(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 101855631 B
(45)授权公告日 2016.06.29

(21)申请号 200780101491.7
(22)申请日 2007.11.08
(85)PCT国际申请进入国家阶段日
2010.05.07
(86)PCT国际申请的申请数据
PCT/CN2007/071033 2007.11.08
(87)PCT国际申请的公布数据
WO2009/059481 EN 2009.05.14
(73)专利权人 上海惠普有限公司
地址 100022 北京市建国路112号惠普大厦
专利权人 惠普发展公司，有限责任合伙企业
(72)发明人 张立 冯是聪 熊宇红
(74)专利代理机构 北京德琦知识产权代理有限公司 11018
代理人 归莹 罗正云
(51)Int.Cl.
G06F 17/30(2006.01)
审查员 张雯

权利要求书2页 说明书5页 附图4页

(54)发明名称

用于聚焦爬行的导航排名

(57)摘要

公开了用于聚焦爬行的导航排名的系统和方法。在一示例性实施例中，一种方法可包括使用分类器来将至少一个目标网页与网站上的其它网页区分开来。该方法还可包括通过有向图G＝(V，E)来对网站上的网页建模，其中每一个网页都由顶点(V)表示，而两个网页之间的链接由边(E)表示。该方法还可包括基于分类器来为V中的每一个网页(u)分配权重p(u)以计算指示网页的相关性的导航排名。

CN 101855631 B

1.一种用于聚焦爬行的导航排名的方法,包括:

使用分类器来将至少一个目标网页与网站上的其它网页区分开来;

通过有向图G=(V,E)来对所述网站上的网页建模,其中每一个网页都由顶点(V)表示,而两个网页之间的链接由边(E)表示;以及

基于分类器来为V中的每一个网页(u)分配权重p(u)以计算指示网页通向目标网页的可能性的度量的导航排名,

其中所述导航排名根据静态模型来计算,并且基于为生成所述网站的图而从所述网站下载的所有网页来根据所述静态模型计算所述导航排名;并且

其中计算所述导航排名使用平均方法。

2.如权利要求1所述的方法,其特征在于,较高的权重p(u)对应于所述网页的较高的相关性。

3.如权利要求1所述的方法,其特征在于,最高的权重p(u)对应于所述至少一个目标网页。

4.如权利要求1所述的方法,其特征在于,所述权重p(u)是二进制数或实数。

5.如权利要求1所述的方法,其特征在于,所述静态模型通过以下迭代过程来定义:

1.对于所有u,NR(u)(0)←1

2.t←0

3.对于所有u,NR(u)(t+1)=d*p(u)+(1-d)avg[NR(w)(t)/Ni(w)]

4.归一化NR(u)(t+1)以使其平均为1

5.如果对于所有u,|NR(u)(t+1)-NR(u)(t)|<e,则停止,

且令NR(u)=NR(u)(t+1)

6.否则,t←t+1,返回至步骤3

其中NR(u)是根据所述静态模型的顶点u的导航排名;p(u)是分配给u的权重;w表示u所指向的所有顶点;d是衰减因子;Ni(w)是指向w的链接数;而e是误差界限。

6.如权利要求5所述的方法,其特征在于,所述导航排名根据基于所述静态模型的活动模型来计算。

7.如权利要求6所述的方法,其特征在于,所述活动模型在使用所述静态模型来计算所述导航排名后通过以下迭代过程来定义:

1.对于所有u,NR'(u)(0)←1

2.t←0

3.对于所有u,NR'(u)(t+1)=d*NR(u)+(1-d)avg[NR'(v)(t)/No(v)]

4.归一化NR'(u)(t+1)以使其平均为1

5.如果对于所有u,|NR'(u)(t+1)-NR'(u)(t)|<e,则停止,

且令NR'(u)=NR'(u)(t+1)

