



(12) 发明专利

(10) 授权公告号 CN 1879107 B

(45) 授权公告日 2014. 10. 15

(21) 申请号 200480033254. 8

代理人 康建忠

(22) 申请日 2004. 09. 15

(51) Int. Cl.

G06F 17/30 (2006. 01)

(30) 优先权数据

60/507, 617 2003. 09. 30 US

10/748, 664 2003. 12. 31 US

(85) PCT国际申请进入国家阶段日

2006. 05. 11

(56) 对比文件

US 2002/0198875 A1, 2002. 12. 26, 摘要、第36、46-68 段, 权利要求 1-6.

审查员 徐春

(86) PCT国际申请的申请数据

PCT/US2004/030000 2004. 09. 15

(87) PCT国际申请的公布数据

W02005/033978 EN 2005. 04. 14

(73) 专利权人 GOOGLE 公司

地址 美国加利福尼亚

(72) 发明人 阿努拉格·阿查雅 马特·卡特斯

杰弗里·迪安 保罗·哈阿

莫尼卡·亨辛格 厄斯·霍尔泽勒

史蒂夫·劳伦斯 卡尔·菲勒格

奥尔坎·瑟斯诺格鲁 西蒙·佟

(74) 专利代理机构 中国国际贸易促进委员会专

利商标事务所 11038

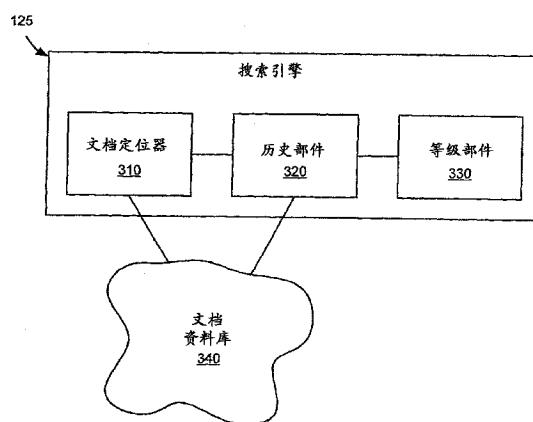
权利要求书7页 说明书14页 附图4页

(54) 发明名称

基于历史数据的信息检索

(57) 摘要

系统 (125) 识别文档并获得与所述文档有关的一种或多种历史数据。系统 (125) 可以至少部分基于一种或多种历史数据, 来生成用于所述文档的分值。



1. 一种计分文档的方法,包括 :

识别文档;

获得与所述文档有关的多种历史数据,所述多种历史数据至少包括:

关于与所述文档关联的初始日期的数据,其中所述初始日期至少基于以下之一:

搜索引擎首次获悉或索引所述文档的日期,

搜索引擎首次发现到所述文档的链接的日期,或者

在另一个文档中首次参考所述文档的日期,

并且其中通过域注册所述文档的日期和所述文档至少包括阈值数目页的日期中的至少一个可以用作所述初始日期;

关于文档内容随时间改变的数据,其中所述关于文档内容随时间改变的数据基于:

更新频率,其基于在一段时间周期内所述文档的内容多久发生改变,和

更新量,其基于在一段时间周期内所述文档的内容改变多少;以及

至少另一种数据,其中所述至少另一种数据至少包括以下之一:

关于一个或多个在先搜索查询的查询分析数据,其中针对所述一个或多个在先搜索查询,所述文档被标识为搜索结果;

关于到或来自所述文档的链接的行为的基于链接的标准;

关于与到所述文档的链接关联的锚文本的数据;

关于与所述文档关联的广告通信量的时间变化特性的数据;

关于所述文档的用户行为数据;

关于与所述文档关联的域的合法性的域相关数据;

关于所述文档的等级历史的数据;

与所述文档关联的用户维护或生成的数据,其中,用户维护或生成的数据与下列中的至少一个有关:与一个用户或多个用户有关的喜好列表、书签、临时文件和缓冲文件;

关于与到所述文档的链接相关联的锚文本中的唯一字、二元语法或短语的数据;

关于独立对等体的连接的数据,或

关于与所述文档关联的随时间变化的文档标题的数据;以及

至少部分基于关于所述初始日期的数据、关于文档内容随时间改变的数据和与所述文档关联的所述至少另一种数据生成用于所述文档的分值,其中,生成分值包括:

确定所述用户维护或生成的数据是否表示用户对所述文档感兴趣;以及

至少部分基于用户维护或生成的数据是否表示用户对所述文档感兴趣,来计分所述文档。

2. 如权利要求 1 所述的方法,其中,所述文档包括多个文档;以及

其中,计分所述文档包括:

基于对应于文档的初始日期,确定每一个文档的寿命,

基于文档的寿命,确定文档的平均寿命;以及

至少部分基于文档的寿命和平均寿命之间的差值,来计分所述文档。

3. 如权利要求 1 所述的方法,其中,生成用于所述文档的分值包括:至少部分基于从对应于所述文档的初始日期测定的逝去时间,来计分所述文档。

4. 如权利要求 1 所述的方法,其中,所述文档内容的多久发生变化是基于下列中的至

少一个 : 变化之间的平均时间、一段时间周期内的变化次数、或者当前时间周期内的变化率与先前时间周期内的变化率的比较。

5. 如权利要求 1 所述的方法, 其中, 所述文档内容的改变多少是基于下列中的至少一个 : 在一段时间周期内与所述文档有关的新页数、与所述文档有关的新页数和与所述文档有关的总页数的比率、或者在一段时间周期期间已经改变的文档内容的百分比。

6. 如权利要求 1 所述的方法, 其中, 所述更新量是基于以下内容确定的 :

基于各部分的重要性的度量, 不同地加权所述文档内容的不同部分 ; 以及将所述文档内容的变化量确定为所述内容的不同加权部分的函数。

7. 如权利要求 1 所述的方法, 其中, 生成分值包括 :

至少部分基于所述更新量, 来计分所述文档。

8. 如权利要求 1 所述的方法, 其中, 所述多种历史数据包括所述查询分析数据 ; 以及其中, 生成分值包括 :

当所述文档被包括在一个搜索结果集中时, 确定随时间所述文档被选择的程度 ; 以及至少部分基于当所述文档被包括在所述搜索结果集中时随时间所述文档被选择的程度, 来计分所述文档。

9. 如权利要求 8 所述的方法, 其中, 计分所述文档包括 : 当在一段时间周期上相比于所述搜索结果集中的其他文档所述文档被更经常选择时, 向所述文档分配更高分值。

10. 如权利要求 1 所述的方法, 其中, 所述多种历史数据包括所述查询分析数据 ; 以及其中, 生成分值包括 :

确定所述文档是否与在搜索查询中随着时间以增加的频率出现的搜索项有关 ; 以及至少部分基于所述文档是否与搜索项有关, 来计分所述文档。

11. 如权利要求 1 所述的方法, 其中, 所述多种历史数据包括所述查询分析数据 ; 以及其中, 生成分值包括 :

确定所述文档是否与随时间大致保持不变但导致随时间改变的结果的查询有关 ; 以及至少部分基于所述文档是否与导致随时间改变的结果的查询有关, 来计分所述文档。

12. 如权利要求 1 所述的方法, 其中, 所述多种历史数据包括所述查询分析数据 ; 以及其中, 生成分值包括 :

确定所述文档是否过期 ; 以及

至少部分基于所述文档是否过期, 来计分所述文档。

13. 如权利要求 12 所述的方法, 其中, 计分所述文档包括 :

当确定所述文档过期时, 确定是否认为该过期文档有利于搜索查询 ; 以及

至少部分基于当确定所述文档过期时是否认为该过期文档有利于搜索查询, 来计分所述文档。

14. 如权利要求 13 所述的方法, 其中, 确定是否认为过期文档有利于搜索查询至少部分基于在用于搜索查询的时间上, 在最近文档上多常选择过期文档。

15. 如权利要求 1 所述的方法, 其中, 所述多种历史数据包括关于基于链接的标准的数据 ; 以及

其中, 生成分值包括 :

确定与所述文档有关的链接行为 ; 以及

至少部分基于与所述文档有关的链接行为,来计分所述文档。

16. 如权利要求 15 所述的方法,其中,链接行为与指向所述文档的一个或多个链接的出现或消失的至少一个有关。

17. 如权利要求 16 所述的方法,其中,一个或多个链接的出现与下列中的至少一个有关:出现到所述文档的新链接的日期、一个或多个链接随时间出现的速率、或者在一段时间周期期间出现的一个或多个链接的数量;而一个或多个链接的消失与下列中的至少一个有关:到所述文档的现有链接消失的日期、一个或多个链接随时间消失的速率、或者在一段时间周期期间消失的一个或多个链接的数量。

18. 如权利要求 15 所述的方法,其中,确定与所述文档有关的链接的行为包括监视下列中的至少一个:与所述文档有关的链接的时间变化行为、在一段时间周期期间出现或消失多少与所述文档有关的链接、或与所述文档有关的现有链接的消失相比是否存在倾向出现与所述文档有关的新链接。

19. 如权利要求 1 所述的方法,其中,所述多种历史数据包括关于基于链接的标准的数据;

其中,生成分值包括:

确定与所述文档有关的链接的新鲜度的度量;

基于所确定的新鲜度的度量,向链接分配权重;以及

至少部分基于分配给与所述文档有关的链接的权重,来计分所述文档。

20. 如权利要求 19 所述的方法,其中,与所述文档有关的链接的新鲜度的度量是基于下列中的至少一个:链接出现的日期、链接变化的日期、与该链接有关的锚文本的出现日期、与该链接有关的锚文本变化的日期、包含该链接的链接文档出现的日期或包含该链接的链接文档变化的日期。

