这些技术可以在搜索服务器系统 160 上实现,或者在生成页面索引 170 并使其可用于搜索服务器系统 160 的单独系统(未示出)中实现。

[0047] 搜索服务器系统 160 被配置为响应于从客户机系统(例如,从搜索模块 126)接收到的各种搜索请求而提供数据。例如,搜索服务器系统 160 可以配备有用于对与给定的查询(例如,基于通过查询中的搜索术语的出现模式而测量的逻辑相关度;上下文标识符;页面赞助(sponsorship)等的组合)相关的网络页面进行处理和分级的搜索相关的算法。

[0048] 基于链接的垃圾检测

[0049] 如图 2 所示,搜索服务器系统 160 与基于链接的垃圾检测器 180 一起工作并将其输出(结果、建议、媒体内容等)提供给基于链接的垃圾检测器 180,其中,该基于链接的垃圾检测器返回网络垃圾页面已被降级或已经从列表中被删除的修正搜索列表。根据本发明的实施例,搜索服务器系统 160 被配置以操作搜索引擎。搜索引擎包括三个部分:一个或多个网络蜘蛛 162、数据库 163、以及工具/应用程序 167。网络蜘蛛 162 在互联网内爬行,用于收集信息;数据库 163 包含网络蜘蛛收集的信息和其他信息;工具/应用程序 167 包括应用程序(例如被用户用于搜索数据库的搜索工具 166)。数据库 167 包含搜索工具使用的页面索引 170。另外,根据本发明实施例的搜索引擎包括垃圾检测器 180。如下所述,垃圾检测器 180 执行各种算法,并且存储用于页面索引 170 中的页面的网络垃圾度量 181。如上所述,根据本发明实施例的垃圾检测器 180 估计与选中项的有效质量对应的度量,并与搜索工具 166 和页面索引 170 共同工作,然后降级那些其有效质量显示它们很可能由基于链接的垃圾人为地提升的那些选中项。给定的网络文档的有效质量的确定依赖于部分地估计给定的网络文档的基于链接的流行度(例如,网页级别)与可信度(例如,信任级别)之间的差异的技术的结合。在一个实施例中,网络垃圾检测器 180 处理页面索引 170 中的所有页面以计算用于索引中的页面的网络垃圾度量 181,并将网络垃圾度量 181 存储在数据库 163 中。度量 181 与使文档被包含在搜索结果中的搜索查询无关。

[0050] 对于给定的网络文档,通过垃圾检测器 180 对垃圾场的有效质量的确定部分地依赖于对给定网络文档的基于链接的流行度(例如,网页级别)与可信度(例如,信任级别)之间差别的估计。对于给定的网络文档的可信度的确定部分地依赖于给定页面离已知为可信的网络文档(即,非垃圾文档)的初始种子集有多远来确定。因此,根据本发明实施例的搜索引擎还包括种子集生成器 184,其与页面索引 170 共同工作,以形成可信网络文档的初始种子集 185。以下将更详细地描述形成网络垃圾度量 181 的垃圾检测器 180 的操作以及形成种子集 185 的种子集生成器 184 的操作。

[0051] 垃圾场、网页级别、和信任级别

[0052] 在该部分中,描述了垃圾场、内部链接页面级别(通常称为“网页级别”)、以及信任级别的概念。垃圾场是人为地创建的指向垃圾目标页面以提高其有效性的页面的集合。信任等级(“信任级别”)是对于高质量页面的子集具有特定远距离传送(teleportation)(即,跳跃)的网页级别的一种形式。利用本文中所描述的技术,搜索引擎可以自动发现坏的页面(网络垃圾页面),更具体地,发现通过创建人工垃圾场(引用页面集)来提升其有效性而创建的那些网络垃圾页面。在具体的实施例中,执行具有统一的远距离传送的网页

级别处理和信任级别处理,并且将它们的结果作为页面或页面集的“垃圾性”的测试的一部分进行比较。另外,下面描述了构造信任级别处理的输入的新方法。

[0053] 本发明的一方面旨在基于对垃圾页面周围的超链接结构的分析来识别(至少一些)垃圾页面。特别地,使用了估计垃圾场大小的新方法。由于非垃圾页面很少指向垃圾,因此,信任级别中的特定威信分配(authority distribution)导致非垃圾页面和垃圾页面之间有一定程度的区别:高质量非垃圾网络页面被认为具有由信任级别赋予的最高分数。