6.否则,t←t+1,返回至步骤3

其中,NR'(u)是根据所述活动模型的顶点u的导航排名;v是指向u的所有顶点;而No(v)是指离v的链接数。

8.如权利要求6所述的方法,其特征在于,根据所述活动模型来计算所述导航排名基于为生成所述网站的图而从所述网站顺序地下载的网站子集。

9.如权利要求1所述的方法，其特征在于，计算所述导航排名使用从子孙到先辈网页的单向分数传播策略。

10.一种用于聚焦爬行的导航排名的系统，包括：

使用分类器来将至少一个目标网页与网站上的其它网页区分开来的装置；

通过有向图G＝(V,E)来对所述网站上的网页建模的装置，其中每一个网页都由顶点(V)表示，而两个网页之间的链接由边(E)表示；以及

基于分类器来为V中的每一个网页(u)分配权重p(u)以计算指示网页通向目标网页的可能性的度量的导航排名的装置，

其中所述导航排名根据静态模型来计算，并且基于为生成所述网站的图而从所述网站下载的所有网页来根据所述静态模型计算所述导航排名；并且

其中计算所述导航排名使用平均方法。

11.如权利要求10所述的系统，其特征在于，较高的权重p(u)对应于所述网页的较高的相关性。

12.如权利要求10所述的系统，其特征在于，最高的权重p(u)对应于所述至少一个目标网页。

13.如权利要求10所述的系统，其特征在于，所述权重p(u)是二进制数或实数。

14.如权利要求10所述的系统，其特征在于，所述导航排名根据基于所述静态模型的活动模型来计算。

15.如权利要求14所述的系统，其特征在于，根据所述活动模型来计算所述导航排名基于为生成所述网站的图而从所述网站顺序地下载的网站子集。

用于聚焦爬行的导航排名

[0001] 背景

[0002] 尽管因特网或万维网(www)上存在大量网站,但用户经常只对来自某些网站的特定网页上的信息感兴趣。例如,学生、专业人士和教育工作者可能想要容易地找到教育资料,如来自特定大学的在线课程。企业的市场营销部门可能想要知道顾客的教育背景、他们的产品和他们的竞争对手的产品之间的比较、以及其他相关产品信息。因此,各种搜索引擎对特定网站可用。

[0003] 一种发现域专用信息的方法是爬行网站的所有网页并使用分类工具来标识所需或"目标"网页。这一方法只有在具有大量计算资源或者网站只有少数网页的情况下才是可行的。一种发现域专用信息的更高效方式被称为聚焦爬行。实现高效聚焦爬行的一个挑战是确定页面可快速通向目标页面的可能性。

[0004] 两个公知示例是HITS算法以及PageRank算法的各种变型,诸如个性化PageRank(PPR)和动态个性化PageRank(DPPR)。这些算法根据主题相关性或个人兴趣来对页面进行排名。这些算法有可能可以在聚焦爬行中使用,即,通过根据HITS或DPPR计算出的分数来设置页面的爬行优先级。然而,这些算法各自具有缺陷。

[0005] 在PageRank算法中,如果链接到一网页的网页具有较高排名,则该网页接收较高排名。PPR是类似的,但还考虑页面相关性。PPR计算出的排名指示网页与特定主题的相关性,但该相关性不是对网页与目标页面的"连接性"的良好度量。例如,最终页面(不具有引出链接的网页)可具有非常高的排名,但它不通向任何其它页面。另外,PageRank及其变型计算合计分数。这对于聚焦爬行是不适当的。例如,考虑两个网页A和B,其中网页A链接到三个目标页面和三个非目标页面,而网页B链接到三个目标页面。如果根据PageRank模型来计算排名,则网页A将接收比网页B更高的排名。然而,从爬行的观点来看,网页B应排名更高,因为它比A"更纯"且通向目标页面。

[0006] 另外,PPR和DPPR是单向(从先辈到子孙)分数传播算法。因此,难以标识中心页面。然而,中心页面通常在聚焦爬行时非常有用,因为中心页面最有可能通向目标页面。

[0007] 另一方面,HITS算法是双向(先辈和子孙之间)分数传播算法,并且该算法可用于标识关于特定主题的中心页面和权威页面两者。直观上,中心页面是应在聚焦爬行时标识并探查的网页。然而,类似于PageRank算法,HITS算法的问题在于它计算合计分数。另外,在HITS算法中,目标页面用作形成包围这些目标页面的子结构的"种子",并且只计算该子结构中的节点的分数。在聚焦爬行时,应计算通常远离目标页面的每一个页面的分数。因此,HITS算法在这种情况下不起作用。

[0008] 附图简述

[0009] 图1是其中可实现用于聚焦爬行的导航排名的示例性联网计算机系统的高级图。

[0010] 图2是示例性网站的组织布局。

[0011] 图3是示出使用静态模型的用于对网站进行聚焦爬行的示例性导航排名的框图。

[0012] 图4是示出使用活动模型的用于对网站进行聚焦爬行的示例性导航排名的框图。

[0013] 详细描述

[0014] 公开了用于聚焦爬行的导航排名的系统和方法。导航排名的示例性实施例通过遵循通过更有可能通向目标页面的网页的链接来主动查找网站中的目标页面。网页通向目标页面的可能性基于网站的链接结构来度量。实现导航排名的聚焦爬行器可通过只探查可用网页的一小部分来发现大多数目标页面，因此减少爬行网站所需的时间和资源(并因此降低成本)。

