(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 101268465 B
(45)授权公告日 2012.05.02

(21)申请号 200680034703.X
(22)申请日 2006.09.07
(30)优先权数据
0552826 2005.09.20 FR
(85)PCT申请进入国家阶段日
2008.03.20
(86)PCT申请的申请数据
PCT/FR2006/050856 2006.09.07
(87)PCT申请的公布数据
WO2007/034096 FR 2007.03.29
(73)专利权人 法国电信公司
地址 法国巴黎
(72)发明人 杰罗姆·高尔蒂尔
(74)专利代理机构 北京市柳沈律师事务所 11105
代理人 刘雅红
(51)Int.Cl.
G06F 17/30(2006.01)
(56)对比文件
GB 2338324 A,1999.12.15, 说明书的10页第30行至第11页第5行、第12页第20-30行、第13页26-32行.
US 5987470 A,1999.11.16, 全文.
US 20020156760 A1,2002.08.24, 全文.
CN 1306650 A,2001.08.01, 全文.
US 5808615 A,1998.09.15, 全文.
Vicenc Torra等.Hierarchical spherical clustering.International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems10 2.2002,10(2),157-172.
审查员 徐淑娴

权利要求书 2 页 说明书 8 页 附图 1 页

(54)发明名称

分类一组电子文档的方法

(57)摘要

本发明关于用于分类一组电子文档的方法，包括以下存在的步骤：针对该组的每对文档{u，v}确定(S110)在文档u和v之间的相关度ω(u，v)；确定在所述组文档和组R^d的球面之间的投影的函数，其中d是正整数，函数X使得对于至少一个文档u，在R^d中两点X(u)和X(v)之间的距离越小，则相关度越高，其中v是在文档u和v之间存在相关度的文档；基于函数X所取的值，关于所述组文档的至少一个部分执行分类操作(S140)。

CN 101268465 B

1. 一种分类一组电子文档的方法，包括步骤：

- 计数在所述组的每对文档 {u，v} 之间出现的超文本链接或联合引用，

- 针对所述组的每对文档 {u，v} 确定文档 u 和 v 之间的相关度 ω (u，v) (S110)，所述相关度取决于在计数步骤完成时所获得的链接数，

- 在 d 维空间 R^d 的球面上产生图解表示法并且通过定位于所述球面上的相关联点 X(u) 表示所述电子文档组的每一个文档 u，其中 R 是一组实数而 d 是正整数 (S120、S125、S130、S135)，

其中，对于所述组的至少一个文档 u1，在所述 d 维空间中相关联点 X(u1) 和 X(u2) 之间的距离越小，则文档 u1 和 u2 之间的相关度就越高，其中 u2 是在文档 u1 和 u2 之间存在相关的文档，

- 将所述组文档的至少部分基于在所述球面上确定的点而分类 (S140)。

2. 如权利要求 1 所述的方法，其中至少一个所述文档包括至少一个到至少一个其他文档的超文本链接，在两个文档 u 和 v 之间的相关度基于出现在文档 u 和 v 之间的超文本链接数、和 / 或联合引用链接数而确定，相关度越高，则该链接数越大，不存在相关与不存在链接相对应。

3. 如权利要求 1 或 2 所述的方法，其中两个文档 u 和 v 之间的相关度取决于文档 u 和 v 的语义内容的接近的测量结果，相关度越高，则所述测量结果越低，不存在相关与测量结果低于预定义的阈值相对应。

4. 如权利要求 1 或 2 所述的方法，包括步骤：

- 定义 (S120) 用于将所述组投影到所述球面的初始函数 X_0，

- 确定用于将所述组投影到所述球面的函数 X，所述投影函数 X 以至少一次迭代、基于初始函数获得，每一次迭代在于基于在上一迭代中获得的函数 X_{i-1}，通过对于所述组的至少一个文档 u，以值 $X_i(u)$ 替换值 $X_{i-1}(u)$ 来确定函数 X_i，从而使得对于属于所述组的任意文档 v，可以优化取决于值 $X_{i-1}(u)$ 和值 $X_{i-1}(v)$ 以及在文档 u 和 v 之间的相关度 ω (u，v) 的预定义准则 (S125、S130、S135)。

5. 如权利要求 4 所述的方法，其中所述预定义的准则的优化在于针对文档 u 将量 Δ (u) 最大化为等于，

$$\Delta(u) = \sum_{\{u,v\}\in E} \delta(u,v)\left\| X_{i-1}(u) - X_{i-1}(v)\right\|^2$$

