

[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 17/30 (2006.01)



[12] 发明专利说明书

专利号 ZL 200510089647.7

[45] 授权公告日 2009 年 12 月 23 日

[11] 授权公告号 CN 100573513C

[22] 申请日 2005.4.30

[21] 申请号 200510089647.7

[30] 优先权

[32] 2004.4.30 [33] US [31] 10/837,540

[73] 专利权人 微软公司

地址 美国华盛顿州

[72] 发明人 B·章 H-J·曾 马维英
陈 正

[56] 参考文献

WO03012684A2 2003.2.13

WO2004025490A1 2004.3.25

Multi - document summarization by sentence
extraction. Goldstein, J et all. [http://acl.ldc.
upenn.edu/W/W00/W00.0405](http://acl.ldc.upenn.edu/W/W00/W00.0405). 2000

审查员 谭李丽

[74] 专利代理机构 上海专利商标事务所有限公司

代理人 李 玲

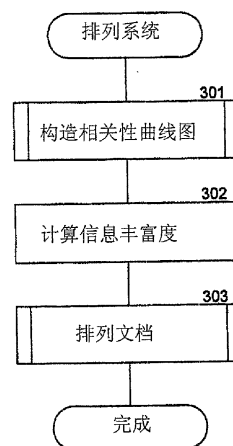
权利要求书 4 页 说明书 10 页 附图 5 页

[54] 发明名称

用于排列搜索结果的文档以改进多样性和信息丰富度的方法和系统

[57] 摘要

一种基于主题的信息丰富度和多样性来排列搜索结果的文档的方法和系统。该排列系统决定在搜索结果中的每一个文档的信息丰富度。该排列系统基于它们的关联性而将搜索结果的文档分组，意味着它们被指向相似的主题。该排列系统将文档排序以保证最高排列文档可以包含覆盖每一个主题的至少一篇文档，那就是说，来自每一个组的一篇文档。该排列系统从在该组中具有最高信息丰富度的文档的每一组中选择文档。当这些文档以某个排列顺序提供给用户时，用户将在搜索结果的第一页中发现覆盖各种类型的主题的主题，而不仅仅是单一的受欢迎主题。



1. 一种在计算机系统中用于排列一个搜索结果的文档的方法，该方法包括：

为该搜索结果的每一篇文档，基于用于该文档的信息丰富度初始化一个相关性排列，其中所述相关性衡量一篇文档中的信息被包含在另一篇文档的信息内容中的程度，并且其中所述信息丰富度是衡量一篇文档包含有多少与它的主题相关的信息的尺度；和

对于每一组相似的文档，调整该组中的文档的相关性排列以使除具有最高相关性排列的文档之外的文档的相关性排列低于在该组中的文档的最高初始的相关性排列。

2. 如权利要求 1 所述的方法，其中，用于该组中的文档的相关性排列的调整包括：减少该组中的每一篇文档的相关性排列，除了在该组中具有最高相关性排列的文档的相关性排列。

3. 如权利要求 2 所述的方法，其中相比于与具有最高相关性排列的文档较少相似的文档，与具有最高相关性排列的文档较多相似的一篇文档具有被减少更多的相关性排列。

4. 如权利要求 1 所述的方法，其中用于该组中的文档的相关性排列的调整包括：从该组中移走该具有最高相关性排列的文档，并减少该组中剩余的文档的相关性排列，其中文档的移走顺序代表了该搜索结果的文档的排列。

5. 如权利要求 1 所述的方法，包括用于每一篇文档的，基于该已调整的相关性排列和一个基于搜索的相关性来计算文档的一个相关性。

6. 一种在计算机系统中用来排序一个搜索结果的文档以增加高排序文档的主题的多样性的方法，该方法包括：

接收搜索请求；

搜索满足所述搜索请求的文档；

将满足所述搜索请求的某些文档指定为搜索结果；

在所述搜索结果被指定后，基于表明文档对之间的相关性的相关性图来识别所述搜索结果的相似的文档的组；

从已识别的每一组中选择一篇文档，其中从每一组中选择出的文档具有在该组中的文档的最高信息丰富度，其中所述信息丰富度是衡量一篇文档包含有

多少与它的主题相关的信息的尺度；和

将所选择的文档排列在所述搜索结果的其它文档之上；

在一有形的介质中存储所选的文档的排列；

其中所述相关性衡量一篇文档中的信息被包含在另一篇文档的信息内容中的程度。

7. 如权利要求 6 所述的方法，其中每一篇文档有一个初始化排列，且来自每个已识别的组中的该选择的文档是具有最高初始化排列的文档。

8. 如权利要求 6 所述的方法，包括基于没有被选择的文档与该组的已选择的文档的相似性再排列该组中所述没有被选择的文档。

9. 如权利要求 8 所述的方法，其中该再排列给予在该组中，与该组中的已选择的文档最相似的还没有选择的文档在该组文档的排列上的最大减少。

10. 如权利要求 9 所述的方法，其中该组中还没有被选择的文档根据它们的再排列而被排列。

11. 如权利要求 9 所述的方法，包括在再排列之后从已被识别的组中的每一组中选择一篇文档，且将那些文档排列在还没有被选择的其他文档之上。

12. 如权利要求 8 所述的方法，其中该再排列应用一个相似性惩罚。

13. 如权利要求 6 所述的方法，其中该组是利用一个相关性曲线图来识别的。

14. 一种在计算机系统中用于计算一个文档的集合中的一篇文档的信息丰富度的方法，该方法包括：

识别在集合中的每一篇文档与该文档的相关性，其中所述相关性衡量一篇文档中的信息被包含在另一篇文档的信息内容中的程度，并且其中信息丰富度是衡量一篇文档包含有多少与它的主题相关的信息的尺度；和

基于在该集合中其他的文档与该文档的相关性决定该文档的信息丰富度，其中的信息丰富度被定义为：

$$InfoRich(d_i) = \sum_{\text{所有 } j \neq i} InfoRich(d_j) \cdot \tilde{M}_{ji},$$