[0054] 信任级别与公知的网络分析算法(即网页级别,其根据指向每个网络页面的其他页面的分数来将数值分数赋给该页面)有关。网页级别使用称为远距离传送的技术:根据所谓的远距离传送分布(其通常是统一分布),将一定数量的总分数传送给一些或所有页面。信任级别只为可信的(非垃圾)网络页面的小集合(即,所谓的“种子集”)提供远距离传送,而不是使用统一的远距离传送分布。实际上,这使得分数只被分配给种子集中的其他页面。

[0055] 以下的描述涉及网络页面。但是,推理、实现、以及算法都同样地可应用于(1)网站(网络内容/页面以及与单个威信相关的其他类型的网络文档的逻辑组),(2)由主机网络(主机级别)代表的近似网站,其中对主机之间的图形边界(例如,一个主机图形,其中如果两个主机包含每个都通过超链接连接的至少一个页面或其他测试,则该两个主机具有链接)有一定限定(3)任何其他网络页面图形集合,和/或(4)具有反映参考强度的相关权重的链接集。

[0056] 垃圾场

[0057] 垃圾场是人为地创建的指向垃圾目标页面以提升其重要性的页面(或者可选地,主机)的集合。图 3A-B 是示出两个简单的垃圾场的示例性示意图。

[0058] 图 3A 示出了垃圾场具有都指向目标垃圾页面 s 的 m 个页面。下面描述了用于获得对垃圾场大小的良好估计的处理。对于每个页面 i , 计算数字 M_i , 其中, 数字 M_i 被称作页面的“有效质量”。对于网络垃圾页面, M 用作提升该页面的垃圾场的大小的良好估计。

[0059] 在简单垃圾场的情况下,有效质量近似为 m 。对于更复杂的场,就实例(图 3B 中所示的垃圾场)来说,有效质量 M 用作指示符,其中,高的 M 值表示垃圾场。应该认识到,虽然描述涉及网络页面,但是这些概念也可以应用于页面群、主机群等。

[0060] 网页级别和信任级别

[0061] 在网络页面的分析中网页级别的概念是有用的。在有关网页级别的许多可能的定义中,使用了以下的网页级别的线性系统定义:

[0062] $x = cT^Tx + (1-c)v$. (方程 1)

[0063] 在方程 1 中:

[0064] T 为转移矩阵,如果存在从页面 i 指向页面 j 的链接 $i \rightarrow j$, 则其元素 $T_{ij} = 1/\text{outdeg}(i)$, 否则其元素为 0。这里, $\text{outdeg}(i)$ 是页面 i 上的外部链接的数量,用作归一化因子,以使矩阵 T 是随机的。

[0065] c 为远距离传送常数,一般在范围 0.7-0.9 内取值,

[0066] $x = (x_i)$ 是威信矢量,其中,下标 i 贯穿所有 n 个页面, $i = 1:n$ (n 为网络页面的数量),

[0067] $v = (v_i)$ 是远距离传送矢量,假设其为概率分布, $0 \leq v_i \leq 1$, $v_1 + \dots + v_n = 1$ 。

[0068] 求解方程 1 的迭代方法是已知的。方程 1 具有定义了相对于远距离传送矢量而言为线性的威信矢量的优点。

[0069] 对于网页级别, p 是将提供对应于统一远距离传送 (即, 当 $v_i = 1/n$ 时) 的方程 1 的解的威信矢量。对于信任级别, t 是将提供对应于特定远距离传送 (即, 其中, v 使得 v 的 k 个元素为非零并且其余的为零, 其中, 非零元素具有可信集合中的对应下标 i) 的方程 1 的解的威信矢量。

[0070] 有效质量的估算

[0071] 网络页面的有效质量被用作指示符, 以帮助确定网络页面是否是垃圾页面。

[0072] 估算的构建

[0073] 对于任一网络页面 i 中潜在的垃圾网页 s , 可以从数学上示出,

$$[0074] \quad p_s - t_s = p_s^{boost} + b \cdot p_s^{leak} + (1 - c)/n, \quad (\text{方程 } 2)$$