其中 δ (u，v) ＝ 1－ω (u，v)，0 ≤ ω (u，v) ≤ 1，在文档 u 和 v 之间不存在相关时 ω (u，v) ＝ 0，

其中 $Y(U) = \sum_{v\in V-\{u\}} \delta(u,v) X_{i-1}(v)$：

若 Y(U) ≠ 0，则 $X_i(u) = -Y(U)/||Y(U)||$，

若 Y(U) ＝ 0，则值 $X_i(u)$ 等于值 $X_{i-1}(u)$。

6. 如权利要求 1 所述的方法，进一步包括在于产生所述球面的图解表示法以及位于所述球面上的点 X(u) 的图解表示法的步骤。

7. 如权利要求 6 所述的方法，进一步包括步骤存在于：

- 在终端上显示所述图解表示法，

- 向终端用户提供合适的图形选择工具，用于至少部分所述图解表示法的图形选择，

- 获取限定通过所述用户选择的至少一个部分的数据，

- 基于所述数据对于所述组的文档执行分类。

8. 一种数据处理装置，包括：

- 计数在所述组的每对文档 {u，v} 之间出现的超文本链接或联合引用的数据处理部件，

- 针对所述组的每对文档 {u，v} 确定文档 u 和 v 之间的相关度 ω(u，v)(S110) 的数据处理部件，所述相关度取决于在计数步骤完成时所获得的链接数，

- 针对所述组的每个文档 u 确定位于组 R^d 的球面上的相关联点 X(u)(S120、S125、S130、S135) 的数据处理部件，其中 R 是一组实数而 d 是正整数，对于所述组的至少一个文档 u1，在 R^d 中相关联点 X(u1) 和 X(u2) 之间的距离越小，则文档 u1 和 u2 之间的相关度就越高，其中 u2 是在文档 u1 和 u2 之间存在相关的文档，

- 将所述组文档的至少部分基于在所述球面上确定的点而分类 (S140) 的数据处理部件。

分类一组电子文档的方法

技术领域

[0001] 本发明属于电信领域，并且特别地属于用于搜索电子文档的搜索引擎领域。

[0002] 更精确地，本发明涉及分类一组电子文档的方法。例如，由用户通过因特网类型的网络上的搜索引擎执行搜索而产生这样的一组，在这种情况中的电子文档是 Web 页面（“万维网”的简称），其通过本地存储介质而本地访问，或经由网络而被远程访问。

背景技术

[0003] 搜索引擎利用用于评价 (rate) 或分类从搜索中出现的页面的几种技术。在用于探查 (explore) 一组 Web 页面的公知技术中，一些技术依靠语义学，如果一个页面包含所搜索的词的大量出现 (occurrence)，则页面被评价为更加相关。这些技术敏感于名字“兜售信息 (spamming)”公知的作法，其旨在使因特网用户在他们的搜索查询中通常使用的词在给定页面中展示大量次数，这具有使所述页面当作相关的而频繁出现的效果。

[0004] 其他技术基于 Web 的拓扑结构。这些技术同时考虑在所考虑的页面之间的现存链接以及这些页面本身的性质，如在 Web 的网络域或子域中页面的从属关系 (membership)。这些技术通常基于要处理的页面的图型表示法。它们适于满足在所述图中给出的拓扑性质的页面的分类。这些技术敏感于“兜售信息”的方法的变体，其旨在多次引用给定页面，这具有本地伪造 Web 的图的拓扑特性的效果。

[0005] 一些利用 Web 的拓扑结构的技术存在于通过向各种页面分配依赖于页面和其他页面之间的关系的排名 (rank) 来实现 Web 页面的分类。

[0006] 这样的过程的示例，公知为术语“PageRank”，用在 Google™ 搜索引擎的实现中并在文档中被描述：“The PageRank Citation Ranking:Bringing Order onthe Web”，by L.Page，S.Brin，R.Motwani and T.Winograd;Technical Report，Computer Science Department，Stanford University，1998。

[0007] PageRank 过程基于 Web 上页面可见性来排序页面。在该过程中，模拟通过跟随超文本链接在 Web 上随机的逐页浏览。该浏览对应于当后者随机地激活位于所查看页面中的超文本链接之一时通过用户访问 Web 而造成的浏览，以便访问另一页面。该过程承担该模拟的浏览的概率分析，以便确定在诸如此的随机逐页浏览期间位于给定页面的用户的概率。一个页面被其他页面引用的次数越多，页面的排名就越高。