21. 如权利要求 19 所述的方法,其中,分配给链接的权重是基于下列中的至少一个:与包含该链接的文档关联的信任的度量、包含该链接的文档的权威的度量、或包含该链接的文档的新鲜度的度量。

22. 如权利要求 19 所述的方法,其中,计分文档包括:

确定指向所述文档的每个链接的寿命;

基于链接的寿命,来确定与链接有关的寿命分布;以及

至少部分基于与链接有关的寿命分布,来计分文档。

23. 如权利要求 1 所述的方法,其中,所述多种历史数据包括关于锚文本的数据;以及其中,生成分值包括:

识别与到所述文档的链接有关的锚文本中随时间的变化;以及

至少部分基于与到所述文档的链接有关的锚文本的变化,计分所述文档。

24. 如权利要求 1 所述的方法,其中,所述多种历史数据包括关于锚文本的数据;以及其中,生成分值包括:

确定文档内容是否改变使得所述内容不同于与到所述文档的一个或多个链接有关的锚文本;以及

至少部分基于所述文档的内容是否改变使得所述内容不同于与到所述文档的一个或多个链接有关的锚文本,来计分所述文档。

25. 如权利要求 1 所述的方法,其中,所述多种历史数据包括关于锚文本的数据;以及其中,生成分值包括:

确定与到所述文档的一个或多个链接有关的锚文本的新鲜度的度量;以及至少部分基于与到所述文档的一个或多个链接有关的锚文本的新鲜度的度量,来计分所述文档。

26. 如权利要求 25 所述的方法,其中,与到所述文档的链接有关的锚文本的新鲜度的度量是基于下列中的至少一个:锚文本的出现日期、锚文本的改变日期、与锚文本有关的链接的出现日期、与锚文本有关的链接的改变日期、所述文档的出现日期或所述文档的改变日期。

27. 如权利要求 1 所述的方法,其中,所述多种历史数据进一步包括关于文档通信量的时间变化特性的数据;以及

其中,生成分值包括:

确定与文档有关的通信量的特性;以及

至少部分基于与所述文档有关的通信量的特性,来计分所述文档。

28. 如权利要求 27 所述的方法,其中,确定与所述文档有关的通信量的特性包括:分析与所述文档有关的通信量模式以便识别通信量模式随时间的变化。

29. 如权利要求 1 所述的方法,其中,所述多种历史数据包括用户行为数据;以及其中,生成分值包括:

确定与文档有关的用户行为;以及

至少部分基于与文档有关的用户行为,来计分所述文档。

30. 如权利要求 29 所述的方法,其中,用户行为与在搜索结果集内文档被选择的次数以及一个或多个用户访问所述文档所花费的时间量中的至少一个有关。

31. 如权利要求 1 所述的方法,其中,所述多种历史数据包括域相关数据;以及其中,生成分值包括:

分析对应于与文档有关的域随时间的域相关信息;以及

至少部分基于分析结果,来计分所述文档。

32. 如权利要求 31 所述的方法,其中,计分所述文档包括:

确定与所述文档有关的域是否合法;以及

至少部分基于与所述文档有关的域是否合法,来计分所述文档。

33. 如权利要求 31 所述的方法,其中,域相关信息与下列中的至少一个有关:域的届满日期、与域有关的域名服务器记录、或与域有关的名称服务器。

34. 如权利要求 1 所述的方法,其中,所述多种历史数据包括关于等级历史的数据;以及

其中,生成分值包括:

确定所述文档的先前等级历史;以及

至少部分基于所述文档的先前等级历史,来计分所述文档。

35. 如权利要求 34 所述的方法,其中,计分所述文档包括:

确定在一段时间周期上所述文档在等级方面移动的数量或速率;以及

至少部分基于所述文档在等级方面移动的数量或速率,来计分所述文档。

36. 如权利要求 34 所述的方法,其中,先前等级历史是基于下列中的至少一个 : 随时间所述文档被选择为搜索结果的查询数量、随时间所述文档被选择为搜索结果的速率、季节性、突发性或者对 URL 查询对,分值随时间的变化。

37. 如权利要求 34 所述的方法,其中,确定文档的先前等级历史包括监视随时间文档等级的等级峰值。

38. 如权利要求 1 所述的方法,其中,计分所述文档包括 :

分析随时间用户维护或生成的数据,来识别下列中的至少一个 : 增加或移出文档的趋势、所述文档增加到用户维护或生成的数据或从中移出的速率、或者所述文档是增加到用户维护或生成的数据、从用户维护或生成的数据删除还是通过用户维护或生成数据被访问 ; 以及

至少部分基于分析结果,来计分所述文档。

39. 如权利要求 1 所述的方法,其中,所述多种历史数据包括关于锚文本的数据 ; 以及其中,生成分值包括 :

确定与到所述文档的一个或多个链接有关的锚文本的增长图 ; 以及

至少部分基于与到所述文档的一个或多个链接有关的锚文本的增长图,来计分所述文档。

40. 如权利要求 1 所述的方法,其中,所述多种历史数据包括与独立对等体的连接有关的数据 ; 以及

其中,生成分值包括 :

确定包括到所述文档的链接的独立对等体的数量增长 ; 以及

至少部分基于独立对等体的数量,来计分所述文档。

41. 如权利要求 1 所述的方法,其中,所述多种历史数据包括关于文档主题的数据 ; 以及

其中,生成分值包括 :

执行与所述文档有关的主题提取 ;

监视文档主题随时间的变化 ; 以及

至少部分基于文档主题的变化,来计分所述文档。

42. 如权利要求 1 所述的方法,进一步包括 :

获得搜索查询,其中,将所识别的文档识别为与该搜索查询有关 ; 以及

基于所述文档与搜索查询有多相关,生成用于所述文档的相关分值 ; 以及

其中,生成用于所述文档的分值至少部分基于所述多种历史数据和相关分值。

43. 一种用于计分文档的系统,包括 :

用于识别文档的装置 ;

用于获得与所述文档有关的多种历史数据的装置,所述多种历史数据至少包括 :

关于与所述文档关联的初始日期的数据,其中所述初始日期至少基于以下之一 :

搜索引擎首次获悉或索引所述文档的日期,

搜索引擎首次发现到所述文档的链接的日期,或者

在另一个文档中首次参考所述文档的日期,

并且其中通过域注册所述文档的日期和所述文档至少包括阈值数目页的日期中的至

少一个可以用作所述初始日期；

关于文档内容随时间改变的数据,其中所述关于文档内容随时间改变的数据基于：

更新频率,其基于在一段时间周期内所述文档的内容多久发生改变,和

更新量,其基于在一段时间周期内所述文档的内容改变多少;以及

至少另一种数据,其中所述至少另一种数据至少包括以下之一：

关于一个或多个在先搜索查询的查询分析数据,其中针对所述一个或多个在先搜索查询,所述文档被标识为搜索结果；

关于到或来自所述文档的链接的行为的基于链接的标准；

关于与到所述文档的链接关联的锚文本的数据；

关于与所述文档关联的广告通信量的时间变化特性的数据；

关于所述文档的用户行为数据；

关于与所述文档关联的域的合法性的域相关数据；

关于所述文档的等级历史的数据；

与所述文档关联的用户维护或生成的数据,其中,用户维护或生成的数据与下列中的至少一个有关:与一个用户或多个用户有关的喜好列表、书签、临时文件和缓冲文件；

关于与到所述文档的链接相关联的锚文本中的唯一字、二元语法或短语的数据；

关于独立对等体的连接的数据,或

关于与所述文档关联的随时间变化的文档标题的数据;以及

至少部分基于关于初始日期的数据、关于文档内容随时间改变的数据和与所述文档关联的所述至少另一种数据来生成用于所述文档的分值的装置,用于生成分值的装置包括:

用于确定所述用户维护或生成的数据是否表示用户对所述文档感兴趣的装置,以及

用于至少部分基于用户维护或生成的数据是否表示用户对所述文档感兴趣,来计分所述文档的装置。

44. 一种用于计分文档的系统,包括:

历史部件,配置成获得与文档有关的多种历史数据,所述多种历史数据包括:

关于与所述文档关联的初始日期的数据其中所述初始日期至少基于以下之一:

搜索引擎首次获悉或索引所述文档的日期,

搜索引擎首次发现到所述文档的链接的日期,或者

在另一个文档中首次参考所述文档的日期,

并且其中通过域注册所述文档的日期和所述文档至少包括阈值数目页的日期中的至少一个可以用作所述初始日期；

关于文档内容随时间改变的数据,其中所述关于文档内容随时间改变的数据基于:

更新频率,其基于在一段时间周期内所述文档的内容多久发生改变,和

更新量,其基于在一段时间周期内所述文档的内容改变多少;以及

至少另一种数据,其中所述至少另一种数据至少包括以下之一:

关于一个或多个在先搜索查询的查询分析数据,其中针对所述一个或多个在先搜索查询,所述文档被标识为搜索结果；

关于到或来自所述文档的链接的行为的基于链接的标准；

关于与到所述文档的链接关联的锚文本的数据；

关于与所述文档关联的广告通信量的时间变化特性的数据；
关于所述文档的用户行为数据；
关于与所述文档关联的域的合法性的域相关数据；
关于所述文档的等级历史的数据；
与所述文档关联的用户维护或生成的数据，其中，用户维护或生成的数据与下列中的至少一个有关：与一个用户或多个用户有关的喜好列表、书签、临时文件和缓冲文件；
关于与到所述文档的链接相关联的锚文本中的唯一字、二元语法或短语的数据；
关于独立对等体的连接的数据，或
关于与所述文档关联的随时间变化的文档标题的数据；以及
等级部件，配置成：
至少部分基于关于初始日期的数据、关于文档内容随时间改变的数据和与所述文档关联的所述至少另一种数据来生成用于所述文档的分值，其中，生成分值包括：
确定所述用户维护或生成的数据是否表示用户对所述文档感兴趣，以及
至少部分基于用户维护或生成的数据是否表示用户对所述文档感兴趣，来计分所述文档。