[0015] 图1是其中可实现用于聚焦爬行的导航排名的示例性联网计算机系统100(例如，经由因特网)的高级图示。联网计算机系统100可包括诸如局域网(LAN)和/或广域网(WAN)等将一个或多个主机130(例如，服务器130a-c)处的一个或多个网站120连接到一个或多个用户140(例如，客户机计算机140a-c)的一个或多个通信网络110。

[0016] 此处所使用的术语“客户机”(例如，客户机计算机140a-c)指的是一个或多个用户140可用来访问网络110的一个或多个计算设备。客户机可包括各种各样的计算系统中的任一种，仅举几个例子，如独立个人台式或膝上型计算机(PC)，工作站、个人数字助理(PDA)、或电器。每一个客户机计算设备都可以包括存储器、存储、以及至少足以管理直接或间接到网络110的连接的一定程度的数据处理能力。客户机计算设备可经由诸如拨号、电缆、或经由因特网服务提供商(ISP)的DSL连接等通信连接来连接到网络110。

[0017] 此处所描述的聚焦爬行操作可由主机130(例如，主存网站120的服务器130a-c)或由联网计算机系统100中的第三方爬行器150(例如，服务器150a-c)来实现。在任一种情况下，服务器可执行允许对联网计算机系统100中的一个或多个网站120进行聚焦爬行的程序代码。然后可存储结果(例如，由爬行器150或网络中的别处)并按需访问这些结果以便在搜索网站120时协助用户140。

[0018] 如此处所使用的术语“服务器”(例如，服务器130a-c或服务器150a-c)指的是具有计算机可读存储的一个或多个计算系统。可经由诸如拨号、电缆、或经由因特网服务提供商(ISP)的DSL连接等通信连接来在网络110上提供服务器。服务器可经由网络110直接访问，或经由网络站点来访问。在一示例性实施例中，网站120还可包括第三方地点(例如，商用因特网站点)上的web门户，其便于经由后端链路或其他直接链路来连接到一个或多个服务器。服务器还可向其他计算或数据处理系统或设备提供服务。例如，服务器还可为用户140提供交易处理服务。

[0019] 当服务器“主存”网站120时，此处该服务器被称为主机130，而不管该服务器是来自服务器群集130a-c还是服务器群集150a-c。同样，当服务器正在执行用于聚焦爬行的程序代码时，此处该服务器被称为爬行器150，而不管该服务器是来自服务器群集130a-c还是服务器群集150a-c。

[0020] 程序代码可执行此处所描述的用于聚焦爬行的导航排名的示例性操作。在各示例性实施例中，这些操作可具体化为一个或多个计算机可读介质上的逻辑指令。当在处理器上执行时，这些逻辑指令使得通用计算设备被编程为实现所述操作的专用机器。在一示例性实现中，可使用各附图中所描绘的组件和连接。

[0021] 在进行聚焦爬行时，程序代码需要高效地标识目标网页。这通常难以做到，因为目标网页通常“远离”网站的主页。例如，大学课程的网页离大学的主页平均八个网页，如图2所示。

[0022] 图2是诸如图1所示的网站120等示例性网站的组织布局200。在该示例中，网站是

具有主页210的大学网站，该主页210具有到不同的子网页220a-c的多个链接215a-e。这些子网页中的至少某一些还可链接到子网页，诸如网页230并且然后链接到网页240-260，等等。通过网页260链接到目标网页270a-c。

[0023] 此处，可以看到，从大学主页210(“根”)到包含课程信息(例如，关于CS(计算机科学)1)的目标网页270a的最短路径是<主页><学院划分><工程和应用科学><计算机科学><学术><课程网站><CS1>。根据此处所描述的系统和方法，聚焦爬行器能够通过跟随该从根开始的最短路径并为每一个网页分配导航排名来发现目标页面270a。导航排名在以下更详细地描述，并且可用于确定每一页面有多大可能通向目标页面。

[0024] 在一示例性实施例中，可如下确定导航排名。假设分类器可用于将目标页面与其它网页区分开来，可由有向图G＝(V,E)来对网站上的网页建模。即，每一网页都由顶点V表示，而两个网页之间的链接由边E表示。由分类器来为V中的每一个网页或节点u分配权重p(u)，以指示该页面的相关性。权重越高，相关性就越高(即，网页是目标页面)。取决于分类器，权重可以是二进制或实数。