[0008] 这种过程提供了对于由用户执行的搜索非必要相关的评价排名。最好评价的页面（最高排名）不必然是最好地与用户的期望对应的页面。

[0009] 此外，该过程不可能在一组文档中识别主题一致 (community) 或兴趣一致，不能够将用户更加迅速地指引到感兴趣的页面。

[0010] 最终，在用户在一组呈现的文档中识别他特别感兴趣的文档的情况下，不可能通过使用仅基于其排名而排序的一列文档来容易地确定与感兴趣的文档接近或无论如何与后者链接的其他文档是否出现在该组文档中。

发明内容

[0011] 因此本发明的目标特别是通过提出用于分类电子文档（如网页）的技术来解决现有技术的上述缺点，特别使检测兜售信息的问题成为可能，其适用于一大组文档且易迅速实现，并且不是通过分配排名，而是不管基于文档的语义内容、还是在这些文档之间的超文本链接或完全为一些其他方式定义所述接近的概念，完全通过构造彼此相近的文档或文档的子组的一致使简单分类文档成为可能。

[0012] 由于这个目的，本发明的主题是，根据第一方面，分类一组电子文档的方法，包括步骤：

[0013] -计数出现在所述组的每对文档{u，v}之间的超文本链接或联合引用(cocitation)，

[0014] -对于所述组的每对文档{u，v}，确定在文档u和v之间的相关度ω(u，v)，所述相关度取决于在完成计数步骤时获得的链接数，

[0015] -对于所述组的每个文档u，确定位于组R^d的球面的关联点X(u)，其中R是一组实数而d是一个正整数，对于所述组的至少一个文档u1，在R^d中关联点X(u1)和X(u2)之间的距离越小，文档u1和u2之间的相关度就越高，其中u2是在文档u1和u2之间存在相关的文档，

[0016] -将所述组文档的至少部分基于在所述球面上确定的点而分类。

[0017] 使用球面来确定点的位置的事实是独创的，在于这使得对于所获得的点-因此对于相关联的文档-可以简单地定义这些点关于彼此的相对位置。具体地，在这种表示法模式中，没有点关于另一点是有利的(favor)。因此，两点之间的相对位置，并且因此这两点之间的距离，可以用于表示与这两点相关联的两个文档之间的相关度。由此获得的表示法反映了涉及的文档之间的相关或链接。

[0018] 采用球面上（例如在三维空间中的球面）的一组文档的表示法的事实，进一步使得设想任意类型的分类操作成为可能：通过选择、评价(rate)、过滤、分级，并且简单地实行这些，这是由于每个文档从此以后在N维空间中由一简单的n重数(tuple)坐标来表示（例如通过三维空间中的三个坐标来表示）。

[0019] 存在所述方法的多种应用：文档的簇的构建、文档的评价或选择。在空间R^d中基于文档的投影的空间位置或基于距离测量结果来执行这些操作，也就是说通过考虑它们的相关度或接近度来确定。

[0020] 根据本发明的方法可以例如用于执行从搜索引擎执行的搜索中产生的WEB页面的任意种类的分类操作、评价、分级，最初的页面，也就是说距离其他页面最远的页面，被评价为例如最高。

[0021] 作为选择或组合，所述页面通过群(group)来分类，每一个群与一组页面相对应，所述组页面通过函数X的投影位于空间R^d的球面的预定义空间区域中。优选地，在这个变体中，限定该球面到空间区域的分割，并且根据它们在分割的空间区域之一中的投影的从属关系而评价所述文档。

[0022] 根据本发明的方法还可以用于检测“兜售信息”的出现，也就是彼此指向的页面，这是因为在球面S上的所有这些页面的投影将基本上彼此靠近。

[0023] 根据本发明的方法还可以用于产生从由搜索引擎执行的搜索中产生的WEB页面

的可视表示 (visual representation)。

[0024] 根据所述方法的第一变体 (variant),其中至少文档之一呈现出到至少另一文档的至少一个超文本链接,两个文档 u 和 v 之间的相关度基于出现在文档 u 和 v 之间的超文本链接数和 / 或联合引用链接数而确定,相关度越高,则该数目就越大,不存在相关对应于不存在链接。

[0025] 该第一变体有利地允许了考虑文档之间的超文本链接或联合引用链接的 分类操作的实现。

[0026] 根据所述方法的第二变体,两个文档 u 和 v 之间的相关度基于文档 u 和 v 的语义内容的接近的测量结果而确定,相关度越高,该测量结果就越低,不存在相关对应于测量结果低于预定义的阈值。