基于历史数据的信息检索

技术领域

[0001] 本发明通常涉及信息检索系统，以及更具体地说，涉及用于至少部分基于与相关文档有关的历史数据，来生成搜索结果的系统和方法。

背景技术

[0002] 万维网（“网页”）包含大量信息。搜索引擎帮助用户通过编目录网页文档，来定位该信息的所需部分。通常，响应用户的请求，搜索引擎返回到与该请求有关的文档的链接。

[0003] 搜索引擎可以将用户兴趣的确定基于由用户提供的搜索项（被称为搜索查询）。搜索引擎的目标是基于搜索查询，来识别到高质量相关结果的链接。典型地，搜索引擎通过匹配搜索查询中的术语与预存储的网页文档的资料库来实现此目标。包含用户搜索项的网页文档被视为“命中”并返回给用户。

[0004] 理想地，搜索引擎将响应指定用户搜索查询，为用户提供最相关结果。一种搜索引擎基于比较搜索查询术语与包含在文档中的词来识别相关文档。另一种搜索引擎使用除文档中存在搜索查询术语之外的因素来识别相关文档。一个这种搜索引擎使用与到或来自文档的链接有关的信息来确定文档的相对重要性。

[0005] 这两种搜索引擎力求提供高质量的搜索查询结果。存在会影响由搜索引擎生成的结果质量的几种因素。例如，一些网站生产商使用垃圾邮件技术来人为地抬高他们的等级。同时，可以使“过期”文档（即长时间未更新的那些文档，从而包含过期数据）等级高于“较新”文档（即最近更新的那些文档，从而包含更新的数据）。在一些特定环境下，较高等级的过期文档降低了搜索结果。

[0006] 因此，仍然需要提高由搜索引擎生成的结果的质量。

发明内容

[0007] 与本发明的原理相符的系统和方法可以至少部分基于与文档有关的历史数据来给文档计分。该计分可以用来提高连同搜索查询生成的搜索结果。

[0008] 根据与本发明的原理相符的一个方面，提供一种用于计分文档的方法。该方法可以包括识别文档并获得与所述文档有关的一种或多种历史数据。该方法可以进一步包括至少部分基于一种或多种历史数据，来生成用于所述文档的得分。

[0009] 根据另一方面，提供一种用于计分文档的方法。该方法可以包括确定与所链接的文档有关的连接数据的寿命，以及基于该连接数据的寿命的衰减函数，来分级所链接的文档。

附图说明

[0010] 包含并构成本说明书的一部分的附图示例性本发明的实施例，以及结合说明书，解释本发明。在图中：

[0011] 图 1 是可以实现与本发明的原理相符的系统和方法的示例性网络图；

- [0012] 图 2 是根据与本发明的原理相符的实现,图 1 的客户机和 / 或服务器的示例图 ;
[0013] 图 3 是根据与本发明的原理相符的实现,图 1 的搜索引擎的示例性功能框图 ;以及
[0014] 图 4 是根据与本发明的原理相符的实现,用于计分文档的示例性处理的流程图。

具体实施方式

[0015] 本发明的下述详细描述参考附图。不同图中的相同参考数字可以识别相同或类似的元件。同时,下述详细描述不限制本发明。

[0016] 与本发明的原理相符的系统和方法可以使用例如与所述文档有关的历史数据来计分文档。系统和方法可以使用这些得分来提供高质量搜索结果。

[0017] “文档”如在此所使用的,广泛解释成包括任何机器可读和机器可存储的作品。文档可以包括电子邮件、网站、文件、文件组合、具有与其他文件的嵌入链接的一个或多个文件、新闻组布告、博客、网页广告等等。在因特网的情况下,公用文档是网页。网页通常包括文本信息并可以包括嵌入的信息(诸如元信息、图像、超级链接等等)和 / 或嵌入的指令(诸如 Java 脚本等等)。网页可以对应于文档或部分文档。因此,单词“网页”或“文档”在某些情况下可以互换使用。在其他情况下,网页可以指部分文档,诸如子文档。网页对应于不止单个文档也是可能的。

[0018] 在下述描述中,可以将文档描述为具有到其他文档的链接和 / 或来自其他文档的链接。例如,当文档包括到另一文档的链接时,链接可以被称为“前向链接”。当文档包括来自另一文档的链接时,该链接可以被称为“后向链接”。当使用术语“链接”时,可以指后向链接或前向链接。

[0019] 网络结构的例子

[0020] 图 1 是网络 100 的示例性图,其中,可以实现与本发明的原理相符的系统和方法。网络 100 可以包括经网络 150 连接到多个服务器 120-140 的多个客户机 110。网络 150 可以包括局域网 (LAN)、广域网 (WAN)、电话网,诸如公用交换电话网 (PSTN)、内联网、互联网、存储器设备、另一类型的网络或网络组合。为简化起见,两个客户机 110 和三个服务器 120-140 示例为连接到网络 150。实际上,可以有更多或更少的客户机和服务器。同时,在一些实例中,客户机可以执行服务器的功能,以及服务器可以执行客户机的功能。

[0021] 客户机 110 可以包括客户实体。实体可以被定义为设备,诸如无线电话、个人计算机、个人数字助理 (PDA)、膝上型电脑或另一计算或通信设备、在这些设备的一个上运行的线程或过程和 / 或能由这些设备的一个执行的对象。服务器 120-140 可以包括以与本发明的原理相符的方式,收集、处理、搜索和 / 或维护文档的服务器实体。客户机 110 和服务器 120-140 可以经有线、无线和 / 或光学连接而与网络 150 相连。

[0022] 在与本发明的原理相符的实现中,服务器 120 可以包括可由客户机 110 使用的搜索引擎 125。服务器 120 可以扒 (crawl) 文档的资料库(例如网页)、索引文档以及存储与所扒的文档库中的文档有关的信息。服务器 130 和 140 可以存储或维护可以由服务器 120 扒的文档。尽管服务器 120-140 被示为单独实体,但也可以服务器 120-140 的一个或多个执行服务器 120-140 的另一个或多个的功能的一个或多个。例如,两个或多个服务器 120-140 实现为单个服务器是可能的。也可以将服务器 120-140 的单个实现为两个或多个独立(以及可以分布式)设备。

[0023] 示例性客户机 / 服务器体系结构

[0024] 图 2 是根据与本发明的原理相符的实现, 客户机或服务器实体 (在下文中称为“客户机 / 服务器实体”) 的示例性图, 可以对应于一个或多个客户机 110 和服务器 120-140。客户机 / 服务器实体可以包括总线 210、处理器 220、主存储器 230、只读存储器 (ROM) 240、存储设备 250、一个或多个输入设备 260、一个或多个输出设备 270 以及通信接口 280。总线 210 可以包括一个或多个导线, 允许客户机 / 服务器实体的部件间的通信。

[0025] 处理器 220 可以包括解释和执行指令的一个或多个传统处理器或微处理器。主存储器 230 可以包括随机存取存储器 (RAM) 或另一种动态存储设备, 存储信息和指令以便由处理器 220 执行。ROM 240 可以包括传统 ROM 设备或另一种静态存储设备, 存储用于由处理器 220 使用的静态信息和指令。存储设备 250 可以包括磁性和 / 或光学记录介质及其相应驱动。

[0026] 输入设备 260 可以包括一个或多个传统的机构, 允许操作者将信息输入客户机 / 服务器实体, 诸如键盘、鼠标、笔、语音识别和 / 或生物机构等等。输出设备 270 可以包括一个或多个传统的机构, 向操作者输出信息, 包括显示器、打印机、扬声器等等。通信接口 280 可以包括收发信机类机构, 允许客户机 / 服务器实体与其他设备和 / 或系统通信。例如, 通信接口 280 可以包括用于经网络, 诸如网络 150 与另一设备或系统通信的机构。

[0027] 如下文详细所述, 与本发明的原理相符, 客户机 / 服务器实体执行某些搜索相关操作。客户机 / 服务器实体可以响应执行包含在计算机可读介质, 诸如存储器 230 中的软件指令的处理器 220, 而执行这些操作。计算机可读介质可以被定义为一个或多个物理或逻辑存储设备和 / 或载波。

[0028] 软件指令可以从另一计算机可读介质, 诸如数据存储设备 250, 或经通信接口 280, 从另一设备读入存储器 230 中。包含在存储器 230 中的软件指令可以使处理器 220 执行将在下文所述的过程。另外, 可以使用硬布线电路来代替或结合软件指令来实现与本发明的原理相符的过程。因此, 与本发明的原理相符的实现可以不限于硬布线电路和软件的任何特定组合。

[0029] 示例性搜索引擎

[0030] 图 3 是根据与本发明的原理相符的实现, 搜索引擎 125 的示例性功能框图。搜索引擎 125 可以包括文档定位器 310、历史部件 320 和等级部件 330。如图 3 所示, 文档定位器 310 和历史部件 320 的一个或多个可以连接到文档资料库 340。文档资料库 340 可以包括与例如在由搜索引擎 125 可访问的数据库中先前扒、索引和存储的文档有关的信息。历史数据, 如在下文中更详细地描述, 可以与文档资料库 340 中的每一个文档相关联。历史数据可以存储在文档资料库 340 或其他地方中。