其中 $InfoRich(d_i)$ 是文档 d_i 的信息丰富度， $\tilde{M}_{ij}$ 是规格化矩阵的一个元素。

15. 如权利要求 14 所述的方法，其中每一篇文档的相关性的识别包括产生一个相关性曲线图。

16. 如权利要求 14 所述的方法，其中相关性是衡量一篇文档中的信息内

容被包含在另一篇文档中的程度。

17. 如权利要求 14 所述的方法, 其中相关性被定义为:

$$aff(d_i, d_j) = \frac{\vec{d}_i \cdot \vec{d}_j}{\|\vec{d}_i\|}$$

其中 $aff(d_i, d_j)$ 是文档 d_i 对于文档 d_j 的相关性, $\vec{d}_i$ 代表文档 d_i 的向量, $\vec{d}_j$ 代表文档 d_j 的向量, $\|\vec{d}_i\|$ 代表文档向量 $\vec{d}_i$ 的长度。

18. 如权利要求 14 所述的方法, 其中信息丰富度是衡量一篇文档中的信息内容包含其它文档的信息内容的程度。

19. 一种用于将搜索结果中的文档排序的计算机系统, 包括:

用于对于每一篇文档, 基于文档的信息丰富度初始化一个相关性排列的装置, 其中所述相关性衡量一篇文档中的信息被包含在另一篇文档的信息内容中的程度, 并且其中信息丰富度是衡量一篇文档包含有多少与它的主题相关的信息的尺度; 和

用于当一篇文档具有一个高的相关性排列时, 减少与其相关的文档的相关性排列的装置,

其中每一篇文档的相关性排列代表该文档的排列。

20. 如权利要求 19 所述的计算机系统, 其中一篇文档的信息丰富度是基于每一对文档的相关性而被计算的。

21. 如权利要求 20 所述的计算机系统, 其中信息丰富度被定义为:

$$InfoRich(d_i) = \sum_{\text{所有 } j \neq i} InfoRich(d_j) \cdot \tilde{M}_{ji},$$

其中 $InfoRich(d_i)$ 是文档 d_i 的信息丰富度, $\tilde{M}_{ji}$ 是规格化矩阵的一个元素。

22. 如权利要求 20 所述的计算机系统, 其中, 该相关性被定义为:

$$aff(d_i, d_j) = \frac{\vec{d}_i \cdot \vec{d}_j}{\|\vec{d}_i\|},$$

其中 $aff(d_i, d_j)$ 是文档 d_i 对于文档 d_j 的相关性, $\vec{d}_i$ 代表文档 d_i 的向量, $\vec{d}_j$ 代表文档 d_j 的向量, $\|\vec{d}_i\|$ 代表文档向量 $\vec{d}_i$ 的长度。

23. 如权利要求 19 所述的计算机系统, 其中相比于与具有高相关性排列的该文档较少相似的文档, 与具有高相关性排列的该文档较多相似的文档具有

被减少更多的相关性排列。

24. 如权利要求 19 所述的计算机系统, 包括用于基于文档的相关性排列和用于文档的基于搜索的相关性为每一篇文档计算用于该文档的一个相关性的装置。

25. 一种用于计算存在于一个文档的集合中的一篇文档的信息丰富度的计算机系统, 包括:

识别每一篇在集合中的文档与该文档的相关性的部件, 其中所述相关性衡量一篇文档中的信息被包含在另一篇文档的信息内容中的程度, 并且其中信息丰富度是衡量一篇文档包含有多少与它的主题相关的信息的尺度; 和

基于在该集合中的其他的文档与该文档的相关性确定该文档的信息丰富度的部件, 其中的信息丰富度被定义为:

$$InfoRich(d_i) = \sum_{\text{所有 } j \neq i} InfoRich(d_j) \cdot \tilde{M}_{ji},$$

其中 $InfoRich(d_i)$ 是文档 d_i 的信息丰富度, $\tilde{M}_{ij}$ 是规格化矩阵的一个元素。

26. 如权利要求 25 所述的系统, 其中该用于识别的部件产生一个相关性曲线图。

27. 如权利要求 25 所述的计算机系统, 其中信息丰富度是衡量一篇文档中的信息内容包含其它文档的信息内容的程度。

28. 如权利要求 1 所述的方法, 其中通过产生每对文档的相关性衡量并根据相关性衡量计算文档的固定概率分布, 来确定信息丰富度, 所述固定概率分布是从马尔可夫链得出的, 并且马尔可夫链是从一过程中导出的, 在该过程中一文档的信息丰富度流入与该文档具有相关性的其它文档。

29. 如权利要求 25 所述的计算机系统, 其中确定该文档的信息丰富度的部件根据固定概率分布来进行确定, 固定概率分布从集合中其它文档和该文档的相关性中导出, 所述固定概率分布是从马尔可夫链得出的, 并且马尔可夫链是从一过程中导出的, 在该过程中一文档的信息丰富度流入与该文档具有相关性的其它文档。

用于排列搜索结果的文档以 改进多样性和信息丰富度的方法和系统

技术领域

所述的技术一般涉及由提交给一个搜索引擎装置的一个搜索请求所识别的一个搜索结果的文档的排列。

背景技术

许多搜索引擎装置，例如 Google 和 Overture，提供用来搜索经由 Internet 可以被访问的信息。这些搜索引擎装置允许用户搜索用户关心的显示页，例如 web 页。在用户提交一个包含搜索条件的搜索请求后，该搜索引擎装置识别可能与这些条件相关联的 web 页。为了快速地识别相关的 web 页，该搜索引擎装置可以保持一个 web 页的关键词映射。该映射依靠“爬行”该 web（即，环球信息网）以提取每一个 web 页的关键词来产生。为了爬行该 web，一个搜索引擎装置可以利用根 web 页的列表来识别所有的可以通过这些根 web 页而被访问的 web 页。任何特定 web 页的关键词可以使用各种公知的信息检索技术被提取，例如识别一个标题的词、在 web 页的元数据中所提供的词、突出显示的词，等等。该搜索引擎装置可以计算一个关联性分数，该关联性分数指出每一个 web 页与基于每一个匹配的接近性、web 页普及性（例如，Google 的 PageRank）等等的搜索请求在多大程度上相关联。该搜索引擎装置然后用基于这些 web 页的关联性的一个顺序显示给用户这些 web 页的链接。搜索引擎可能更普遍地提供用于任何文档的集合中的信息的搜索。例如，该文档的集合可以包括所有的美国专利、所有的联邦法庭的意见、一个公司的所有存档文档等等。

由一个基于 web 的搜索引擎装置提供的搜索结果的最高排列的 web 页可能被全部指向相同的受欢迎的话题。例如，如果一个用户利用搜索条件“Spielberg”提出一个搜索请求，然后该搜索结果的最高排列的 web 页将可能与 Steven Spielberg 相关。然而，如果用户对 Steven Spielberg 不感兴趣，而是对定位于一个具有同姓的数学教授的主页感兴趣的话，则该 web 页的排列对用户是没有帮助的。尽管该教授的主页可能被包含在搜索结果中，但该用户仍然需要去浏览