[0027] 该第二变体有利地允许了考虑文档的语义内容的分类操作的实现。

[0028] 根据第三变体,相关度基于由多个用户定义的喜爱页面而确定。在这种情况下,由于每个用户均与一组文档相关联 (他的喜爱页面),所以两个文档 u 和 v 之间的相关度作为文档 u 和 v 所属于的这样的组的数目而确定。

[0029] 该第三变体有利地使得在确定页面之间的相关度时考虑用户简档 (profile) 成为可能。

[0030] 所述三种变体可以进一步联合在一起,以便确定同时考虑超文本链接、语义内容和 / 或更适宜用户的相关度。两个文档之间的任意其他类型的链接也用于定义相关度。

[0031] 根据特定实施例,所述方法进一步包括步骤 :

[0032] - 定义用于将所述组投影到所述球面的初始函数 X_0,

[0033] - 确定用于将所述组投影到所述球面的初始函数 X,所述投影函数 X 以至少一次迭代、基于初始函数 X_0 而获得,每一次迭代在于基于在上一迭代中获得的函数 X_{i-1} 来对于所述组的至少一个文档 u,通过以值 $X_i(u)$ 值替换 $X_{i-1}(u)$ 来确定函数 X_i,从而使得对于属于所述组的任意文档 v,优化取决于值 $X_{i-1}(u)$ 和值 $X_{i-1}(v)$ 以及文档 u 和 v 之间的相关度 $\omega(u,v)$ 的预定义准则成为可能。

[0034] 根据本发明的方法有助于函数 X 的迭代确定,由此简化了其实现并且使得精确控制所述方法的收敛成为可能。

[0035] 优选地,以随机方式定义函数 X_0。统计地从随机函数开始的事实改进了收敛到所期望的函数 X 的速度,实行这些无需关于要获得的函数的先验知识。

[0036] 在该实施例中,预定义准则的优化存在于对于文档 u 将量 $\Delta(u)$ 的值最大化等于 :

[0037] $$\Delta(u)=\sum_{\{u,v\}\in E}\delta(u,v)\left\|X_{i-1}(u)-X_{i-1}(v)\right\|^2$$

[0038] 其中 $\delta(u,v)=1-\omega(u,v)$,$0\leqslant\omega(u,v)\leqslant 1$,在文档 u 和 v 之间不存在相关时 $\omega(u,v)=0$,

[0039] 值 $X_i(u)$ 等于

[0040] $X_i(u)=-Y(U)/||Y(U)||$ 其中

[0041] 若 $Y(U)\neq 0$,则 $$Y(U)=\sum_{v\in V-\{u\}}\delta(u,v)X_{i-1}(v)$$

[0042] 若 $Y(U)=0$,则值 $X_i(u)$ 等于 $X_{i-1}(u)$。

[0043] 本发明的主题也是由计算机系统可读的信息介质上的计算机程序，所述程序包括当该程序被载入然后由计算机系统执行时，用于实现诸如以上简要定义的根据本发明的方法的指令。

[0044] 本发明的主题也是数据处理装置，包括用于根据本发明的方法的步骤的执行的数据处理部件。这样的装置设备例如是实现文档搜索引擎的计算机服务器。

[0045] 本发明的主题也是由计算机系统可读的记录介质，包括程序，所述程序包括当所述程序由计算机系统执行时，用于实现根据本发明的方法的程序代码指令。

[0046] 附图说明

[0047] 本发明的其他目的、特性和优点将通过下面的完全以非限制示例给出的、并参照附图提供的描述而变得更加明显，其中

[0048] 图 1 是根据本发明的方法的实施例的流程图。

[0049] 具体实施方式

[0050] 根据本发明的方法应用于一组电子文档，特别是一组 WEB 页面，对于它们中的一些，包括到一个或多个其他页面的一个或多个超文本链接。

[0051] 在图解的所选择的实施例中，一组文档 V 的两个文档 u 和 v 之间的相关度基于存在于文档 u 和 v 之间的超文本链接和联合引用链接的数目而确定。

[0052] 对于两个文档之间超文本链接数的确定，不考虑超文本链接的意义而考虑“对称的”超文本链接，也就是说将相同的处理应用到文档 u 包括到文档 v 的链接的情况和文档 v 包括到文档 u 的链接的情况。

[0053] 如果存在至少一个其他文档 w 使得：

[0054] - 存在至少一个从 w 指向 u 的超文本链接，以及

[0055] - 存在至少一个从 w 指向 v 的超文本链接，

[0056] 则两个文档 u 和 v 具有联合引用链接。

[0057] 现在通过参照图 1 更详细地描述根据本发明的方法的步骤。

[0058] 步骤 S100 存在于对于所述组 V 的文档的任意对 {u，v} 确定取决于文档 u 和 v 之间的超文本链接数的权重 $\omega_1(u,v)$。优选地，函数 $\omega_1(u,v)$ 是文档 u 和 v 之间的超文本链接数的增函数。