[0025] 给定这一具有顶点权重的图，对于V中的每一个顶点u，通过以下迭代过程来计算该顶点u的导航排名NR(u)：

[0026] 1.对于所有u,NR(u)(0)←1

[0027] 2.t ←0

[0028] 3.对于所有u,NR(u)(t+1)＝d＊p(u)+(1-d)avg[NR(w)(t)/Ni(w)]

[0029] 4.归一化NR(u)(t+1)以使其平均为1

[0030] 5.如果对于所有u,|NR(u)(t+1)-NR(u)(t)|＜e,则停止，且令NR(u)＝NR(u)(t+1)

[0031] 6.否则，t ←t+1，返回至步骤3

[0032] 在这些计算中，w表示u所指向的所有顶点；d是衰减因子(通常是一小常数，诸如0.2)；Ni(w)是指向w的链接数；而e是误差界限，此处出于说明的目的将其选为10-5。

[0033] 上文中，步骤1和2初始化该过程。在步骤3中，将导航排名作为初始相关性评级p和从在上一次迭代中计算的邻居的导航排名导出的值的线性组合来计算。在步骤4和5中，确定是否已满足收敛条件。

[0034] 直观上，每一个节点由于指向具有高分的节点而受到奖励并且由于指向具有低分的节点而受到惩罚，其中分数是递归地定义的。对于任何权重p，上述迭代过程通常收敛，并且该收敛通常是迅速的。

[0035] 上述过程可以参考以下说明来更好地理解。图3是可用于示出使用静态模型的示例性导航排名计算的网站图(G)300，其中下载来自该网站的所有网页以生成图300。在图300中，节点A是根而节点D和F是目标网页。

[0036] 表1示出了p的值、上述过程中的前两次迭代后的NR的中间值、以及最终NR计算(d＝0.2)。在表1中，对于t的每一个值存在两行，其中上行示出步骤3之后的NR值，而下行示出步骤4之后的NR的归一化值。

[0037] 表1：示例性NR计算

[0038]

	A	B	C	D	E	F
P	0	0	0	1	0	1
t=1	0.67	0.40	0.53	0.20	0.00	0.20
	2.00	1.20	1.60	0.60	0.00	0.60
t=2	0.83	0.24	0.16	0.20	0.00	0.20
	3.05	0.89	0.59	0.74	0.00	0.74
t=∞	0.64	0.31	0.21	0.20	0.00	0.20
	2.46	1.20	0.80	0.77	0.00	0.77

[0039] 如可以从表1中的示例性计算中看到的，虽然节点D和F是仅有的两个相关页面，但它们的NR值相对较低。的确，NR度量页面有多大可能通向目标页面，而不是度量页面有多大可能是目标页面。

[0040] 在该示例中，首先必须下载网站中的所有页面以获取图3所示的图300。被称为“静态”模型的该模型在其中在许多类似站点上实现域专用爬行的情况下是合适的。例如，爬行器可实现该静态模型以爬行来自所有大学网站的课程页面。可下载若干整个网站并且可使用这些网站通过上述过程来计算导航排名。然后，可调用机器学习过程以发现导航排名和诸如网页URL名称、锚文本或内容等网页特征之间的关系。然后，对于新网站，使用所习得的结果来逼近在爬行过程中遇到的每一个页面的导航排名。在爬行期间展开具有较高排名的网页。

[0041] 在其中不期望为了获取诸如图3所示的图300等图G而下载网站中的所有页面的其它情况下，可根据“活动”模型来计算导航排名。在活动模型中，通过在爬行每一个单独站点时动态调整节点的导航排名来确定该站点的结构。导航排名更准确地反映网站的结构。

[0042] 然而，在直接在活动模型中应用静态模型中的NR定义方面存在问题。即，对于尚未下载的每一个URL，不存在“引出”链接，因为尚未下载和解析对应的网页。这些页面的NR值将根据NR的静态定义而始终为0，这对于辨别哪一个页面更有可能通向目标页面是没用的。

[0043] 为了解决该问题，可实现附加计算以便在爬行期间传播NR分数。如下计算对于活动模型的导航排名(指定为NR')：

[0044] 1.对于所有u，NR'(u)(0)←1

[0045] 2.t←0

[0046] 3.对于所有u，NR'(u)(t+1)＝d＊NR(u)+(1-d)avg[NR'(v)(t)/No(v)]