链接于该搜索结果的 web 页的许多页，以定位该教授的主页的链接。通常，当没有被识别为搜索结果的第一页时，对于用户来说定位一个期望的文档是困难的。此外，当用户不得不翻阅多页搜索结果以找到感兴趣的文档时，他们会感到很灰心。

人们会期望一种用于排列文档的技术，它可以提供更多样化的存在于最高排列文档中的主题，人们会更进一步地期望每个这样的最高排列文档具有与它的主题相关的丰富的信息内容。

发明内容

一种基于主题的信息的丰富度和多样性而排列搜索结果的文档的系统。一种排列系统基于它们的关联性而将搜索结果的文档分组，意味着它们被指向类似的主题。该排列系统为文档排序以保证最高排列文档包含覆盖每一个主题的至少一篇文档。该排列系统然后从在该组中具有文档的最高信息丰富度的每一组中选择文档，作为最高排列文档中的一篇。

附图说明

图 1 是说明在一个实施例中的一个相关性曲线图的图表。

图 2 是说明在一个实施例中的排列系统的部件的方块图。

图 3 是说明在一个实施例中的排列系统的全部处理的流程图。

图 4 是说明在一个实施例中的一个构造相关性曲线图部件的处理的流程图。

图 5 是说明在一个实施例中的一个排列文档部件的处理的流程图。

具体实施方式

一种用于基于主题的信息的丰富度和多样性来排列搜索结果的文档的方法和系统被提供。在一个实施例中，一个排列系统决定在搜索结果中的每一个文档的信息的丰富度。信息的丰富度是一个文档包含有多少与它的主题相关的信息的尺度。具有高信息丰富度的文档（例如，web 页）可能包含包含有与同一主题相关但却具有更低的信息丰富度的文档信息的信息。该排列系统基于它们的关联性而将搜索结果的文档分组，意味着它们被指向类似的主题。该排列系统将文档排序以保证最高排列文档可以包含覆盖每一个主题的至少一篇文档，也就是说，来自于每一个组的一篇文档。该排列系统从在该组中具有文档的最高信息丰富度的每一组中选择文档。当这些文档以排列顺序被提供给用户时，

用户可能将在搜索结果的第一页中发现覆盖各种主题的文档，而不仅仅是单一的受欢迎主题。例如，如果搜索请求包含搜索条件“Spielberg”，则在搜索结果的第一页中的一篇文档可能与 Steven Spielberg 相关，而在搜索结果的第一页中的另一篇文档可能与 Spielberg 教授相关。这样，用户很可能在搜索结果的第一页被呈现覆盖多样化主题的文档，且当感兴趣的主体不是与搜索请求关联的最受欢迎的主题时，用户将不会太沮丧。此外，因为该排列系统排列具有更高信息丰富度的文档高于具有更低信息丰富度的文档，因此用户将很可能在搜索结果的第一页给出的文档中找到期望的信息。

在一个实施例中，该排列系统根据一个相关性曲线图计算搜索结果的文档的信息丰富度。相关性是衡量一篇文档中的信息被包含在另一篇文档的信息中的程度。例如，一篇描述 Spielberg 的电影中的一部电影的文档与所有详细描述 Spielberg 的电影的文档表面上可能具有一个高的相关性。相反地，所有详细描述 Spielberg 的电影的文档对这篇表面上描述 Spielberg 的电影中的一部电影的文档可能具有一个相对低的相关性。与不同主题相关联的文档彼此之间没有相关性。每一篇文档与每一篇其他文档的相关性的汇集表示为相关性曲线图。一篇具有许多其他的与它具有高相关性的文档的文档将可能具有高的信息丰富度，因为它的信息包含许多其他文档的信息。此外，如果那些具有高的相关性的其他文档自身也有相对高的信息丰富度的话，则该文档的信息丰富度也将很高。

在一个实施例中，该排列系统还利用一个相关似性曲线图来帮助保证该搜索结果的高排列文档的多样性。该排列系统根据一个传统的排列技术（例如，关联性）、一种信息丰富度技术或者一些其他的排列技术可以具有文档的初始排列分数。该排列系统最初选择具有最高初始排列分数的文档作为具有最高最终排列分数的文档。该排列系统然后减少具有与已选择的文档高相关性的每一篇文档的排列分数。因为那些文档的内容可能被已选择的文档所包含且代表了多余的信息，所以该排列系统减少该排列分数。该排列系统然后选择余下的具有其后更高排列分数的文档中的文档。该排列系统减少具有与新的已选择的文档高相关性的每一篇文档的排列分数。该排列系统重复这样的处理直到期望数目的文档具有一个最终的排列分数、所有的文档都有一个最终的排列分数或者一些其他的中止条件被满足。在一个实施例中，多样性代表了在文档的集合中的不同的主题的数量，在集合中的文档的信息丰富度表示与整个集合相关的文档

的信息度。

本领域的普通技术人员能够理解该搜索结果的文档可以基于单独的信息丰富度或单独的多样性而被排列，而不是根据信息丰富度和多样性的结合。例如，一个搜索引擎装置可以单独利用信息丰富度，通过识别与相似的主题相关的多组文档并确定在它的组中的每一篇文档的信息丰富度。该搜索引擎装置然后将已确定的信息丰富度分解为该文档的排列，因而它们组的具有最高的信息丰富度的文档将比他们组中的其他的文档排列得更高。例如，该搜索引擎装置可能单独利用多样性，通过识别与相似主题相关的多组文档并保证来自每一组的至少一篇文档在与它的信息丰富度无关的搜索结果中被排列得很高。例如，该搜索引擎装置可以选择在搜索结果的第一页显示来自于在组中具有最高关联性的每一组中的文档。

相关性曲线图表示作为结点的文档和作为在结点之间的有向边的权的相关性值。该排列系统代表一个相关性曲线图，它通过一个将每一篇文档映射到在文档集合中的每一个其他文档的矩形矩阵表示。该排列系统将该矩阵元素的值设置为相应文档的相关性。如果 M 是该矩阵，那么 M_{ij} 代表文档 i 到文档 j 的相关性。该排列系统依靠将每一篇文档表示为一个向量来计算文档的相关性。该向量表示文档的信息化内容。例如，每一个向量可以包含该文档的最重要的 25 个关键词。该排列系统可以根据下述公式计算相关性：

$$\text{aff}(d_i, d_j) = \frac{\vec{d}_i \cdot \vec{d}_j}{\|\vec{d}_i\|} \quad (1)$$