[0059] 优选地，$\omega_1(u,v)$ 的值位于预定义的最小值（一般是 0）和预定义的最大值（一般是 1）之间。在这种情况下，最小值与在文档 u 和 v 之间不存在超文本链接相对应，而最大值例如与在文档 u 和 v 之间出现预定义的最小数目的超文本链接相对应。

[0060] 根据第一示例，在不存在超文本链接时 $\omega_1(u,v)$ 的值选为等于 0，而在文档 u 和 v 之间至少出现一个超文本链接时 $\omega_1(u,v)$ 的值选为等于 1。

[0061] 根据第二示例，在不存在超文本链接时 $\omega_1(u,v)$ 的值选为等于 0，在文档 u 和 v 之间出现单个超文本链接时 $\omega_1(u,v)$ 的值选为等于 0.5，而在文档 u 和 v 之间出现两个或多个超文本链接时 $\omega_1(u,v)$ 的值选为等于 1。

[0062] 根据第三示例，将 $\omega_1(u,v)$ 的值定义为在文档 u 和 v 之间的超文本链接数 N_h 的连续增函数，例如：

[0063] $$\omega_1(u,v)=\sqrt{\frac{N_h}{N_{h\max}}}$$

[0064] 其中 N_{hmax} 是超文本链接数 N_h 的最高阈值。

[0065] 步骤 S105 存在于对于所述组 V 的文档的任意对 {u，v} 确定取决于文档 u 和 v 之间的联合引用链接数的权重 $\omega_2(u,v)$。优选地，函数 $\omega_2(u,v)$ 是文档 u 和 v 之间的联合引用链接数的增函数。

[0066] 针对 $\omega_1(u,v)$ 给出的函数定义示例可换位于 $\omega_2(u,v)$。例如，在不存在联合引用链接时 $\omega_2(u,v)$ 的值选为等于0，而在文档u和v之间至少存在一个联合引用链接时 $\omega_2(u,v)$ 的值选为等于 1。

[0067] 步骤 S110 存在于对于文档的任意对 (u，v) 确定与对 {u，v} 通过以下关系相关联的相关度 $\omega(u,v)$。

[0068] $\omega(u,v) = k_1\omega_1(u,v)+k_2\omega_2(u,v)$，

[0069] k1 和 k2 是实系数使得，

[0070] $0 \leqslant k_1 \leqslant 1$，

[0071] $0 \leqslant k_2 \leqslant 1$，

[0072] $k_1+k_2 = 1$。

[0073] 由此相关度 $\omega(u,v)$ 取得位于 0 和 1 之间的实数值，值 0 与不存在链接相对应。

[0074] 由于希望强调超文本链接的存在，所以赋予系数 k1 的值将被选择得更高。相反，由于希望强调联合引用链接的存在，所以赋予系数 k2 的值将被选择得更高。

[0075] 用于确定文档之间的相关度的该过程使得在根据本发明的、用于分类文档的方法中，在文档之间考虑两种类型的链接：超文本链接和联合引用链接成为可能。

[0076] 该过程对其他类型的链接是普遍的。例如，如果存在一个或多个超文本链接使得从 u 传递到 v 成为可能，则可以定义两个文档 u 和 v 通过间接超文本链接而互连，在这种情况下的超文本链接数大于或等于 2。

[0077] 根据另一示例，可以考虑文档之间的语义类型的链接。在这种情况下，基于两个文档的语义内容的分析和比较来执行两个文档之间的相关度的确定。就这个目的而言，用于比较语义内容的公知过程是适用的。然后相关度表示两个文档之间语义接近的测量。例如可以基于在每个文档中所包括的词语的统计分析和比较来确定语义相关度。作为变体，可以定义两个文档之间的距离以及将相关度作为所定义的距离的减函数而定义，以这种方式使得两个文档之间的距离越小，则这些文档之间的相关度越高。

[0078] 将该过程最终概括为链接的任意值，而不管它们的类型。然后将两个文档之间的相关度确定为加权的各初等相关度之和，例如取决于两个文档之间的联合引用链接数的相关度与取决于两个文档的语义内容的相关度的和。所述过程使得同时考虑文档间的超文本链接所提供的信息以及文档的语义内容所提供的信息成为可能。