[0047] 4.归一化NR'(u)(t+1)以使其平均为1

[0048] 5.如果对于所有u，|NR'(u)(t+1)-NR'(u)(t)|＜e，则停止，且令NR'(u)＝NR'(u)(t+1)

[0049] 6.否则，t ←t+1，返回至步骤3

[0050] 在这些计算中，NR'(u)是由第一个算法计算的顶点u的导航排名；v是指向u的所有顶点；而No(v)是指离v的链接数。

[0051] 迭代非常类似于上述过程。区别在于颠倒分数传播的方向。先前，从外邻居(指离u的邻居)中取平均值；在上述过程中，从内邻居(指向u的邻居)中取平均值。

[0052] 可在聚焦爬行时实现活动模型以便实时(即，在下载网站的子集时)计算导航排名。图4是可用于示出使用活动模型的示例性导航排名计算的网站图(G')400，其中从该网站顺序地下载网页的子集(例如，子集410和420)以生成图400。同样，节点A'是子集的根节

点而节点D'和F'是目标网页。注意，节点A'可以是主页，但不必是该网站的主页。因此，即使在整个网页不可用时也可更快且更高效地实现导航排名。

[0053] 首先使用标准广度优先搜索方法来下载第一页面子集410。然后调用分类器并计算该图400的子集中的每一个节点的导航排名。爬行器然后通过跟随第一子集410中的具有较高导航排名的网页上的链接来下载更多网页(例如，第二子集420)，直到新页面的数量达到阈值。基于设计考虑(例如，处理能力、所需完成时间等)，该阈值可以是网页的任何合适数量。然后在展开图上重新计算每一个网页的导航排名并重复爬行过程(下载更多子集并计算NR)，直到定位到足够的目标页面。

[0054] 在一替换实施例中，类似于静态模型，可确定导航排名和网页特征之间的关系，并且将结果用于引导爬行。在第二步传播后，将在第一步中计算的NR分数分发给跟随链接的节点。如果尚未下载这些页面，则可为这些页面分配较高的NR分数并在下一爬行周期中首先爬行。

[0055] 导航排名的示例性实施例还可在迭代计算中实现平均分数而不是总和，以确定页面排名。使用先前的现有技术图示，在网页u具有作为目标网页的两个子网页以及作为噪声的三个子网页的情况下，u排在5个单位；而在网页v具有三个子网页且所有三个子网页都是目标网页的情况下，v排在3个单位，这两种情况都使用合计方法。因此，网页u被误选为目标网页。然而，根据此处的教导，使用平均方法来将网页u排在2/5(即，总共五个子网页中有两个目标网页)；而网页v将排在3/3(或1，即，总共三个子网页中有三个目标网页)，因此比网页u更高。因此使用平均方法的导航排名在聚焦爬行期间提供更准确的结果。

[0056] 同样在示例性实施例中，导航排名可实现从子孙到先辈的单向分数传播策略。如果一网页指向具有高分的页面，则该网页排名更高。因此，可有效地标识中心页面。或者，导航排名可实现双向且两步分数传播策略。同样，网页在其指向具有高分的页面的情况下排名更高以使得可有效地标识中心页面。接着，将在第一个步骤中获取的分数从先辈分发到子孙。因此，将很可能爬行可能的目标页面，因为其是高分网页所指向的。此外，该两步分数传播比HITS中使用一步分数传播更有效。

[0057] 可以理解，此处所示出和描述的实施例只是出于说明示例性系统和方法的目的并且不旨在是限制性的。另外，此处所示出和描述的操作和示例是为了说明用于聚焦爬行的导航排名的示例性实现而提供的。注意，操作不限于所示操作。也可实现其他操作。还构想用于聚焦爬行的导航排名的还有一些其他实施例，如本领域内的普通技术人员将在熟悉此处的教导后容易地理解的那样。