其中 $\text{aff}(d_i, d_j)$ 是文档 d_i 到文档 d_j 的相关性， d_i 代表文档 d_i 的向量， d_j 代表文档 d_j 的向量，同时 $\|\vec{d}_i\|$ 代表向量 d_i 的长度。公式 1 设定了从 d_j 到 d_i 的投影的长度的相关性。本领域的技术人员可以理解该相关性可以以许多种方式来定义。例如，一篇文档对另一篇文档的相关性可以基于这一篇文档中的关键词存在于其他文档的关键词之中的百分比而被定义。在设置理论条件时，一篇文档对另一篇文档的相关性可以被表示成存在于被其他文档中的关键词的数目所分割的两篇文档的交集的关键词的数目。矩阵 M 的每一个元素代表从一篇文档的结点到另一篇文档的结点的相关性曲线图中的有向边。在一个实施例中，该排列系统设定一个低于一个相关性门限值（例如，.2）到 0 的相关性值。概念地，这意味着在相关性为低时，在相关性曲线图中没有从一篇文档的结点到另一篇文档

的结点的有向边。该相关性矩阵可以表示如下：

$$M_{ij} = \begin{cases} \text{aff}(d_i, d_j) & \text{如果 } \text{aff}(d_i, d_j) \geq \text{aff}_i \\ 0 & \text{否则} \end{cases} \quad (2)$$

其中， M_{ij} 是矩阵的一个元素， aff_i 是相关性门限值。在它们之间具有许多边的一组结点可以代表一个单独的主题，因为在该组中的许多文档具有一个大于它们彼此之间的门限相关性的相关性。相反地，在他们之间没有链接的结点代表指向不同的主题的文档。

通过将边分析算法应用到相关性曲线图该排列系统为每一篇文档计算信息的丰富度。该排列系统规格化该相关性矩阵，从而在每一行中值被增加到 1。该规格化相关性矩阵可以表示为如下：

$$\tilde{M}_{ij} = \begin{cases} M_{i,j} / \sum_{j=1}^n M_{ij}, & \text{如果 } \sum_{j=1}^n M_{ij} \neq 0 \\ 0 & \text{否则} \end{cases} \quad (3)$$

其中， $\tilde{M}_{ij}$ 是该规格化矩阵的一个元素。该排列系统根据如下公式计算信息的丰富度：

$$\text{InfoRich}(d_i) = \sum_{\text{所有 } j \neq i} \text{InfoRich}(d_j) \cdot \tilde{M}_{ji} \quad (4)$$

其中 $\text{InfoRich}(d_i)$ 是文档 d_i 的信息丰富度。因此，信息丰富度被递归定义。公式 4 可以按如下表示为矩阵形式：

$$\lambda = \tilde{M}^T \lambda \quad (5)$$

其中 $\lambda = [\text{InfoRich}(d_i)]_{n \times 1}$ 是该规范化相关性矩阵 $\tilde{M}^T$ 的特征向量。由于该规范化相关性矩阵 $\tilde{M}$ 典型地为一个稀疏矩阵，所以全 0 的行可能在它里面出现，这意味着一些文档没有其他的文档与它们有有意义的相关性。为了计算一个有意义的特征向量，该排列系统使用一个卸载因子（例如，.85），它可以是基于文档普及性的一个文档排列。使用卸载因子的该信息丰富度可以表示如下：

$$\text{InfoRich}(d_i) = c \cdot \sum_{\text{所有 } j \neq i} \text{InfoRich}(d_j) \cdot \tilde{M}_{ji} + \frac{(1-c)}{n} \quad (6)$$

其中， c 是卸载因子， n 是在集合中的文档的数目。公式 6 可以用矩阵形式表示如下：

$$\lambda = c \tilde{M}^T \lambda + \frac{(1-c)}{n} \bar{e} \quad (7)$$

其中， $\bar{e}$ 是一个具有所有元素都为 1 的单位向量。该信息丰富度的计算可以被类推为一个信息流程和接收器模型。根据该模型，在每一次迭代时，信息

在结点间流动。文档 d_i 具有一组与它具有相关性的文档 $A(d_i)$ ，文档 $A(d_i)$ 可以如下表示：

$$A(d_i) = \{d_j \mid \forall j \neq i, \text{aff}(d_i, d_j) > \text{aff}_i\} \quad (8)$$

在每一次迭代中，信息可以按照下列的一种规则流动：

1. 根据概率 c （即，该卸载因子），该信息可以流入 $A(d_i)$ 中的一篇文档，同时，流入文档 d_j 的概率与 $\text{aff}(d_i, d_j)$ 成比例。

2. 根据概率 $1-c$ ，该信息可以随机地流入该集合中的任何文档。

从上述的处理中能够推导出一个马尔可夫链，其中，状态由文档给出，而转换（或者流动）矩阵由下式给出

$$c\tilde{M}^T + \frac{(1-c)}{n}U \quad (9)$$

其中 $U = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{n \times n}$ 。每一种状态的固定概率分布由该转换矩阵的首要的特征向量给出。

在一个实施例中，该排列系统通过将信息丰富度与相似性惩罚相组合，来计算一个相关性排列，从而指向相同主题的多篇文档没有全部被很高地排列而排斥指向其他主题的文档。该相似性惩罚的使用导致了在大多数高排列文档中的主题多样性的增加。该排列系统可以在一篇文档的初始相关性排列被设置为它的信息丰富度时，利用一个迭代贪婪算法来计算该相似性惩罚。在每一次迭代中，该算法选择具有次高相关性排列的文档，并通过一个相似性惩罚减少该指向相同主题的文档的相关性排列。因此，一旦一个文档被选择，所有的其他的指向该相同主题的文档将使它们自己的相关性排列减少，以改进代表不同主题的最高排列文档的机会。该排列系统可以根据下式减少文档的相关性排列：

$$AR_j = AR_j - \tilde{M}_{ji} \bullet \text{InfoRich}(d_i) \quad (10)$$

其中， AR_j 表示文档 j 的相关性排列， i 是被选择的文档。因为相似性惩罚是基于相关性矩阵的，一个文档与选择的文档越相似，它的相似性惩罚就越大。

在一个实施例中，该排列系统将一个基于文本的排列（例如，传统的关联性）与一个相关性排列相结合，以产生一个全排列。该排列可以基于分数或者排列而被结合。对于该组合的分数，该基于文本的分数被与相关性排列组合，以给出一个代表该文档的最终分数的全分数。该组合的分数可基于一个基于文本的分数和该相关性排列的线性组合。因为该分数可能具有不同等级的顺序，