[0079] 返回图 1，下面的步骤 S120 到 S135 在于确定用于在文档的组 V 和组 R^d (R 的 d 次笛卡儿幂，其中 R 表示一组实数，而 d 是正整数）的球面 S 之间投影的函数 X。最好将 d 选为等于 2 或 3。

[0080] 所确定的函数 X 是这样的，使得对于至少一个文档 u，两点 X(u) 和 X(v) 之间在 R^d 中的距离越小，相关度就越高，其中 v 是在文档 u 和 v 之间存在相关的文档。

[0081] 根据特定实施例，迭代处理用于确定函数X。该迭代处理的每一次迭代在于基于在前一步骤得到的函数 X_{i-1} 来对于所述组 V 的至少一个文档 u，通过 以 $X_i(u)$ 值替换 $X_{i-1}(u)$

值确定函数 X_i，从而使得优化预定义的准则成为可能；该准则一方面取决于对于所考虑的文档 u 获得的 $X_{i-1}(u)$ 值以及对于所述组 V 的任意文档 v 获得的 $X_{i-1}(u)$ 值，而另一方面取决于在所述组 V 的文档 u 和任意文档 v 之间的相关度 ω(u，v)。选择所述准则，以便使一系列函数 X_i 收敛于呈现上述特性的函数 X。

[0082] 优选地，所述预定义准则的优化在于对于给定的文档 u 最大化数量 Δ(u) 的值，Δ(u) 等于，

[0083] $$\Delta(u) = \sum_{\{u,v\}\in E} \delta(u,v)\left\|X(u) - X(v)\right\|^2$$

[0084] 其中 δ(u，v) ＝ 1－ω(u，v)，0 ≤ ω(u，v) ≤ 1，并且在文档 u 和 v 之间不存在相关时 ω(u，v) ＝ 0。

[0085] 在步骤 S120，确定初始投影函数 X0。优选地，初始函数 X_0 取得球面 S 上的随机值。其后将迭代处理应用于当前函数 $X_i = X_0$。

[0086] 确定投影函数 X 的迭代处理从步骤 S125 开始。迭代与步骤 S125、S130 以及 S135 的执行相对应。迭代由索引 i 来表示。在步骤 S120 的结尾，索引 i 取得其初始值并且等于 0。

[0087] 在步骤 S215，该索引递增：i ＝ i+1。

[0088] 在步骤 S130，针对至少一个文档 u 执行下面的操作：

[0089] －确定 $Y(U) = \sum_{v\in V-\{u\}} \delta(u,v)X_{i-1}(v)$ 的值，

[0090] －若 Y(U) ≠ 0，则通过 $X_i(u) = -Y(U)/\|Y(U)\|$ 通过 Y(U) 计算 $X_i(u)$，

[0091] －若 Y(U) ＝ 0，则 $X_i(u)$ 采取等于 $X_{i-1}(u)$。

[0092] 在步骤 S135，确定迭代处理是否结束。优选地，该处理被迭代充分多的次数，用于函数 X 针对组 V 的每一个文档 u 修改至少一次。

[0093] 即使使用随机开始函数，一系列函数 X_i 也能快速收敛，可以对于一组文档迭代有限次数。

[0094] 终止迭代的判断还可以基于：

[0095] －已经执行的迭代次数，

[0096] －在每次迭代之后执行的函数收敛性的测量结果。

[0097] 该收敛性的测量结果可以通过在每次迭代之后以如下方式计算和 Δ_i：

[0098] $$\Delta_i = \sum_{u\in V} \left\|X_i(u) - X_{i-1}(u)\right\|$$

[0099] 并且通过固定阈值来执行，所述阈值可能取决于组 V 的文档 u 的数目，低于 所述阈值则迭代处理终止。

[0100] 在步骤 S135，如果采取终止迭代处理的判断，则在其之后执行步骤 S140；否则从步骤 S125 开始，执行以下迭代。

[0101] 在步骤 S140，基于在上次迭代获得的函数 X 所取得的值，对于文档的组 V 的至少一部分执行分类操作。

[0102] 依靠所确定的投影函数 X，球面 S 上的点 X(u) 的位置取决于文档 u 和其他文档之间的链接。特别地，两点之间的距离表示与这两个点对应的文档之间的相关度。