[0075] 其中, 方程右边的第一项是由于来自支持垃圾场的 (在没有垃圾页面的情况下, 该场为空或不存在) 对页面的提升而产生的, 而第二项是由于对有时错误地指向垃圾页面的非垃圾页面的威信遗漏而引起的。在图 3A-B 中, 这种遗漏用表示从网络的剩余部分到给定页面的不同的意外超链接的虚线箭头来示出。对于垃圾页面 s , 第一项是非常主要的, 因为垃圾制作者制造垃圾场的目的就是使 s 的网页级别变高。对于简单的场,

$$[0076] \quad p_s^{boost} = \frac{m \cdot c(1 - c)}{n}, \quad (\text{方程 } 3)$$

[0077] 类似的公式对于其他结构的场也是有效的。例如, 对于具有反向链接的场,

$$[0078] \quad p_s^{boost} = \frac{m \cdot c(1 - c)}{(1 - c^2)n}, \quad (\text{方程 } 4)$$

[0079] 在下列条件下

$$[0080] \quad p_s^{leak} \ll p_s^{boost} \quad (\text{方程 } 5)$$

[0081] 从方程 (2) 和 (3) 如下构造关于简单垃圾场的大小 m 的良好估算

$$[0082] \quad M_s = \frac{n(p_s - t_s)}{c(1 - c)}, \quad (\text{方程 } 6)$$

[0083] 方程 6 定义了对任一网络页面 i 来说都可以被计算的有效质量 M_i 。如上所述, 如果 i 是通过简单垃圾场提升的垃圾页面, 则 M_i 近似为实际的场大小 m , 并且对于其他结构的场, M_i 与实际场大小只相差一个常数, 如方程 4 所示。考虑到实际垃圾场相当大 (例如, 欺骗性地创建数百万个提升页面) 的事实, 这样的差别是不重要的。

[0084] 对于非垃圾页面, M_i 会是绝对项不会很大或者相对于 p_i 而言不会很大的某些数字。根据本发明实施例的基于链接的垃圾检测将揭示该问题, 并且基于作为指示符的 M_i , 不会将这样的页面推荐为可能的网络垃圾页面。

[0085] 垃圾检测处理

[0086] 接下来的示例性处理用于检测基于链接的垃圾。该处理是极为简单和有效的, 因为它旨在找到具有最高有效质量的页面。然而, 仅当满足方程 5 时, 有效质量才能提供对垃圾大小的良好近似, 从而确保由于可信的网络页面的流行度的分配而产生的页面的基于链接的流行度远远小于由于通过垃圾页面的人为提升而产生的页面的基于链接的流行度。在

方程 5 的条件下,垃圾检测处理能区分真正受欢迎的页面和那些通过链接垃圾场而使其受欢迎的页面。根据本发明实施例的技术确保了方程 5 的条件得以满足。这在下面的步骤 C 中实现,其中, $\eta > 1$ 为用作阈值的算法参数。可以看出,步骤 C 中的大比值对应于满足方程 5 的页面。总的来说,示例性处理包括以下步骤:

[0087] A. 对于列表(例如,与查询相关的选中项的列表,或页面索引的列表)中的所有页面(主机等) i ,根据方程(6)得到其有效质量 M_i ;

[0088] B. 按照 M_i 的降序对页面 i 进行排序,并且保留或识别有序的列表的顶部部分。可选地,可以保留整个列表,尽管这样可能需要过多资源,因此,不保留低 M_i 的页面更加有效。该识别和/或保留可以在任一步骤完成。选择处理的一部分旨在选择同时具有高的 M_i 和高的 M_i/p_i 的页面;

[0089] C. 得到保留在列表中的所有页面 i 的比值 M_i/p_i ;

[0090] D. 从列表中删除 $M_i/p_i < \eta$ 的页面 i ;

[0091] E. 所保留的页面构成垃圾。

[0092] 在实验中,这样检测到的垃圾页面实际上在大多数情况下(通过人为判定)被确认为垃圾。这意味着使用这些技术,误报率可能很低。

[0093] 种子集

[0094] 上述处理取决于信任级别,即方程(1)的解,其具有与所谓的种子集相关的特定远距离传送分布。种子集是一组已知为非垃圾的 k 个高质量网页。本发明实施例的一个方面旨在找到可信的(即,非垃圾)页面或站点的适当种子集。识别可信的网络页面的种子集的一种方法是根据人力编辑判断来指定某些网络页面。然而,人为估计是昂贵并且耗时的。在保留人工地选择种子集作为可行的备选方案的选择的同时,以下描述了半自动地构建种子集的另一种技术。