该排列系统规格化该分数。该组合的分数可以表示如下：

$$\text{Score}(q, d_i) = \alpha \cdot \frac{\text{Sim}(q, d_i)}{\overline{\text{Sim}}_\Theta(q)} + \beta \cdot \frac{\log \overline{AR}_\Theta}{\log AR_i}, \quad \forall d_i \in \Theta \quad (11)$$

其中， $\alpha + \beta = 1$ ， Θ 代表用于搜索请求 q 的搜索结果， $\text{Sim}(q, d_i)$ 代表搜索请求 q 的文档 d_i 的相似性，和

$$\overline{\text{Sim}}_\Theta(q) = \text{Max}_{\forall d_i \in \Theta} \text{Sim}(q, d_i) \quad (12)$$

$$\overline{AR}_\Theta = \text{Max}_{\forall d_i \in \Theta} AR_i \quad (13)$$

利用组合排列，该基于文本的排列与该相关性排列相结合，以提供一个文档的最终排列。该组合排列可以基于一个基于文本的排列和该相关性排列的线性组合。该组合排列可以表示如下：

$$\text{Score}(q, d_i) = \alpha \cdot \text{Rank}_{\text{Sim}(q, d_i)} + \beta \cdot \text{Rank}_{AR_i}, \quad \forall d_i \in \Theta \quad (14)$$

其中， Score 代表用于搜索请求 q 的文档 d_i 的最终排列。 $\text{Rank}_{\text{Sim}(q, d_i)}$ 代表该基于文本的排列， Rank_{AR_i} 代表该相关性排列。在两个组合算法中的 α 和 β 都是可以被调整的参数。当 $\alpha = 1$ 和 $\beta = 0$ 时，没有再排列被执行，而该搜索结果根据基于文本的搜索而被排列。当 $\beta > \alpha$ 时，在再排列时，更多的权被增加给该相关性排列。当 $\beta = 1$ 和 $\alpha = 0$ 时，该再排列单独地基于该相关性排列而被执行。

图 1 是说明在一个实施例中的一个相关性曲线图的图表。该相关性曲线图 100 包括结点 111-115、结点 121-124 和结点 131，它们每一个代表一篇文档。在结点之间的有向边表示一个结点与另一个结点的相关性。例如，结点 111 与结点 115 具有一个相关性，但是结点 115 与结点 111 没有相关性（或者有一个低于门限水平的相关性）。在这个例子中，结点组 110 包括指向同样的主题的结点 111-115，因为在该结点组中的结点之间有许多边。类似地，结点组 120 包括指向同一的主题的结点 121-124。结点组 130 只有一个结点，因为那个结点与其他任何结点都没有相关性，也没有结点与它有相关性。结点 115 可能具有在结点组 110 中的所有结点的最高信息丰富度，而结点 124 也可能具有在结点组 120 中的所有结点的最高的信息丰富度，因为每一个结点都有最大数目的与它有相关性的结点。

图 2 是说明在一个实施例中的排列系统的部件的方块图。该排列系统 200 包括数据存储器 201-204 和部件 211-216。该文档存储器 201 包含文档的集合且可代表所有经由 Internet 的可用的 web 页。该产生相关性曲线图部件 211 基于文

档存储器中的文档产生一个相关性曲线图。该产生相关性曲线图部件在相关性曲线图存储器 202 中存储该相关性。该计算信息丰富度部件 212 输入来自相关性曲线图存储器的相关性曲线图，并为每一篇文档计算一个信息丰富度分数。该部件将已计算的信息丰富度分数存储在信息丰富度存储器 203 中。在一个实施例中，该产生相关性曲线图部件和该计算信息丰富度部件可以在一个搜索进行之前脱机执行以产生该相关性曲线图和信息丰富度分数。进行搜索部件 213 从用户接收一个搜索请求并从文档存储器的文档中识别搜索结果。该进行搜索部件在搜索结果存储器 204 中存储该搜索结果以及搜索结果的每一篇文档与搜索请求的关联性的一个表示。该计算相似性惩罚部件 214 基于该搜索结果存储器、相关性曲线图存储器和信息丰富度存储器的信息计算一个相似性惩罚以提供给该相关性排列。该计算相关性排列部件 215 为搜索结果中的每一篇文档产生一个相关性排列。该计算相关性排列部件在文档的信息丰富度、相关性曲线图分数和搜索结果中分解。该计算最终分数部件 216 结合该相关性排列和关联性分数来计算最终分数。

在其上该排列系统被执行的该计算装置可以包括一个中央处理单元、存储器、输入装置（例如，键盘和指示装置）、输出装置（例如，显示装置）和存储装置（例如，磁盘驱动器）。该存储器和存储装置是包括执行该排列系统的指令的计算机可读介质。此外，该数据结构和信息结构可以被存储或者经由一个数据传输介质例如一个在通讯链路上的信号而被传送。各种各样的通讯链路可以被使用，例如 Internet 局域网、广域网或者点对点拨号上网连接器。

该排列系统可以在各种各样的操作环境中被执行。各种公知的适合于使用的计算系统、环境和配置包括个人计算机、服务器计算机、手提式或者膝上型装置、多处理机系统、基于微处理器的系统、可编程消费电子装置、网络 PC、小型计算机、大型计算机，包含任何上述系统和装置的分布式计算环境等等。

该排列系统可以被描述为普通的计算机可执行指令的内容，例如，由一个或多个计算机或者其他装置执行的程序模块。通常，程序模块包括执行特定任务或者执行特定的抽象数据类型的常规程序、程序、对象、组件、数据结构等等。典型地，该程序模块的功能可以是在各种实施例中期望的组合式或者分布式的。

图 3 是说明在一个实施例中的排列系统的全部处理的流程图。该排列系统

被提供了一个可以代表一个搜索结果的文档的集合。在块 301 中，该部件为该文档的集合构造了一个相关性曲线图。该部分还可以构造覆盖一个在脱机的文档的语言资料库中（例如，所有的 web 页）的所有文档或者仅仅覆盖实时采集的文档的相关性曲线图。在块 302 中，该部件计算该集合的每一篇文档的信息丰富度。在块 303 中，该部件排列该集合的文档，而后结束。