[0058] 除了此处明确阐述的具体实施例之外，考虑此处所公开的说明书，其它方面和实现将对本领域的技术人员是显而易见的。

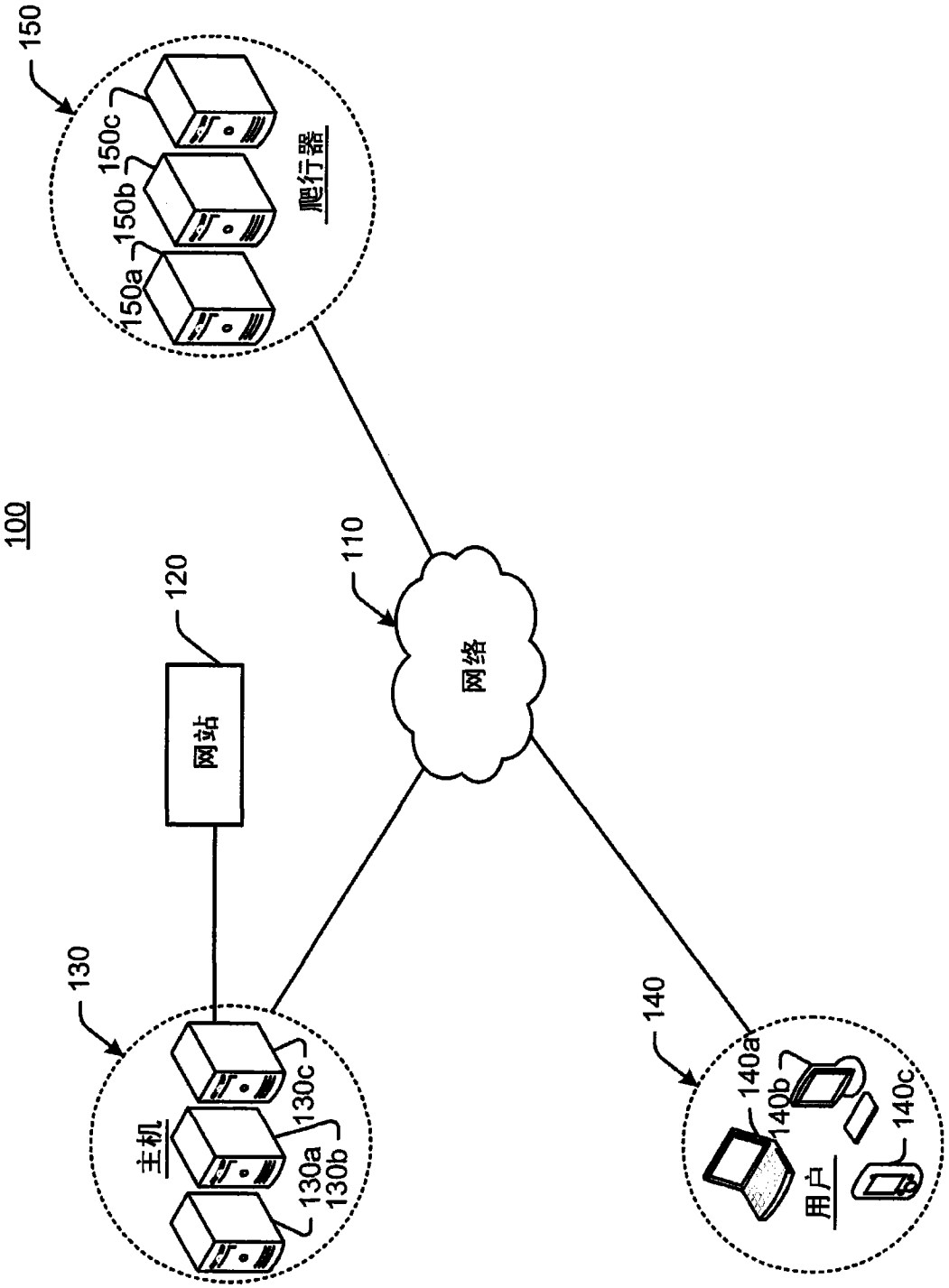
150
150c
150b
150a
爬行器
100
110
网络
120
网站
130
主机
130c
130a
130b
140
140a
140b
140c
用户