[0095] 种子选择处理取决于对种子页面应该具有两个重要特征的观察,即:1)从种子页面开始并且反复地跟随所遇到的网络页面上的外部链接,应该可以到达大量其他页面;即,种子页面应该提供高的覆盖范围,以及2)种子网页的质量应该非常高,因此,遇到从非垃圾到垃圾的链接的可能性应该是最小的。

[0096] 为了确保第一特征,产生了所有页面(即,在页面索引中的页面)的级别。为此,使用由方程 7 示出的以下线性系统。

[0097] $y = cU^T y + (1-c)v$, (方程 7)

[0098] 在该系统中

[0099] ● U 为逆转移矩阵,如果存在链接 $j \rightarrow i$,则其元素为

[0100] $U_{ij} = 1/\text{indeg}(i)$,否则,其元素为零。这里, $\text{indeg}(i)$ 是到页面 i 的内部链接的数量,用作归一化因子,以使矩阵 U 为随机的,

[0101] ● c 为远距离传送常数,一般在范围 0.7-0.9 内取值,

[0102] ● $y = (y_i)$ 为威信矢量,其中,下标 i 贯穿所有 n 个页面, $i = 1:n$,

[0103] ● $v = (v_i)$ 为远距离传送矢量,假设其为概率分布, $0 \leq v_i \leq 1, v_1 + \dots + v_n = 1$ 。

[0104] 值得注意的是,除了方程 7 使用逆转移矩阵 U 而不是正规转移矩阵 T 之外,由方程 7 所描述的系统类似于由方程 1 所描述的系统。逆转移矩阵对应于具有反向链接的方向性的网络图形。为此,具有统一远距离传送的方程 7 的解 y 被称为逆网页级别。逆网页级别

是通过跟随页面上的外部链接而从该页面可以到达的网络的数量的测度。

[0105] 为了确保种子页面的第二特征,人力编辑进一步处理具有最高逆网页级别的页面。人力编辑选择哪些候选者实际上是高质量非垃圾页面(提供如由逆网页级别测量的高覆盖范围的页面)。然后,如上所述,将人力编辑选择的页面包括进种子集中并用于信任级别计算。

[0106] 示例性种子集构造处理被概括如下:

[0107] A. 对于所有页面(主机等) i ,根据方程(7)得到其逆网页级别 y_i

[0108] B. 按照 y_i 的降序对页面 i 进行排序,并且保留排序列表的顶部,或者否则识别并保留一组最高级别的页面

[0109] C. 利用人力编辑来估计列表中所保留的页面的质量

[0110] D. 从列表中删除编辑认为不合适的页面

[0111] E. 所保留的页面构成种子集。

[0112] 实验结果表明,所得到的种子集适于信任级别计算和基于从网页级别和信任级别得出的质量估计的垃圾检测。

[0113] 本文中所描述的实施例可以涉及(reference)网站、链接、以及特定于由万维网(或其子集)充当搜索语料库的实例的其他术语。可以理解,本文中所描述的系统和方法可以适用于不同的搜索语料库(例如,电子数据库或文档储存库),并且结果可以包括内容以及对可以找到内容的位置的链接或引用。