图 4 是说明在一个实施例中的一个构造相关性曲线图部件的处理的流程图。该部件通过了一个文档的集合并构造一个用于那些文档的相关性曲线图。在块 401-403 中，该部件为文档的集合中的每一篇文档循环产生文档向量。在块 401 中，该部件选择在集合中的下一篇文章档。在决定块 402 中，如果在集合中的所有文档已经被选择，然后，该部件继续到块 404，否则该部件继续到块 403。在块 403 中，该部件为已选择的文档产生文档向量，然后循环到块 401 以选择集合中的下一篇文章档。在块 404-408 中，该部件为集合中的每一对文档计算相关性。在块 404 中，该部件从第一篇文章档开始选择在集合中的下一篇文章档。在决定块 405 中，如果所有的文档都已经被选择，则该部件返回该相关性曲线图，否则该部件继续到块 406。在块 406-408 中，该部件循环挑选集合中的每一篇文章档。在块 406 中，该部件从第一篇文章档开始挑选在集合中的下一篇文章档。在决定块 407 中，如果在集合中的所有文档已经被挑选，则该部件循环到块 404 以选择集合中的下一篇文章档，否则该部件继续到块 408。在块 408 中，该部件根据公式 1 计算从选择的文档到已挑选的文档的相关性，然后循环到块 406 以挑选集合中的下一篇文章档。

图 5 是说明在一个实施例中的一个排列文档部件的处理的流程图。该部件通过了一个已经具有它的已产生的相关性曲线图和已计算过的每一篇文章档的信息丰富度的文档的集合。在块 501-503 中，该部件循环初始化集合中的每个文档的相关性排列到它的信息丰富度。在块 501 中，该部件选择集合中的下一篇文章档。在决定块 502 中，如果所有的文档都已经被选择，则该部件继续到块 504，否则该部件继续到块 503。在块 503 中，该部件设置已选择的文档的相关性排列到已选择的文档的信息丰富度，然后循环到块 501 以选择在集合中的下一篇文章档。在块 504-508 中，该部件循环识别多对文档并通过一个相似性惩罚调整相关性排列。在块 504 中，该部分件选择具有最高相关性排列的下一篇文章档。在决定块 505 中，如果一个中止条件被达到，则该部件返回已排列的文档，否则该

部件继续到块 506。在块 506-508 中，该部件循环挑选文档并用一个相似性惩罚调整相关性排列。在块 506 中，该部件挑选在相关性曲线图中，具有相对已选择的文档的相关性被指示为非 0 值的下一篇文章档，用于从已挑选的文档到已选择的文档的相关性。在决定块 507 中，如果所有的文档已经被挑选，则该部件循环到块 504 以选择具有最高相关性排列的下一篇文章档。在块 508 中，该部件根据公式 10 用一个相似性惩罚为已挑选的文档调整相关性排列。该部件然后循环到块 506 以挑选具有与已选择的文档的相关性的下一篇文章档。

本领域的技术人员可以理解尽管在这里已经被描述的本排列系统的特定实施例是用于说明的目的，但在不脱离本发明的精神和范围的前提下，可以做各种各样的改变。在一个实施例中，该排列系统可以在一块接一块的基础上计算相关性和信息丰富度而不是在文档接文档的基础上。一个块代表通常与一个单一主题相关的 web 页的信息。该 web 页的排列可以部分基于一个块对它的 web 页的重要性。该块的重要性被描述在美国专利申请号_____题目为“用于计算在显示页中的块的重要性的方法和系统”并在_____公开，在这里仅结合作为参考。因此，除了附加的权利要求之外，本发明没有被限制。