[0103] 可以想到使用其他数学准则来使得初始随机函数收敛到这样的函数。

[0104] 在组 V 是从由搜索引擎执行的搜索中产生的一组 WEB 页面的情况下，该分类操作可以针对：

[0105] - 通过检测具有距离其他投影最远的投影的页面来选择最初始的页面；

[0106] - 通过检测其投影基本上接近于一群页面的投影的页面来过滤包含“兜售信息”的页面（指向彼此的页面）；

[0107] - 选择其投影满足所确定的准则的页面。

[0108] 根据第一变体，分类操作包括以下操作：

[0109] - 针对组 V 的任意对 {u，v} 的距离值 d(u，v) = ‖X(u)-X(v)‖ 的计算，

[0110] - 组 V 的至少一个子组 V_1 的确定，其中值 d(u，v) 满足预定义的准则，例如高于或低于预定义的阈值。

[0111] 该第一变体使得检测球面上的点的簇成为可能，并且因此使得确定对应的文档簇成为可能。

[0112] 根据第二变体，分类操作包括存在于确定子组 V_1 的操作，对于所述子组任意点 X(u) 属于所确定的组，例如属于 R^d 中的空间中的预定义区域。

[0113] 该区域可以是例如球面的、立方体的内部体积，或在 R^d 的球面 S 上限定的其它表面。通过针对几个预定义的区域重复这些处理，可以构建所述组文档的划分或分割。

[0114] 根据本发明的方法由此使得基于由所确定的函数 X 所取的值，关于一组文档执行任意种类的分类操作成为可能。

[0115] 此外，可以证明用于确定函数 X 的处理快速收敛。

[0116] 此外，当相关度基于超文本链接数的而确定时，用于该处理的迭代的计算时间正比于该超文本链接数。本发明的所述方法因此可以用于大量页面。

[0117] 最后，要是修改该组电子文档，（通过文档添加、文档删除或文档之间链接的修改），它能够从对于未修改的组而获得的函数 X 开始，然后对于一些所选择的文档（优选地，至少针对已经经历修改或已经被添加的文档）进行步骤 130 的执行，以便确定经校正的、考虑经修改的电子文档组的函数 X。因此本发明特别适用于包含大量文档的组的处理，所述文档的部分被时常更新。

[0118] 在根据本发明的方法的变体中，产生函数 X 的图解表示法，也就是说所述球面以及位于所述球面上的点 X(u) 的表示法。产生这样的图解表示法的事实使得可能促进由用户对于相关文档组的选择。该表示法可以例如以二维制图 (cartgraphic) 表示法的形式来完成，其中每个文档由与针对该文档所确定的函数 X 的值相对应的绘图符号来区分。

[0119] 由此本发明有助于一实施例，其中在用户计算机终端上显示该图解表示法，所述用户计算机终端包括显示屏以及图形选择工具（例如与指示器结合使用的鼠标，使得在屏幕上限定图像区域成为可能），该工具适于选择图解表示法的至少一部分。

[0120] 然后用户能够执行与他所选择的一个或多个组的文档对应的图解表示法的一个或多个部分。终端经由图形选择工具来获得限定所选择的部分的数据。基于这些数据，终端分类文档的组 V。产生例如减少的一列文档，与其投影位于由用户所选择的部分中的文档对应。可选择地，相反，排除其投影位于所选择的部分中的文档。基于由用户保存的一列文档，可以执行附加的分类操作，这些操作或者基于文档的属性或它们的相关度而自动地执行，或者基于在初始选择的部分之内选择的新的部分而手动地执行。

[0121] 查看由搜索引擎执行的搜索的结果的这种模式对于用户是特别工效的(ergonomic)。它以一组共同接近点的形式来显示文档的一致。

[0122] 诸如本发明中所定义的基于投影的表示法因此使得通过图形选择工具可视地并且手动地,或者根据与在所产生的表示法中的这些文档的位置相链接的预定义的准则来自动地执行分类或分级。

[0123] 根据优选实现,根据本发明的分类电子文档的方法的步骤由计算机程序的指令而确定。

[0124] 这里,“计算机程序”被理解为表示一个或多个计算机程序,当其由合适的计算机系统执行时,形成其目的是实现本发明的一组(软件)。

[0125] 然后当前述程序被载入到集成的计算机装置(例如载入到链接的用户终 端)时,如果适于因特网类型的网络并且装配有因特网浏览器软件,则实现根据本发明的方法。

[0126] 因此,本发明的主题也是这样的计算机程序,特别以存储在信息介质上的软件的形式。这样的信息介质可以包括能够存储根据本发明的程序的任意实体或装置。

[0127] 例如,正被讨论的介质可以包括诸如 ROM 之类的硬件存储装置,例如 CD-ROM 或微电子电路 ROM,或者磁记录装置,例如硬盘。作为变体,信息介质可以是集成了所述程序的集成电路,所述电路适于执行或用在正被讨论的方法的执行中。