[0031] 文档定位器 310 可以识别其内容与用户搜索查询匹配的文档集。文档定位器 310 可以通过将用户搜索查询中的术语与资料库中的文档进行比较, 初始地从文档资料库 340 定位文档。通常, 用于索引文档并搜索索引集合以返回包含搜索项的文档集的过程在本领域非常公知。因此, 在此不再描述文档定位器 310 的该功能。

[0032] 历史部件 320 可以收集与文档资料库 340 中的文档有关的历史数据。在与本发明的原理相符的实现中, 历史数据可以包括与下列有关的数据: 文档初始日期; 文档内容更新 / 改变; 查询分析; 基于链接的标准; 锚文本 (例如嵌入超级链接的文本, 通常在文档中

被加下划线或者高亮) ; 通信量 ; 用户行为 ; 域相关信息 ; 等级历史 ; 用户维护 / 产生的数据 (例如书签) ; 锚文本中的唯一字、二元语法和短语 ; 独立对等的连接和 / 或文档主题。在下文中另外详细地描述这些不同类型的历史数据。在其他实现中, 历史数据可以包括另外或不同类型的数据。

[0033] 等级部件 330 可以向文档资料库 340 中的一个或多个文档分配等级得分 (在此也简单地称为“计分”)。等级部件 330 可以在搜索查询前、与搜索查询无关或结合搜索查询, 来分配等级得分。当文档与搜索查询相关时 (例如识别为与搜索查询有关), 搜索引擎 125 可以基于等级得分来排序文档并将排序后的文档集返回给提交搜索查询的客户机。与本发明的原理相符, 等级得分是试图量化文档质量的值。在与本发明的原理相符的实现中, 得分至少部分基于来自历史部件 320 的历史数据。

[0034] 示例性历史数据

[0035] 文档初始日期

[0036] 根据与本发明的原理相符的实现, 文档初始日期可以用来生成 (或修改) 与那个文档有关的得分。术语“日期”在此广泛使用并可以由此包括时间和日期度量。如下所述, 存在能用来确定文档初始日期的几种技术。这些技术中的一些在它们会受期望提高与文档有关的得分的第三方影响的意义方面是“有偏差”。其他技术无偏差。这些技术中的任何一种、这些技术的组合或其他技术可以用来确定文档的初始日期。

[0037] 根据一种实现, 可以由搜索引擎 125 首次获悉或索引文档的日期, 来确定文档的初始日期。搜索引擎 125 可以通过扒、从“外部”源向搜索引擎 125 提交文档 (或其表示 / 概述)、扒或基于提交的索引技术的组合, 或以其他方式, 来发现所述文档。另外, 可以由搜索引擎 125 首次发现到所述文档的链接的日期, 来确定文档的初始日期。

[0038] 根据另一实现, 通过域注册文档的日期可以被用作文档的初始日期的表示。根据另一实现, 可以使用在另一文档, 诸如新闻文章、新闻组、电子邮件列表或一个或多个这些文档的组合中第一次参考文档的时间来推断文档的初始日期。根据另一实现, 文档至少包括阈值数目页的日期可以被用作文档的初始日期的表示。根据另一实现, 可以使文档的初始日期等于服务器寄存文档的与所述文档有关的时间戳。其他技术, 在此未具体提及的, 或技术组合也能用来确定或推断文档的初始日期。

[0039] 搜索引擎 125 可以将文档的初始日期用于计分文档。例如, 可以假定具有相当近的初始日期的文档将不具有来自其他文档的多个链接 (即后向链接)。对基于到 / 来自文档的链接数的现有的基于链接的计分技术, 该新文档可能得分低于具有更多链接 (例如向后链接) 的较早文档。当考虑文档的初始日期时, 然而, 可以基于文档的初始日期, 来 (正或负地) 修改文档的得分。

[0040] 假定由 10 个后向链接参考的具有初始日期为昨天的文档的例子。所述文档可以由搜索引擎 125 计分高于由 100 个后向链接参考的具有初始日期为 10 年前的文档, 因为前者的链接增长率相对高于后者。尽管后向链接数的增长的尖峰速率 (spiky rate) 可以是由搜索引擎 125 用来计分文档的因素, 但也可能是发尝试信号来向搜索引擎 125 发送垃圾邮件。因此, 在这种情况下, 搜索引擎 125 实际上可以降低文档的分值来降低发送垃圾邮件的影响。

[0041] 因此, 根据与本发明的原理相符的实现, 搜索引擎 125 可以使用文档的初始日期

来确定创建到所述文档的链接的速率（例如作为基于从初始日期以来或在那个周期中的一些窗口创建的链接数的每单位时间的平均值）。然后，能使用该速率来计分所述文档，例如向更常生成链接的文档提供更大权重。

[0042] 在一个实现中，搜索引擎 125 可以修改文档的基于链接的分值如下：

$$H = \frac{L}{\log(F+2)}$$

[0044] 其中，H 指历史调整的链接分值，L 可以指为所述文档提供的链接分值，其可以使用基于到 / 来文档的链接而为文档分配分值的任何已知链接计分技术（例如在 U.S. 专利 No. 6, 285, 999 中所描述的计分技术）来导出，以及 F 可以指从与所述文档有关的初始日期（或该周期内的窗口）测量的逝去时间。

[0045] 对于一些查询，较早文档比新的更有利。因此，可以基于与结果集的平均寿命的差值（寿命方面），来调整文档的分值。换句话说，搜索引擎 125 可以确定结果集中每个文档的寿命（例如使用它们的初始日期），确定文档的平均寿命，以及基于文档的寿命和平均寿命之间的差值，来（正或负）地修改文档的分值。

[0046] 总的来说，搜索引擎 125 可以至少部分基于与文档的初始日期有关的信息，来生成（或修改）与文档有关的分值。

[0047] 内容更新 / 改变

[0048] 根据与本发明的原理相符的实现，与文档内容随时间改变的方式有关的信息可以被用来生成（或修改）与那个文档有关的分值。例如，其内容经常被编辑的文档得分不同于其内容随时间保持不变的文档。同时，相对多内容随时间更新的文档的计分可以不同于随时间更新相对少量内容的文档。

[0049] 在一个实现中，搜索引擎 125 可以生成内容更新得分 (U) 如下：

$$U = f(UF, UA)$$

[0051] 其中，f 可以指函数，诸如求和或加权和，UF 可以指表示多久更新文档（或网页）的更新频率得分，以及 UA 可以指表示文档（或网页）随时间改变多少的更新量得分。UF 可以以多个方式来确定，包括更新之间的平均时间、在指定时限内的更新次数等等。

[0052] UA 也可以确定为一个或多个因素的函数，诸如在一个时间周期内与文档有关的“新”或唯一页的数量。另一因素可以包括一个时间周期内与文档有关的新或唯一页的数量和与那个文档有关的总页数的比率。另一因素可以包括在一个或多个时间周期内更新文档的数量（例如文档的可见内容的 n% 可以随周期 t 改变（例如最近 m 个月）），其可以是平均值。另一因素可以包括在一个或多个时间周期内（例如在最近 x 天内），文档（或网页）改变的数量。

[0053] 根据一个示例性实现，UA 可以确定为文档内容的不同加权部分的函数。例如，当确定 UA 时，认为如果更新 / 改变不重要的内容，诸如 Java 脚本、注释、广告、导航要素、样板资料或日期 / 时间标签，则给予相对小的权重或甚至完全忽略。另一方面，当确定 UA 时，认为如果（例如经常、更近、更广泛等等）更新 / 改变很重要的内容，诸如与前向链接有关的标题或锚文本，则给予比其他内容改变更高的权重。

[0054] UF 和 UA 可以用其他方式来影响分配给文档的分值。例如，能将当前时间周期中的改变率与在另一（例如在前）时间周期中的改变率进行比较，来确定存在加速还是减速趋势。改变率增加的文档可以比改变率稳定的那些文档计分更高，即使那一改变率相当高。

改变量也可以是该计分中的因素。例如，当改变量大于一些阈值时，改变率增加的文档可以得分高于改变率稳定或改变量小于阈值的那些文档。

[0055] 在一些情况下，当监视文档的内容改变时，数据存储资源可能不足以存储那些文档。在这种情况下，搜索引擎 125 可以存储文档的表示并监视这些表示的变化。例如，搜索引擎 125 可以存储文档的“签名”，代替（整个）文档本身以检测文档内容的改变。在这种情况下，搜索引擎 125 可以存储用于文档（或网页）的术语矢量并监视其相对大的改变。根据另一实现，搜索引擎 125 可以存储和监视确定为重要或最频繁发生（除“停止字”外）的文档的相对小部分（例如几个术语）。

[0056] 根据另一实现，搜索引擎 125 可以存储文档的概述或其他表示并监视该信息的变化。根据另一实现，搜索引擎 125 可以生成用于所述文档的相似度散列（可以用来检测文档的较近复制）并监视其变化。相似度散列的变化可以被视为表示其相关文档中的相对大变化。在其他实现中，可以使用其他技术来监视文档的变化。在存在足够数据存储资源的情况下，可以存储和使用整个文档来确定变化，而不是文档的一些表示。

[0057] 对一些查询，具有最近未改变的内容的文档可以比具有最近改变过的内容的文档更有利。因此，可以基于与结果集的平均改变日期的差值来调整文档的分值可能是有利的。换句话说，搜索引擎 125 可以确定结果集中每一个文档的内容最后一次改变的日期，确定所述文档的平均改变日期，并基于文档的改变日期和平均改变日期之间的差值，来修改文档的分值（正或负）。