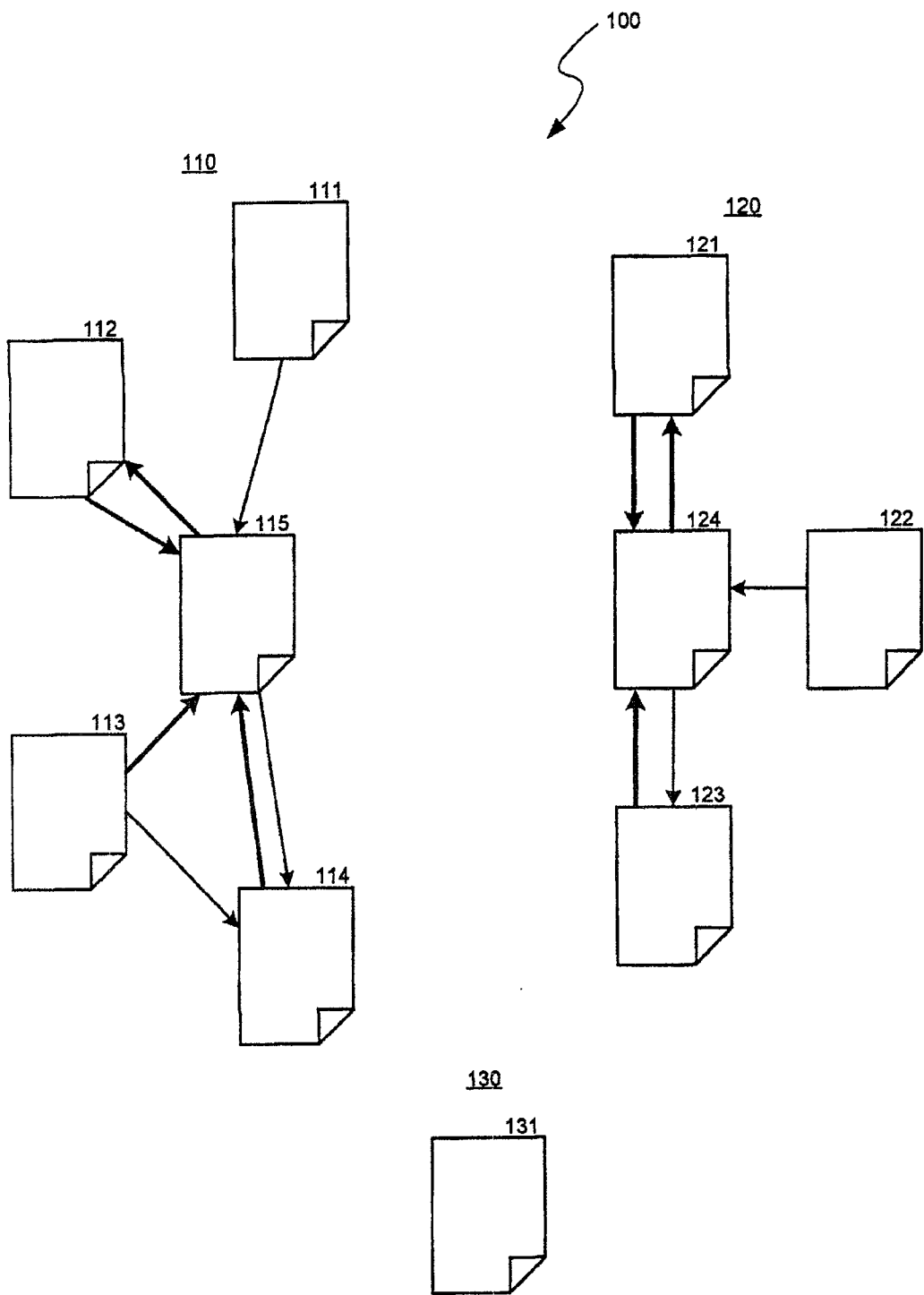


图 1

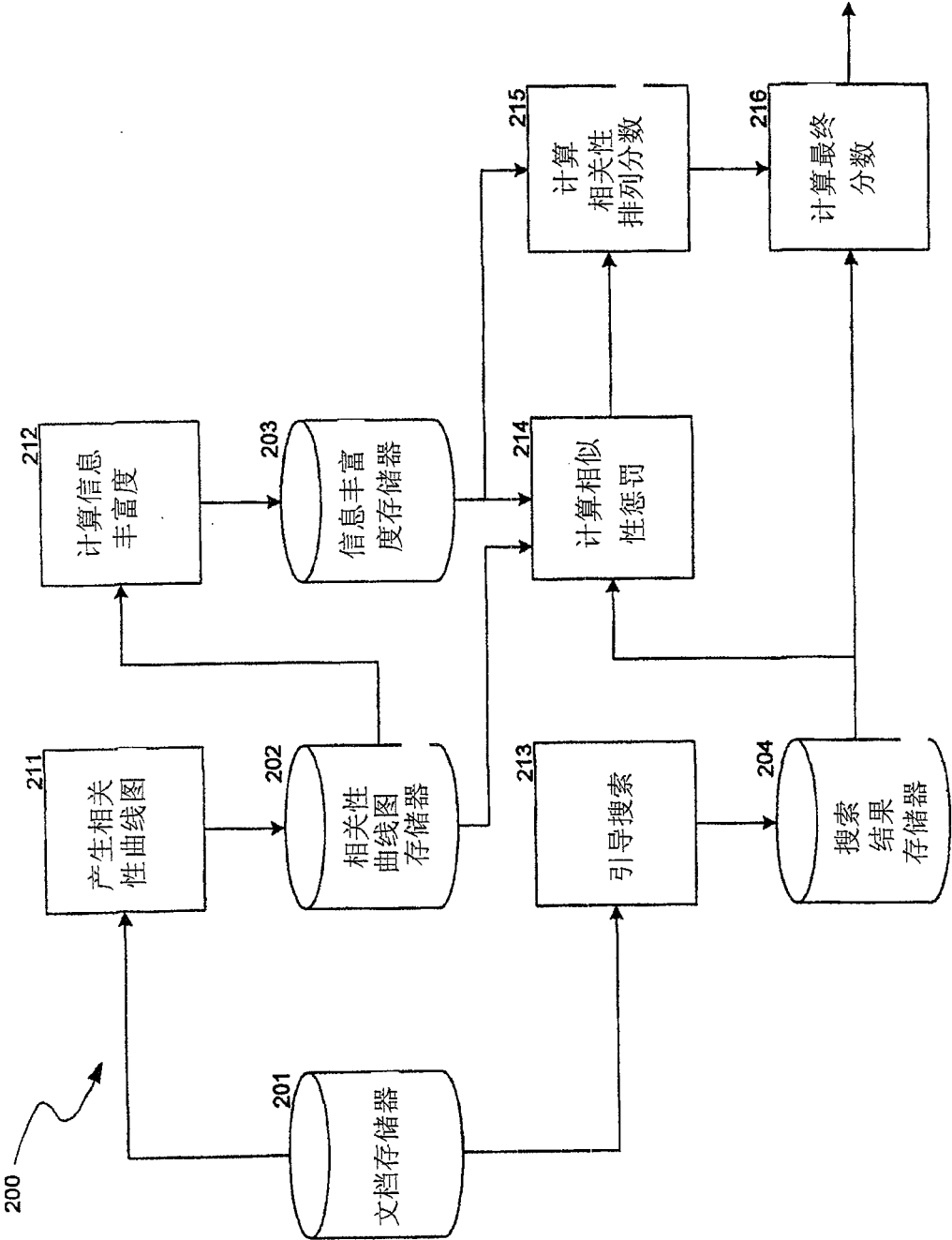


图 2

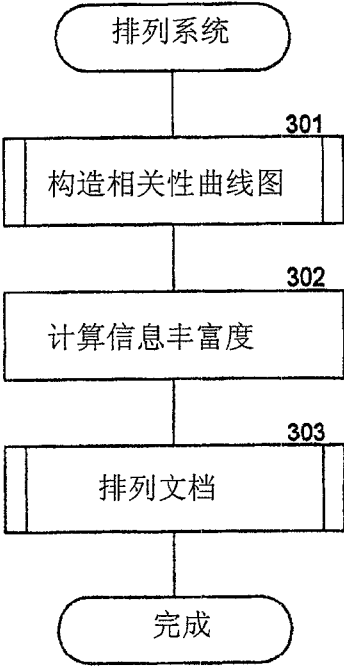