[0114] 因此,虽然已经参考特定实施例描述了本发明,应该理解本发明旨在覆盖权利要求的范围内的所有修改及其等同物。

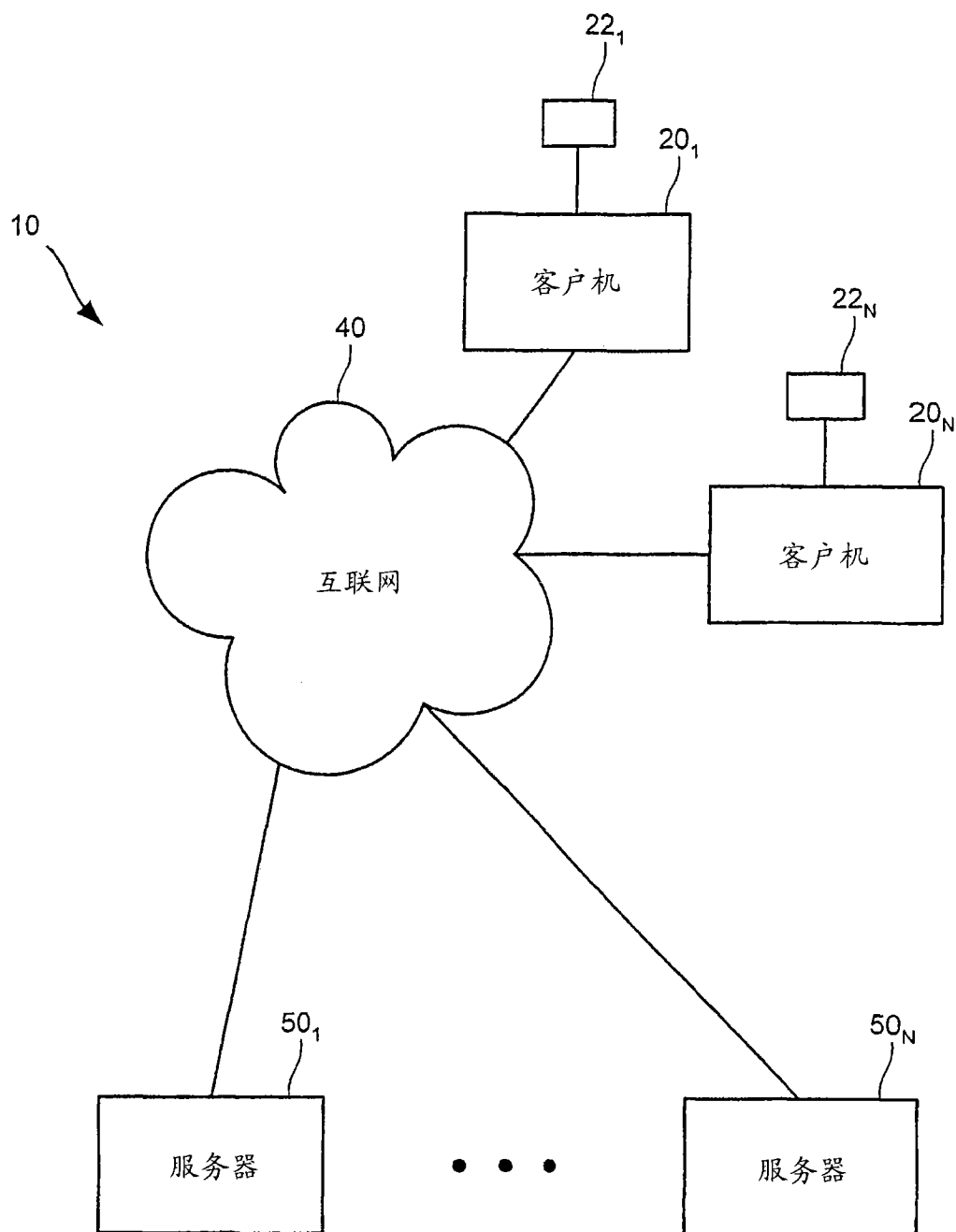


图 1

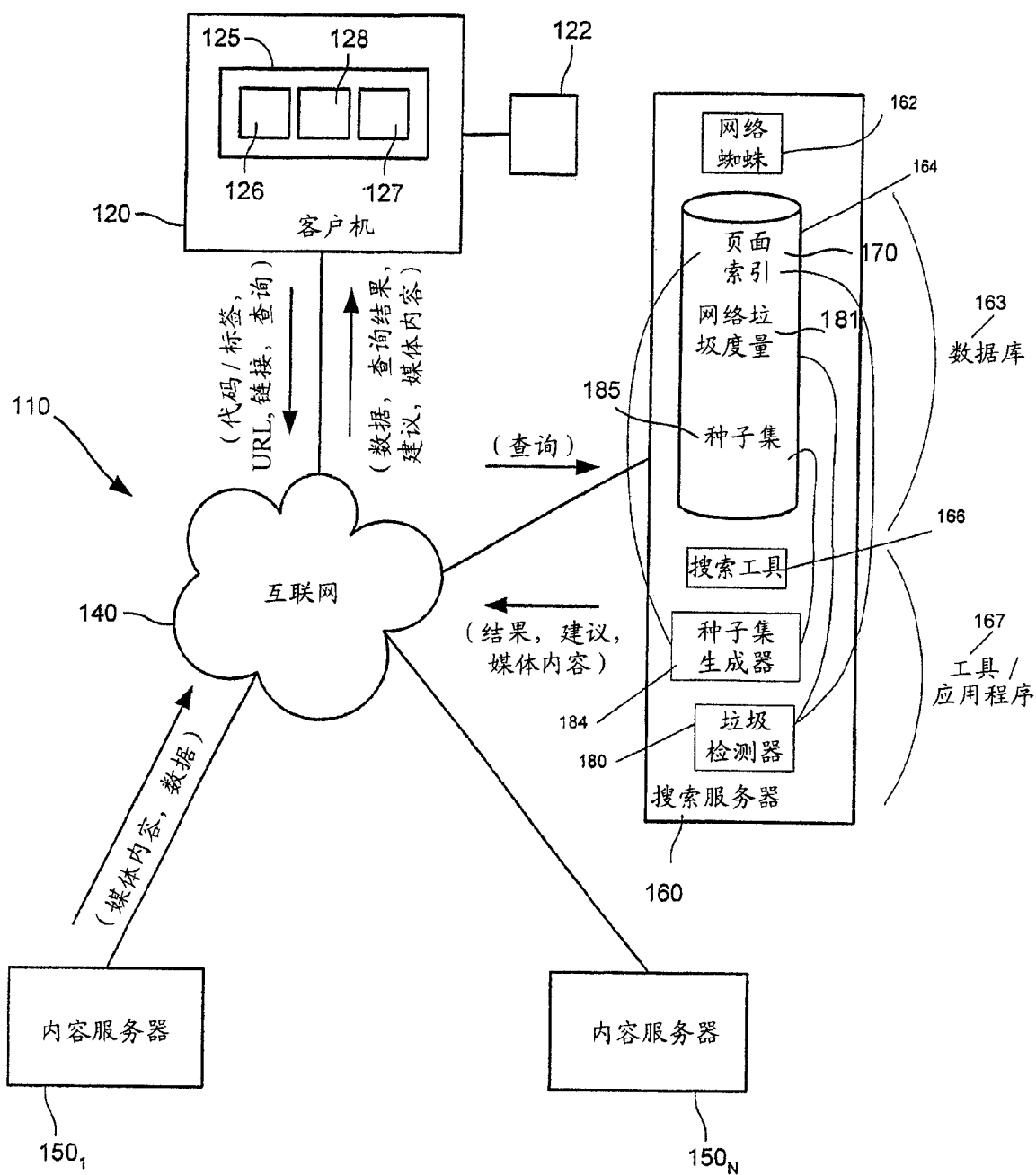