图1

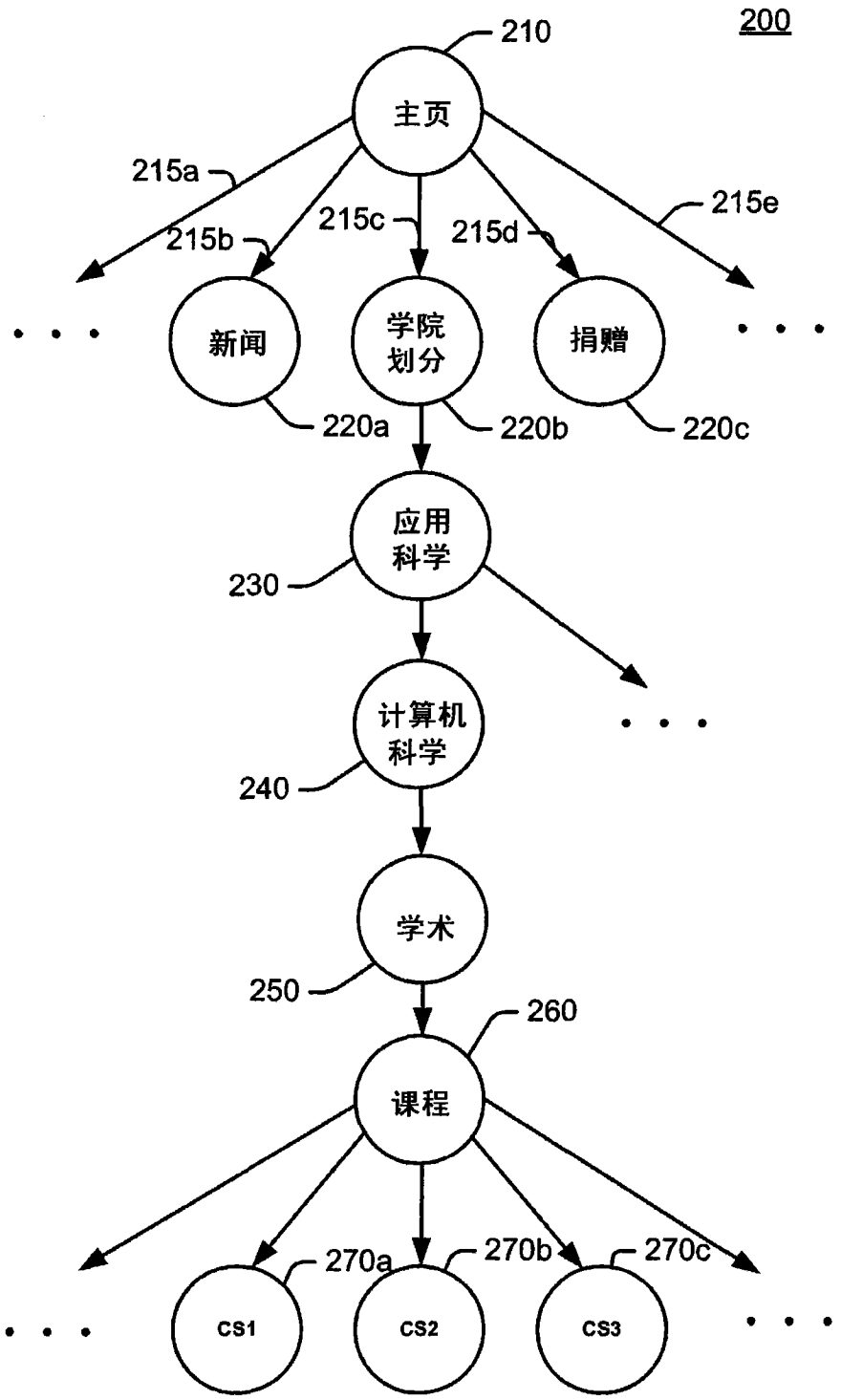
200
210
主页
215a
215b
215c
215d
215e
新闻
学院划分
捐赠
220a
220b
220c
应用科学
230
计算机科学
240
学术
250
260
课程
270a
270b
270c
CS1
CS2
CS3