图 3

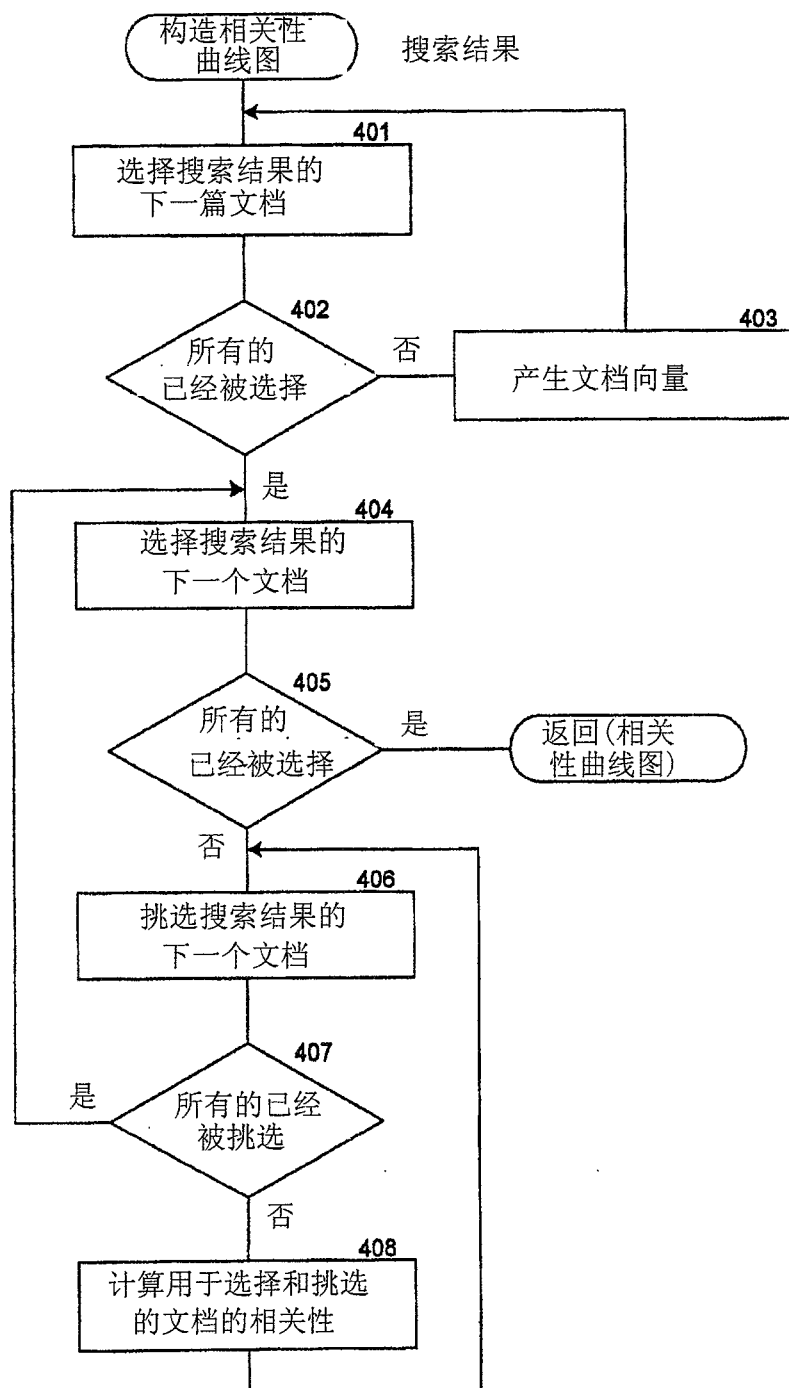


图 4

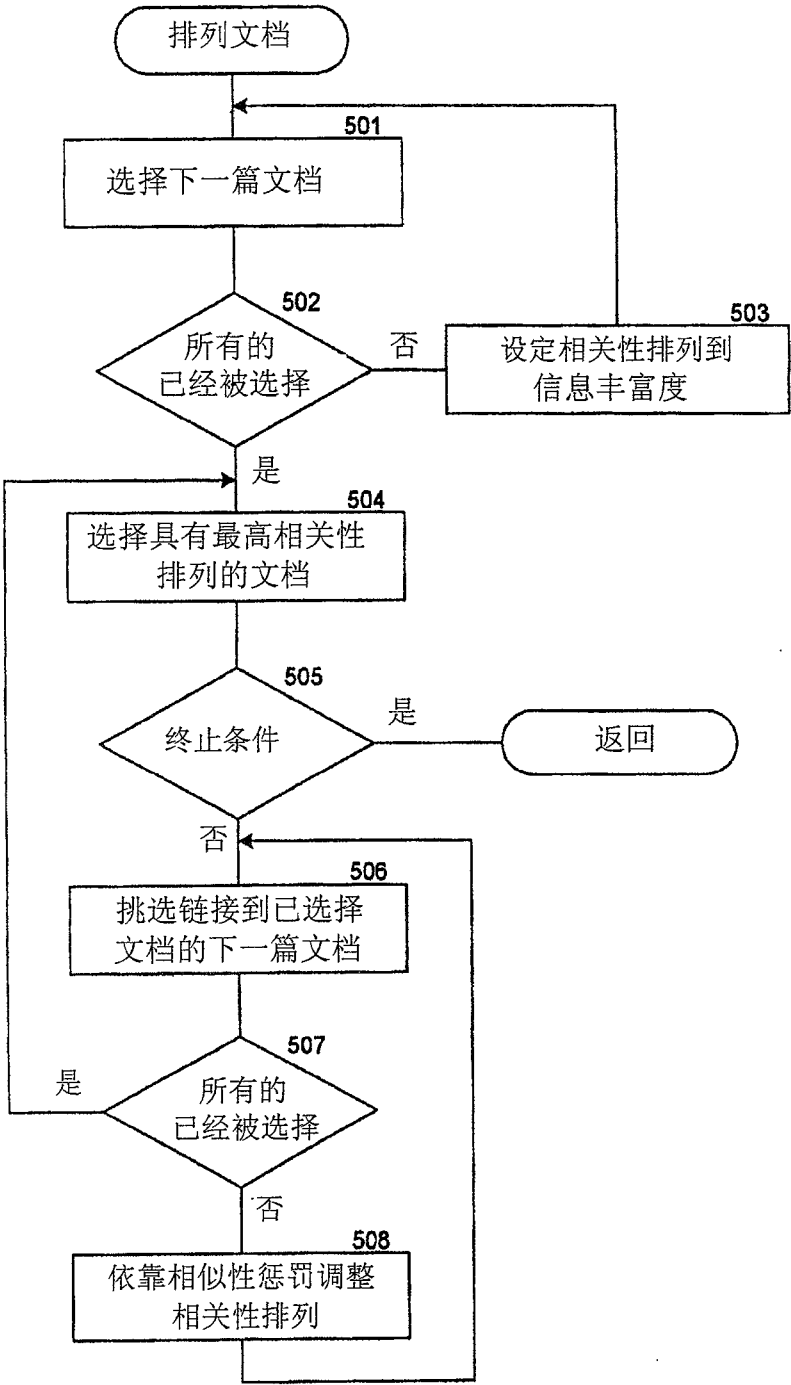


图 5