[0058] 总的来说，搜索引擎 125 可以至少部分基于与文档的内容随时间改变的方式有关的信息，来生成（或修改）与文档有关的分值。对于包括属于多个个人或公司的内容的非常大的文档，分值可以对应于每一个子文档（即，属于单个人或公司或由其更新的内容）。

[0059] 查询分析

[0060] 根据与本发明的原理相符的实现，可以使用一个或多个基于查询的因素来生成（或改变）与文档有关的分值。例如，当文档包括在搜索结果集中时，一个基于查询的因素涉及随时间选择该文档的程度。在这种情况下，搜索引擎 125 可以使用户相对经常 / 日益增加选择的文档的得分高于其他文档。

[0061] 另一基于查询的因素可以涉及在查询中出现的某些搜索项随时间的出现。特定搜索项集可以随时间周期递增地出现在查询中。例如，与正变得 / 已经变为流行的“热门”标题或分裂新闻事件有关的术语将可能在时间周期上频繁地出现。在这种情况下，搜索引擎 125 可以使与这些搜索项（或查询）相关的文档的得分高于不与这些术语有关的文档。

[0062] 另一基于查询的因素可以涉及通过类似查询生成的搜索结果数目随时间的改变。由类似查询生成的搜索结果数的显著增加例如可以表示热门标题或分裂新闻，并使搜索引擎 125 增加与这些查询有关的文档的得分。

[0063] 另一基于查询的因素可以涉及随时间保持相对恒定但会导致随时间改变的结果的查询。例如，与“世界职业棒球锦标赛”有关的查询导致随时间改变的搜索结果（例如与特定队有关的文档控制在特定年或年度内的搜索结果）。该改变能被监视并用来相应地计分文档。

[0064] 另一基于查询的因素可以涉及作为搜索结果返回的文档的“过期”。文档过期可以基于以下因素，诸如文档创建日期、锚增长、通信量、内容变化、前向 / 后向链接增长等等。

对于一些查询，最近文档非常重要（例如如果搜索常问问题（FAQ）文件，则将非常希望最近版本）。搜索引擎 125 可以通过分析用户选择搜索结果中的哪些文档，来学习哪些查询最近变化最重要。更具体地说，搜索引擎 125 可以考虑用户有多经常喜欢等级低于搜索结果中的较早文档的最新文档。另外，如果随时间流逝，特定文档被包括在最关注的查询（例如“世界职业棒球大赛”）对更特定的查询（例如“纽约美国人”）中，那么，该基于查询的因素 – 通过自身或通过在此提到的其他 – 可以用来降低似乎过期的文档的分值。

[0065] 在一些情况下，可以比更新文档更优先考虑过期文档。因此，当生成用于所述文档的分值时，搜索引擎 125 可以考虑随时间选择该文档的程度。例如，如果对指定查询，用户随时间倾向于选择比更高等级的更新文档更低等级、相对过期的文档，则这由搜索引擎 125 用作调整过期文档的分值的指示。

[0066] 另一基于查询的因素可以涉及文档出现在不同查询结果中的程度。换句话说，可以监视用于一个或多个文档的查询熵，并用作用于计分的基础。例如，如果特定文档作为用于不一致查询集的命中而出现，这可以（尽管不一定）看作所述文档是垃圾邮件的信号，在这种情况下，搜索引擎 125 可以相对更低地计分所述文档。

[0067] 总的来说，搜索引擎 125 可以至少部分基于一个或多个基于查询的因素，来生成（或修改）与文档有关的分值。

[0068] 基于链接的标准

[0069] 根据与本发明的原理相符的实现，可使用一个或多个基于链接的因素来生成（或修改）与文档有关的分值。在一种实现中，基于链接的因素可以涉及新链接出现于文档以及现有链接消失的日期。链接的出现日期可以是搜索引擎 125 找到链接的第一日期或文档包含链接的日期（例如，通过链接找到文档的日期或最近更新它的日期）。链接的消失日期可以是包含该链接的文档删除该链接或本身消失的第一日期。

[0070] 这些日期可以由搜索引擎 125 在扒或索引更新操作期间确定。将该日期作为参考，然后，搜索引擎 125 可以监视到文档的链接的时间变化行为，诸如当链接出现或消失时，链接随时间出现或消失的速率、在指定时间周期期间多少链接出现或消失、存在倾向出现新链接还是文档的现有链接消失等等。

[0071] 使用到和 / 或来自文档的链接的时间变化行为，搜索引擎 125 可以相应地计分文档。例如，随时间新链接数量或速率下降趋势（例如基于最近时间周期对较早时间周期中新链接的数量或速率的比较）能信号告知搜索引擎 125 文档是过期的，在这种情况下，搜索引擎 125 可以减少文档的分值。相反地，根据特定情况和实现，向上趋势会信号告知可以被视为更相关的“最新”文档（例如最新创建或更新其内容的文档）。

[0072] 通过分析文档（或页面）的后向链接随时间增加 / 减少的数量或速率的变化，搜索引擎 125 可以导出文档有多新的重要信号。例如，如果这种分析用逐渐下滑的曲线反映，这可以发信号告知文档是过期的（例如不再更新、重要性降低、由另一文档代替等等）。

[0073] 根据一种实现，分析可以取决于文档的新链接的数量。例如，搜索引擎 125 可以监视自首次找到文档以来新链接的数量相比于最近 n 天中文档的新链接的数量。另外，搜索引擎 125 可以确定与找到的第一链接的寿命相比，最新 y% 链接的最早寿命。

[0074] 为示例目的，假定 y = 10 和 100 天前首次发现两个文档（在该例子中为网站）。对于第一网站，发现 10% 的链接少于 10 天前，而对于第二网站，发现 0% 的链接少于 10 天

前（换句话说，更早地发现它们）。在这种情况下，量度导致对网站 A 为 0.1 以及对网站 B 为 0。可以适当地放大度量。在另一示例性实现中，可以通过执行链接日期分布的相对更详细的分析来修改度量。例如，可以构建模型，预测特定分布是否表示特定类型的网站（例如不再更新、流行增加或减少、取代等等的网站）。

[0075] 根据另一实现，分析可以取决于分配给链接的权重。在这种情况下，每个链接可以由随链接的新鲜度而增加的函数来加权。可以由链接的出现 / 改变的日期、与该链接有关的锚文本的出现 / 改变的日期、包含该链接的文档的出现 / 改变日期来确定链接的新鲜度。基于如果链接仍然相关且良好，则当文档更新时良好链接不变的理论，包含链接的文档的出现 / 改变日期可以是链接的新鲜度的更好指示。为了不由文档的细微不相关部分的微小编辑而更新每个链接的新鲜度，可以测试每个更新文档的显著变化（例如文档的更大部分的变化或文档的许多不同部分的改变），并相应地更新（或不更新）链接的新鲜度。

[0076] 可以用其他方式来加权链接。例如，可以基于有多信任包含链接的文档（例如政府文档可以给予较高信任）来加权链接。链接也可以基于包含链接的文档有多少权威性（例如以类似于在 U. S. 专利 No. 6, 285, 999 中所述的方式来确定权威文档）来加权。链接也可以使用确定新鲜度的一些其他特征，基于包含该链接的文档的新鲜度来加权（例如频繁更新的文档（例如 Yahoo 主页）突然删除到文档的链接）。

[0077] 搜索引擎 125 可以提高或降低存在到其的链接的文档的分值作为指向文档的链接的加权和的函数。该技术可以递归采用。例如，假定文档 S 有 2 年。如果到 S 的链接的 n% 是新的或如果包含到 S 的前向链接的文档被视为新的，则将文档 S 视为新。可以通过使用文档的创建日期并递归地应用该技术来校验后者。

[0078] 根据另一技术，分析可以取决于与指向文档的链接有关的寿命分布。换句话说，可以确定创建到文档的链接的日期并输入到确定寿命分布的函数中。可以假定过期文档的寿命分布将非常不同于新文档的寿命分布。因此，搜索引擎 125 可以部分基于与文档有关的寿命分布来计分文档。

[0079] 链接出现的日期也可以被用来检测“垃圾邮件”，其中，文档的所有者或他们的同僚为提高由搜索引擎分配的分值的目的而创建到他们自己的文档的链接。典型的“合理”文档缓慢地吸引后向链接。后向链接数量的大峰值会信号告知关注现象（例如 CDC 网站在爆发诸如 SARS 后，会迅速地发展许多链接），或通过交换链接、购买链接或获得来自文档的链接，而没有有关生成链接的编辑判断，信号尝试向搜索引擎发送垃圾邮件（以便获得较高等级，从而获得搜索结果中的更好位置）。提供链接而没有编辑判断的文档的例子包括访客薄、参考日志和允许任何人增加文档链接的“免费”页。

[0080] 根据另一实现，分析可以取决于链接消失的日期。许多链接消失能表示这些链接所指向的文档过期（例如不再更新或已经由另一文档替代）。例如，搜索引擎 125 可以监视到文档的一个或多个链接消失的日期、在指定时间窗口中消失的链接数，或到文档的链接数（或到包含这些链接的文档的链接 / 更新）的一些其他时间变化减少，来识别可被视为过期的文档。一旦已经确定文档过期，当确定由链接指向的文档的分值时，包含在那个文档中的链接可以由搜索引擎 125 忽视或忽略。

[0081] 根据另一实现，分析可以不仅取决于文档的链接的寿命，而且可以取决于链接的动态化。如此，搜索引擎 125 可以加权除具有非常新的链接外，每天具有不同于（例如降