图2

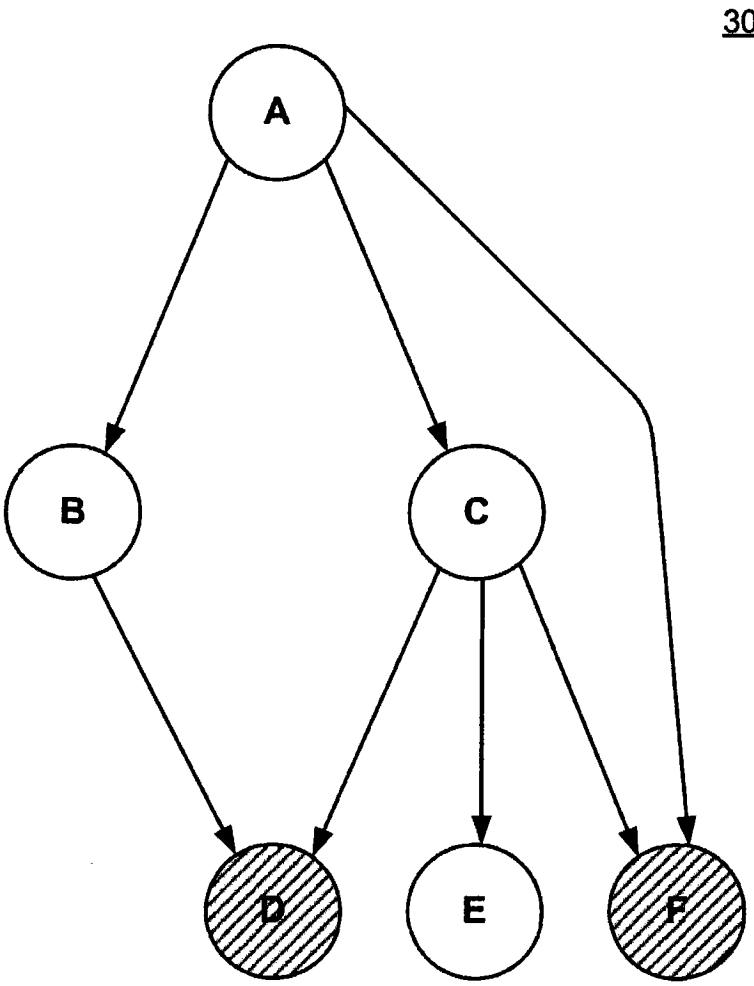

图3

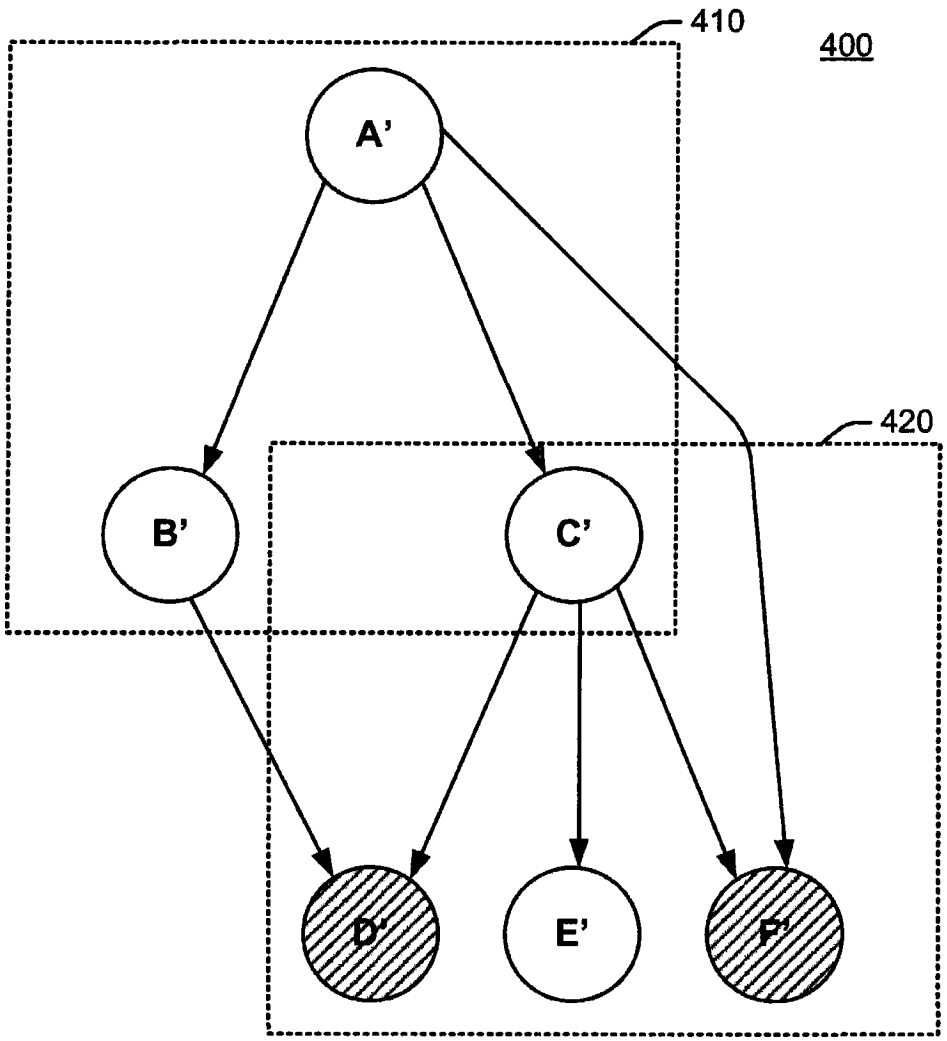

图4