[0128] 此外,信息介质还可以是可传送的非硬件介质,诸如可以通过无线电或其他装置经由电或光缆传输的电或光信号。根据本发明的程序可以特别地从因特网类型的网络中下载。

[0129] 从设计的角度来看,根据本发明的计算机程序可以使用任意编程语言并且可以以源代码、目标代码或在源代码和目标代码之间的中间代码(如部分经编译的代码)的形式,或者以用于实现根据本发明的方法的任意其他所期望的形式。

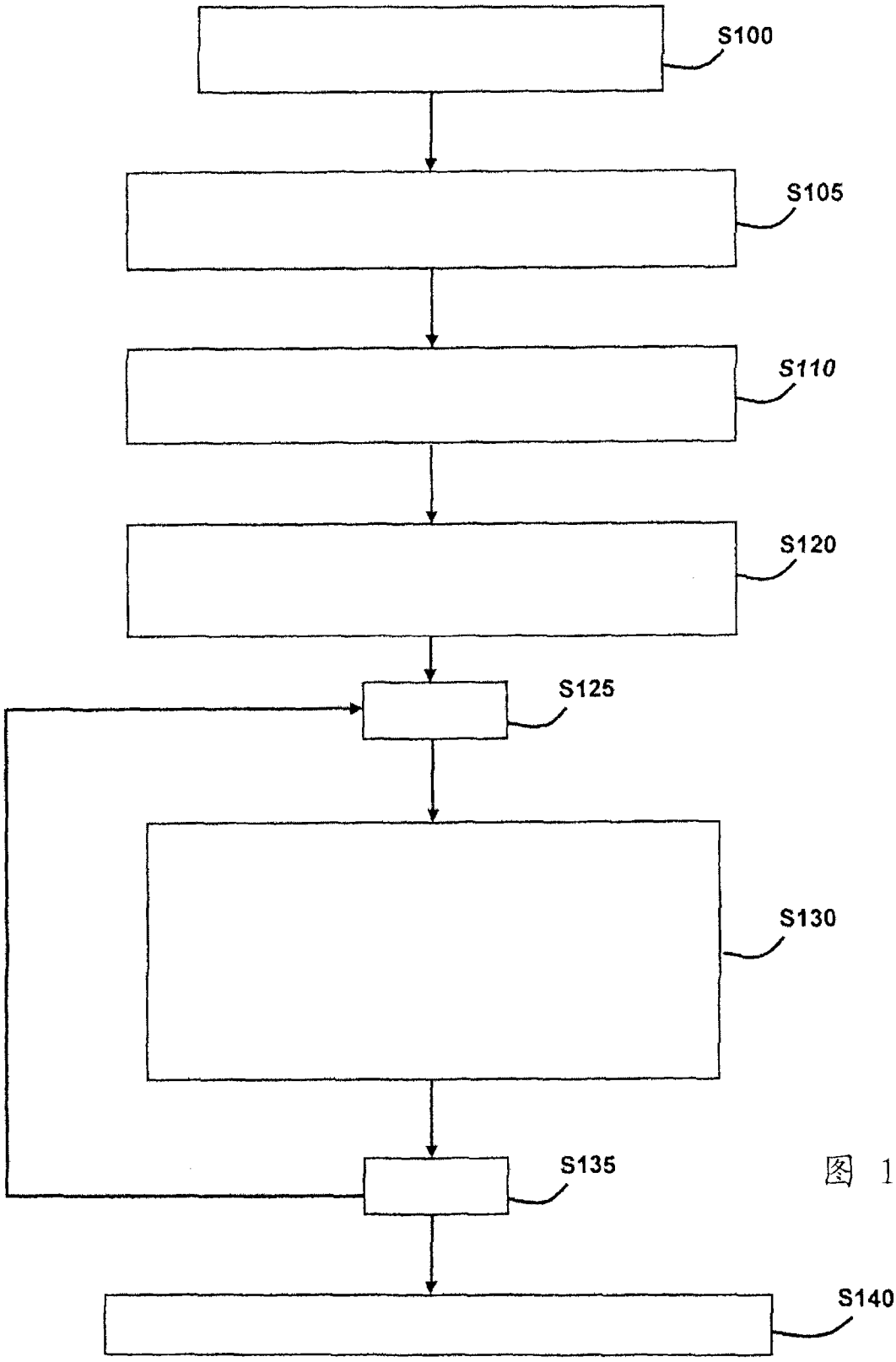

图 1