低)始终更新并始终链接到指定目标文档的文档的不同特征链接的文档。在一个示例性实现中,搜索引擎 125 可以基于在时间窗内,对于所有版本文档,具有到一个文档的链接的各文档的分值,来生成用于该文档的分值。该另一版本可以基于文档的主要更新时间,将减少 / 衰减因子包含在集成中。

[0082] 总的来说,搜索引擎 125 可以部分基于一个或多个基于链接的因素,来生成(或修改)与文档有关的分值。

[0083] 锚文本

[0084] 根据与本发明的原理相符的实现,与锚文本随时间改变的方式有关的信息可以用来生成(或修改)与文档有关的分值。例如,可以将与到文档的链接有关的锚文本随时间的改变用作文档中已经有更新或甚至焦点改变的表示。

[0085] 另外,如果文档的内容改变,使得它显著地不同于与其后向链接有关的锚文本,那么与文档有关的域可以显著地(完全)从前身改变。当域届满和不同方购买该域时这会发生。因为锚文本通常被视为是其相关链接所指向的文档的一部分,域可以在用于查询的搜索结果中不再在标题上显现。这是不期望的结果。

[0086] 解决该问题的一个方法是估计域改变其焦点的日期。这可以通过确定文档的文本显著改变或锚文本的文本显著改变的日期来完成。然后可以忽略或忽视在那一日期前的所有链接和 / 或锚文本。

[0087] 锚文本的新鲜度也可以被用作计分文档的因素。可以通过例如锚文本的出现 / 改变日期、与锚文本有关的链接的出现 / 改变日期和 / 或相关链接所指向的文档的出现 / 改变日期,来确定锚文本的新鲜度。基于如果锚文本仍然相关且良好,则当文档更新时良好锚文本不变的理论,由链接指向的文档的出现 / 改变日期可以是锚文本的新鲜度的良好指示符。为了不由文档的细微不相关部分的细微编辑而更新锚文本的新鲜度,可以测试每个更新文档的显著变化(例如文档的大部分改变或文档的许多不同部分的改变)并相应地更新(或不更新)锚文本的新鲜度。

[0088] 总的来说,搜索引擎 125 可以至少部分基于与锚文本随时间改变的方式有关的信息,来生成(或修改)与文档有关的分值。

[0089] 通信量

[0090] 根据与本发明的原理相符的实现,有关与文档有关的通信量随时间的信息可以用来生成(或修改)与文档有关的分值。例如,搜索引擎 125 可以监视一个或多个用户到文档的通信量或其他“用途”的时间变化特性。通信量的大的降低可以表示文档为过期(例如不再更新或可能由另一文档替代)。

[0091] 在一种实现中,搜索引擎 125 可以比较最近 j 天(例如其中 $j = 30$) 文档的平均通信量与文档接收最多通信量,可选地,按季节变化调整的月期间,或最近 k 天(例如 $k = 365$) 期间的平均通信量。可选地,搜索引擎 125 可以识别重复通信量模式或通信量模式随时间的变化。可以发现存在文档或多或少流行(例如具有或多或少通信量)的周期,诸如在夏季月期间,周末或在一些其他季节时间周期期间。通过识别重复通信量模式或通信量模式的变化,搜索引擎 125 可以适当地调整在这些周期期间或之外文档的得分。

[0092] 另外,或者,搜索引擎 125 可以监视与用于特定文档的“广告通信量”有关的时间变化特性。例如,搜索引擎 125 可以监视下述因素的一个或多个组合:(1) 随时间,由指定

文档呈现或更新广告的程度或频率；(2) 广告商的质量（例如其广告参考 / 链接到搜索引擎 125 知道随时间具有相对高通信量和信任的文档，诸如 amazon.com 的文档可以被提供比其广告指向低通信量 / 不可靠文档的那些文档，诸如色情网站相对更高的权重）；以及 (3) 广告生成到它们所涉及的文档的用户通信量的程度（例如它们的点击率）。搜索引擎 125 可以使用与广告通信量有关的这些时间变化特性来计分文档。

[0093] 总的来说，搜索引擎 125 可以至少部分基于有关与文档有关的通信量随时间的信息，来生成（或修改）与文档有关的分值。

[0094] 用户行为

[0095] 根据与本发明的原理相符的实现，可以使用对应于随时间与文档有关的个人或集体用户行为的信息，来生成（或修改）与文档有关的分值。例如搜索引擎 125 可以监视从搜索结果集中选择一个文档的次数和 / 或一个或多个用户访问所述文档所花费的时间量。然后，搜索引擎 125 可以至少部分基于该信息来计分所述文档。

[0096] 如果对某一查询返回文档，以及给定相同或类似查询，随时间或在指定时间窗口内，用户在该文档上平均花费或多或少的时间，那么这可以分别被用作该文档新或旧的表示。例如假定查询“Riverview 游泳计划”返回具有标题“Riverview 游泳计划”的文档。进一步假定用户以前花费 30 秒访问它，但现在选择所述文档的每个用户仅花费几秒来访问它。搜索引擎 125 可以使用该信息来确定所述文档为旧（即包含过时游泳计划）并相应地计分所述文档。

[0097] 总的来说，搜索引擎 125 可以至少部分基于与随时间与文档有关的个人或集体用户行为相应的信息，来生成（或修改）与文档有关的分值。

[0098] 域相关信息

[0099] 根据与本发明的原理相符的实现，涉及与文档有关的域的信息可以用来生成（或修改）与所述文档有关的分值。例如，搜索引擎 125 可以监视与在计算机网络（例如互联网、内联网或其他网络或文档数据库）内如何寄存文档有关的信息，并使用该信息来计分文档。

[0100] 尝试欺骗（发送垃圾邮件）搜索引擎的个人通常使用用完即扔或“门口（doorway）”域，并尝试在被抓住前获得尽可能多的通信量。当计分与这些域有关的文档时，关于域的合法性的信息可以由搜索引擎 125 使用。

[0101] 可以使用某些信号来区分非法域和合法域。例如域可以续达 10 年的周期。有用（合法）域通常预先支付几年，而门口（非法）域仅使用 1 年多。因此，当未来域届满时的日期能被用作预测域的合法性，从而预测与之有关的文档的合法性的因素。

[0102] 同样，或者，用于域的域名服务器（DNS）记录可以被监视以预测域是否合法。DNS 记录包含谁注册了域、行政和技术地址以及名称服务器（即将域名解析为 IP 地址的服务器）的地址的详情。通过分析用于域的随时间的该数据，可以识别非法域。例如，搜索引擎 125 可以监视在时间周期上，物理正确的地址信息是否存在，域的联系信息是否相对频繁地改变，在不同名称服务器和寄主公司之间是否存在相当大量的变化等等。在一个实现中，可以识别、存储已知不良联系信息、名称服务器和 / 或 IP 地址的清单，并用于预测域的合法性，从而预测与之相关的文档的合法性。

[0103] 同样，另外，关于与域有关的名称服务器的寿命或其他信息可以用来预测域的合

法性。“良好”名称服务器可以具有来自不同注册器的不同域的混合并具有寄主这些域的历史,而“不良”名称服务器会主要寄主色情或门口域、具有商业词汇的域(垃圾邮件的通用指示符)或主要来自单个注册器的零散域或可能是全新的。名称服务器的新鲜度可以非自动地为确定相关域的合法性的消极因素,而可以结合其他因素,诸如在此所述的。

[0104] 总的来说,搜索引擎 125 可以至少部分基于有关与文档有关的域的合法性的信息,来生成(或修改)与文档有关的分值。

[0105] 等级历史

[0106] 根据与本发明的原理相符的实现,可以使用与文档的先前等级有关的信息来生成(或修改)与文档有关的分值。例如,搜索引擎 125 可以响应提供给搜索引擎 125 的搜索查询,监视文档的时间变化等级。搜索引擎 125 可以确定在许多查询上等级跳跃的文档可能是主题文档,或它可能是发信号试图向搜索引擎 125 发送垃圾邮件。

[0107] 因此,可以使用在时间周期上文档在等级方面移动的数量或速率来影响分配给那个文档的未来分值。在一种实现中,对于搜索结果的每个集合,可以根据它在前 N 个搜索结果中的位置来加权文档。对 N = 30,一个示例函数可以是 $\lceil ((N+1)-SLOT/N) \rceil^4$ 。在这种情况下,第一结果可得到 1.0 的分值,对第 N 个结果,下降到接近 0 的分值。

[0108] 可以重复查询集(例如商业查询),以及可以标记获得等级多于 M% 的文档,或等级的百分比增长被用作确定用于所述文档的分值的信号。例如,如果前面结果的平均(中等)分值相对高以及前面结果逐月存在相当大的变化,则搜索引擎 125 可以确定查询很可能是商业的。搜索引擎 125 也可以监视流入流出(churn)作为商业查询的指示。对商业查询,垃圾邮件的可能性较高,因此,搜索引擎 125 可以相应地处理与之有关的文档。

[0109] 除用于指定查询的文档的位置(或等级)的历史外,搜索引擎 125 可以监视(在页面、主机、文档和 / 或域基础上)一个或多个其他因素,诸如随时间将文档选择为搜索结果的查询数以及速率(增加 / 减少)、季节性、突发性和随时间文档被选择为搜索结果的其他模式和 / 或对于 URL 查询对,分值随时间的变化。

[0110] 另外,或者,搜索引擎 125 可以监视随时间,与基于查询的标准无关的文档(例如 URL)数量。例如,搜索引擎 125 可以监视响应于指定查询或查询集而生成的顶端结果集中的平均分值,并调整响应于指定查询或查询集而生成的结果集和 / 或其他结果的分值。此外,搜索引擎 125 可以监视随时间,为特定查询或查询集生成的结果数。如果搜索引擎 125 确定结果数增加或增长率有变化(例如这种增加可以是“热门主题”或其他现象的表示),搜索引擎 125 可以使那些结果在未来计分更高。