图 2

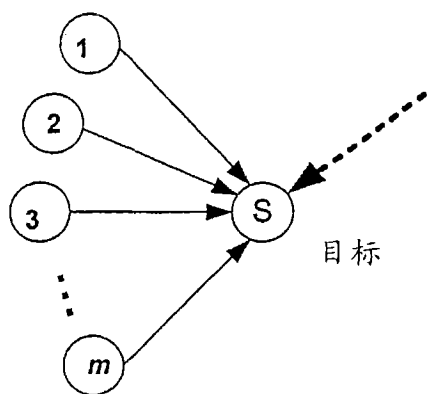


图 3A 简单垃圾场

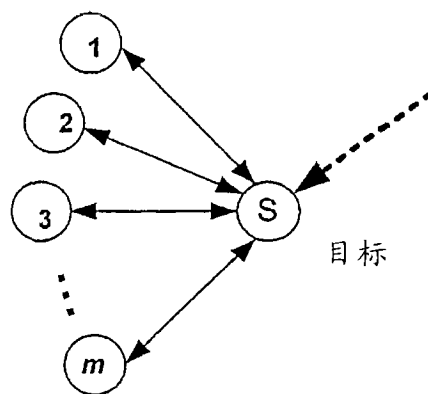


图 3B 具有反向链接的简单垃圾场