[0111] 另外,或者,搜索引擎 125 可以监视随时间的文档等级来检测文档等级中的突然峰值。峰值可以表示主题现象(例如热门主题)或试图通过例如交易或购买链接而向搜索引擎 125 发送垃圾邮件。搜索引擎 125 可以通过利用滞后来允许以某一速率增长等级,采用防止垃圾邮件尝试的措施。在另一实现中,指定文档的等级可以被允许在预定时间窗上增长的某一最大阈值。作为将与主题现象有关的文档与垃圾邮件文档区分的进一步措施,搜索引擎 125 可以基于例如在新闻中将不会提到垃圾邮件文档的理论,考虑在新闻文章、论述组等等中文档的记载。可以使用这些技术的任何一个或组合来减少垃圾邮件尝试。

[0112] 搜索引擎 125 也可以把在一些方面中被确定为权威的文档,诸如政府文档、web 目录(例如 Yahoo)以及随时间已经显示出相对稳定和高等级的文档作为例外。例如,如果到

权威文档的链接的数量或增加率中出现不寻常峰值,那么搜索引擎 125 可以认为所述文档不是垃圾邮件,从而允许相当高或甚至(随时间)对其等级(增长)无阈值。

[0113] 另外,或者,搜索引擎 125 可以将文档等级的显著下降视为这些文档“不受喜欢”或过期的指示。例如,如果文档的等级随时间显著地下降,那么搜索引擎 125 可以将所述文档视为过期并相应地计分所述文档。

[0114] 总的来说,搜索引擎 125 可以至少部分基于与文档的先前等级有关的信息,来生成(或修改)与文档有关的分值。

[0115] 用户维护 / 生成的数据

[0116] 根据与本发明的原理相符的实现,可以使用用户维护或生成的数据来生成(或修改)与文档有关的分值。例如,搜索引擎 125 可以监视由用户维护或生成的数据,诸如“书签”、“喜好”或可以提供用户喜欢或感兴趣的文档的一些指示的其他类型的数据。搜索引擎 125 可以直接(例如经浏览器辅助)或间接(例如经浏览器)获得该数据。然后,搜索引擎 125 随时间分析文档与之有关的多个书签 / 喜好来确定文档的重要性。

[0117] 搜索引擎 125 还可以分析从书签 / 喜好列表增加或移出文档(或更具体地说,文档的路径),增加到书签 / 喜好列表或从其移出文档的速率和 / 或是否增加、删除或通过书签 / 喜好列表访问文档的向上和向下趋势。如果多个用户正将特定文档增加到他们的书签 / 喜好列表中,或通常随时间通过这些列表访问该文档,这可以被视为该文档相对重要的指示。另一方面,如果多个用户正减少访问在他们的书签 / 喜好列表中指示的文档,或正从他们的列表日益删除 / 替代到所述文档的路径,这可以被看作该文档过时、不流行等等的指示。因此,搜索引擎 125 可以相应地计分所述文档。

[0118] 在另一实现中,可以表示用户随时间对特定文档的兴趣增加或减少的其他类型的用户数据可以由搜索引擎 125 使用来计分文档。例如,与用户有关的“临时”或缓冲文件能由搜索引擎 125 监视,以识别随时间添加的文档增加还是减少。类似地,与特定文档有关的 cookie 数据块也可以由搜索引擎 125 监视来确定对文档的兴趣存在向上还是向下趋势。

[0119] 总的来说,搜索引擎 125 可以至少部分基于用户维护或生成的数据,来生成(或修改)与文档有关的分值。

[0120] 锚文本中的唯一字、二元语法(bigram)、短语

[0121] 根据与本发明的原理相符的实现,可以使用关于锚文本中的唯一字、二元语法、短语的信息来生成(或修改)与文档有关的分值。例如搜索引擎 125 可以监视随时间的网站(或链接)图以及它们的行为,并将该信息用于计分、垃圾邮件检测或其他目的。自然开发的网站图通常包含独立的判断。通常表示垃圾邮件意图的合成生成的网站图是基于协调判断,引起锚字 / 二元语法 / 短语的增长图可能相对尖。

[0122] 这种尖峰的一种原因可以是增加了来自许多文档的大量相同锚。另一可能性是增加了来自多个文档的故意不同的锚。搜索引擎 125 可以监视锚并将它们作为计分它们的相关链接所指向的文档的因素。例如,搜索引擎 125 可以改进可疑锚对相关文档分值的影响。另外,搜索引擎 125 可以使用合成生成的似然度的连续换算并导出乘法因子来换算用于所述文档的分值。

[0123] 总的来说,搜索引擎 125 可以至少部分基于关于与指向文档的一个或多个链接有关的锚文本中的唯一字、二元语法和短语的信息,来生成(或修改)与文档有关的分值。

[0124] 独立对等体 (peer) 的连接

[0125] 根据与本发明的原理相符的实现,可以使用关于独立对等体 (例如无关文档) 的连接的信息来生成 (或修改) 与文档有关的分值。

[0126] 具有到各文档的大量链接的明显独立对等体 - 输入和 / 或输出数量的突然增长可以表示潜在虚假网站图,其是试图发送垃圾邮件的指示符。如果增长对应于通常相干或不一致的锚文本,则可以增强该指示。当与基于链接的计分技术一起使用时,能使用该信息来降级这些链接的影响,作为二进制判断项 (例如将分值降级固定量) 或乘法因子。

[0127] 总的来说,搜索引擎 125 可以至少部分基于关于独立对等体的连接的信息,来生成 (或修改) 与文档有关的分值。

[0128] 文档主题

[0129] 根据与本发明的原理相符的实现,可以使用有关文档主题的信息来生成 (或修改) 与文档有关的分值。例如,搜索引擎 125 可以执行主题提取 (例如通过分目录、URL 分析、内容分析、群集、概括、唯一低频字集或一些其他类型的主题提取)。然后,搜索引擎 125 可以监视随时间文档的主题并将该信息用于计分目的。

[0130] 与文档有关的主题集随时间的显著变化可以表示文档已经改变所有者和先前文档指示符,诸如分值、锚文本等等不再可靠。类似地,主题数目中的峰值能表示垃圾邮件。例如,如果特定文档与可以视为“稳定的”时间周期上的一个或多个主题集有关,然后与所述文档有关的主题数目中出现 (突然) 峰值,则这可以是文档已经被取代为“门口”文档的指示。另一指示可以包括与文档有关的初始主题的消失。如果检测到一个或多个这些情形,那么,搜索引擎 125 可以降低这些文档和 / 或链接、锚文本或与所述文档有关的其他数据的相对分值。

[0131] 总的来说,搜索引擎 125 可以至少部分基于与所述文档有关的一个或多个主题的变化,来生成 (或修改) 与文档有关的分值。

[0132] 示例性处理

[0133] 图 4 是根据与本发明的原理相符的实现,用于计分文档的示例性处理的流程图。处理可以从服务器 120 识别文档 (动作 410) 开始。文档可以包括例如与搜索查询有关的一个或多个文档,诸如识别为与搜索查询有关的文档。另外,文档可以包括与任何搜索查询无关的文档资料库或库中的一个或多个文档 (例如通过扒网络而识别并存储在库中的文档)。

[0134] 搜索引擎 125 可以获得与所识别的文档有关的历史数据 (动作 420)。如上所述,历史数据可以采用不同形式。例如,历史数据可以包括与文档初始日期有关的数据;文档内容更新 / 改变;查询分析;基于链接的标准;锚文本;通信量;用户行为;域相关信息;等级历史;用户维护 / 生成的数据 (例如书签和 / 或喜好);锚文本中的唯一字、二元语法和短语;独立对等体的连接和 / 或文档主题。搜索引擎 125 可以获得这些类型的历史数据中的一个或组合。

[0135] 然后,搜索引擎 125 可以至少部分基于历史数据来计分所识别的文档 (动作 430)。当所识别的文档与搜索查询有关时,搜索引擎 125 可以例如基于它们与搜索查询有多相关,来生成用于所述文档的相关分值。然后,搜索引擎 125 可以将历史分值与相关分值组合来获得用于所述文档的总分值。代替组合分值,搜索引擎 125 可以基于历史数据来修改用

于所述文档的相关分值，从而提高或降低分值，或在一些情况下，使分值相同。另外，搜索引擎 125 可以基于历史数据来计分文档，而不生成相关分值。在任一情况下，搜索引擎 125 可以使用历史数据类型的一个或组合来计分文档。

[0136] 当所识别的文档与搜索查询有关时，搜索引擎 125 也可以由计分文档来形成搜索结果。例如，搜索引擎 125 可以基于它们的分值来排序文档。然后，搜索引擎 125 可以形成对这些文档的参考，其中，参考可以包括文档的标题（可以包含当选择时，将用户引导到该真正文档的超级链接）以及来自文档的片断（例如文本摘录）。在其他实现中，可以不同地形成参考。搜索引擎 125 可以将对应于多个高计分文档的参考（例如预定多个文档，具有超出阈值分值的文档，所有文档等等）呈现给提交搜索查询的用户。

[0137] 结论

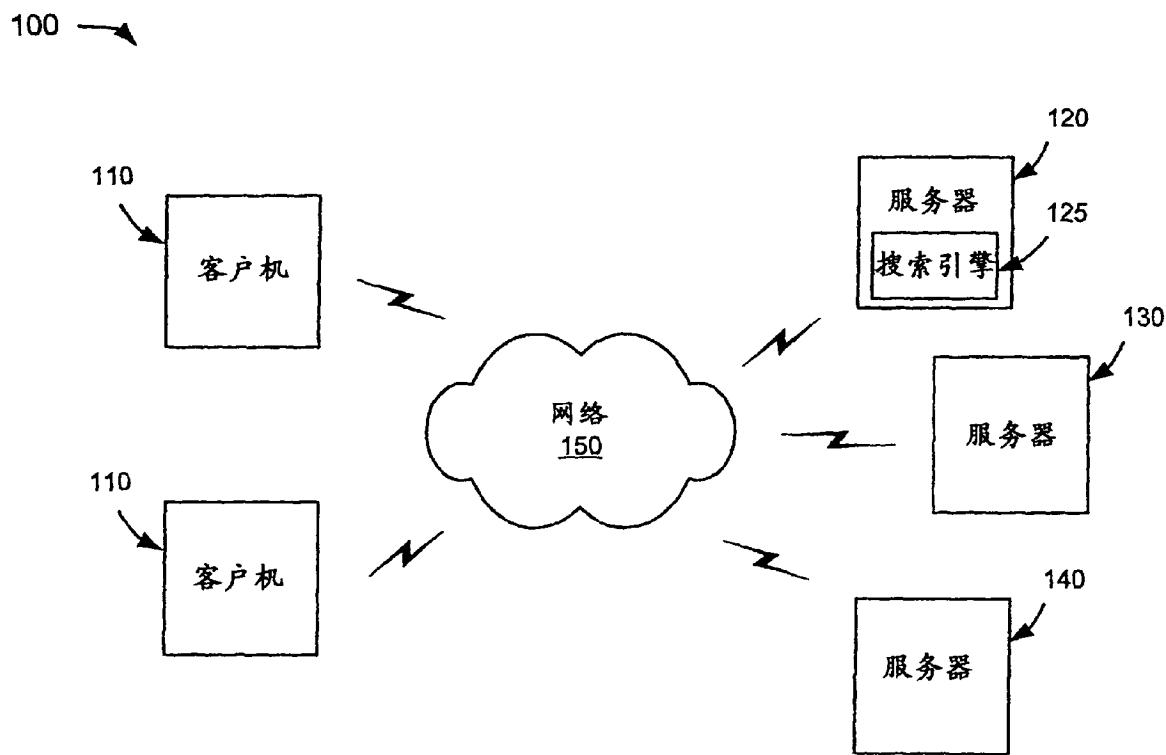
[0138] 与本发明的原理相符的系统和方法可以使用历史数据来计分文档并形成高质量搜索结果。

[0139] 本发明的优选实施例的上述描述提供示例和描述，但不打算排除或将本发明限制到所公开的具体形式。鉴于上述进行教导，修改和改进是可能的，或可以从实施本发明获得。例如，尽管参考图 4 描述了一系列动作，但在与本发明的原理相符的其他实现中，可以修改动作顺序。同时，可以并行执行不相关动作。

[0140] 另外，通常描述服务器 120 来执行参考图 4 的处理描述的大部分动作，如果不是全部的话。在与本发明的原理相符的另一实现中，可以由另一实体，诸如另一服务器 130 和 / 或 140 或客户机 110 来执行一个或多个或所有动作。

[0141] 对本领域的普通技术人员来说，如上所述的本发明的方面可以在图中所示的实现中的软件、固件和硬件的许多不同形式实现是显而易见的。用来实现与本发明的原理相符的方面的真正软件代码或专用控制硬件不是本发明的限制。因此，在不参考特定软件代码的情况下，描述这些方面的操作和行为，应理解到本领域的一个普通技术将能基于在此的说明，设计实现这些方面的软件和控制硬件。

图 1



110-140 →

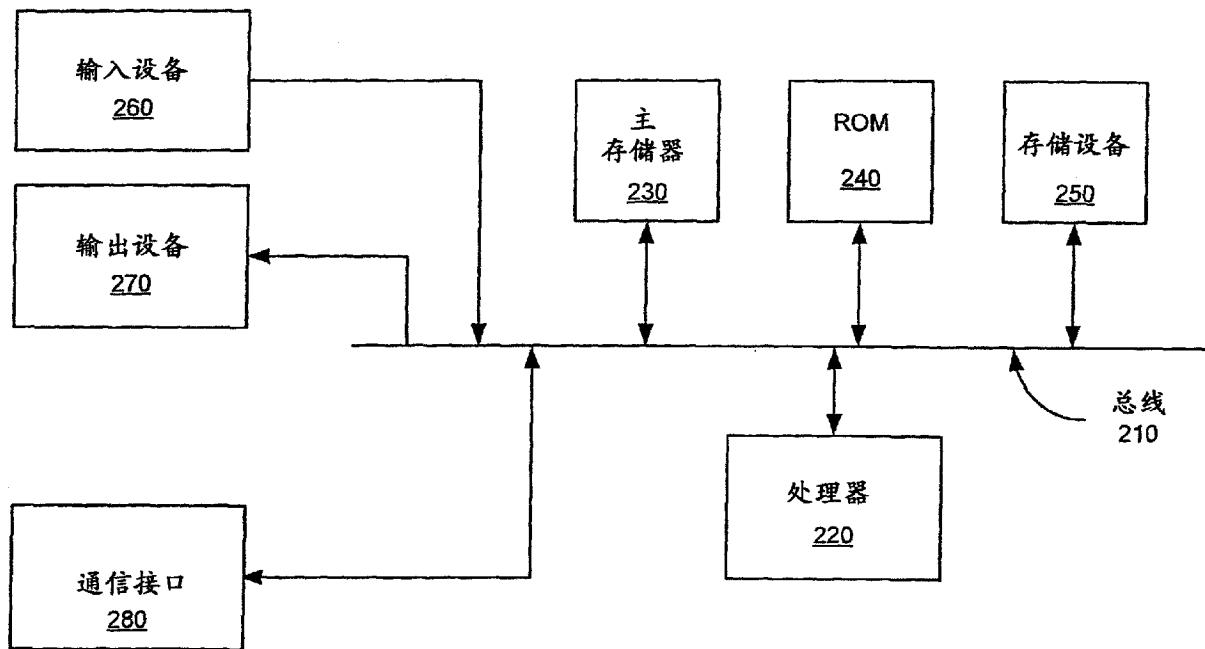


图 2

图 3

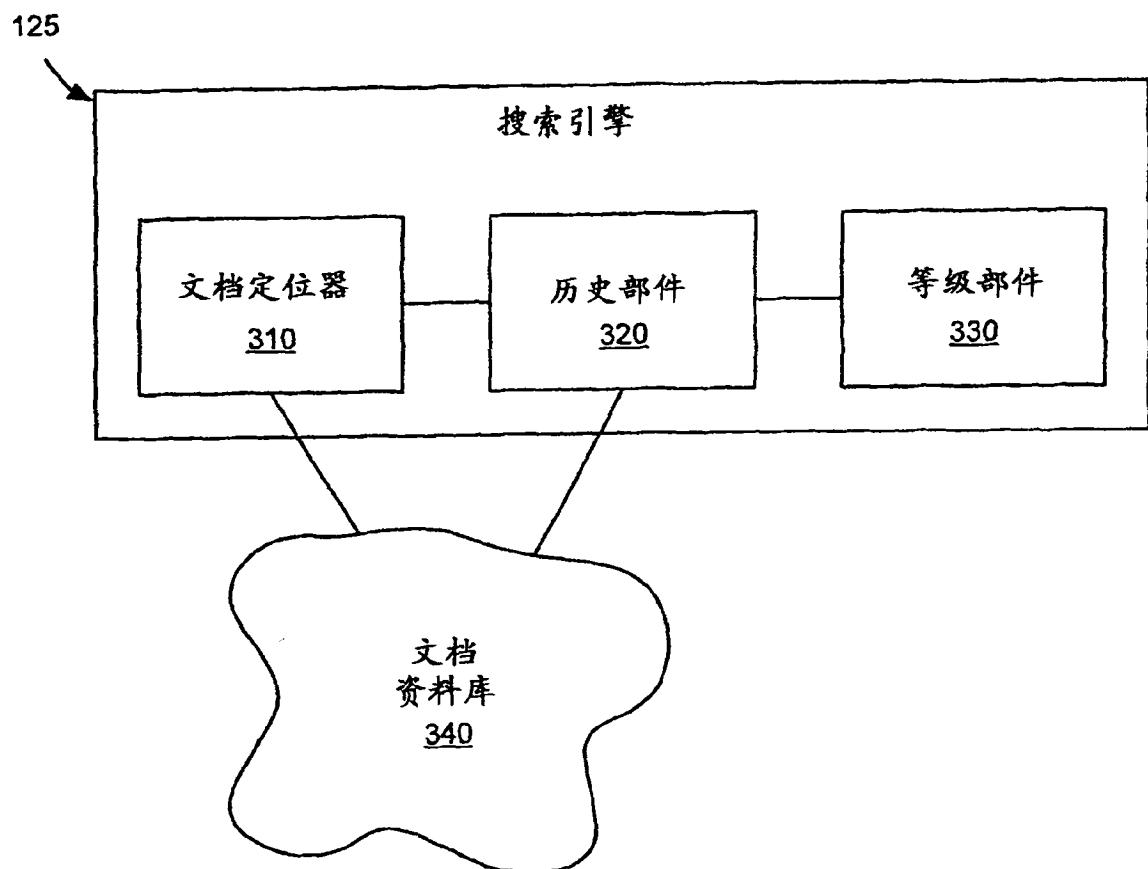


图 4

