[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 200580030640.6

[45] 授权公告日 2009 年 5 月 6 日

[11] 授权公告号 CN 100485677C

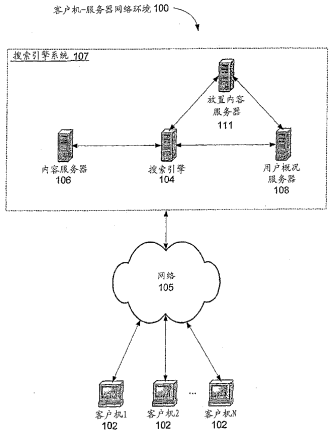
[22] 申请日 2005.7.12
[21] 申请号 200580030640.6
[30] 优先权
[32] 2004. 7.13 [33] US [31] 10/890,854
[86] 国际申请 PCT/US2005/025081 2005.7.12
[87] 国际公布 WO2006/017364 英 2006.2.16
[85] 进入国家阶段日期 2007.3.13
[73] 专利权人 谷歌股份有限公司
地址 美国加利福尼亚州
[72] 发明人 奥伦·E·赞米尔
杰弗里·L·科恩
安德鲁·B·菲克斯
斯蒂芬·R·劳伦斯
[56] 参考文献
CN1324044A 2001.11.28
EP1050830A2 2000.11.8
US5724567A 1998.3.3
US2004/0044571A1 2004.3.4
US5754939A 1998.5.19
US2002/0024532A1 2002.2.28
US6385619B1 2002.5.7
审查员 胡雅娟
[74] 专利代理机构 北京市柳沈律师事务所
代理人 黄小临 王志森
权利要求书 3 页 说明书 27 页 附图 13 页

[54] 发明名称

搜索结果中放置内容排序的个性化

[57] 摘要

利用用户概况排序搜索引擎返回的搜索结果中的放置内容的系统和方法。 用户概况基于用户提交的搜索查询、用户与搜索引擎识别的文档的特定交互、和用户提供的个人信息。 放置内容按至少部分基于特定放置内容与用户概况的相似性的分数排行。 用户概况可以在客户机 – 服务器网络环境的客户机方或服务器方创建和/或存储在其中。



1. 一种使与搜索查询相联系的放置内容个性化的方法，包含:

接收来自用户的搜索查询;

访问与用户相联系的用户概况，其中该用户概况的至少一部分基于多个以前提交的搜索查询中的查询术语;

识别与搜索查询匹配的一组放置内容;

按照用户概况、放置内容的各自出价、和放置内容的各自点击通过率将分数指定给该组放置内容的每个放置内容，其中将分数指定给该组放置内容的每个放置内容包括:

确定用户概况和与每个放置内容相联系的放置内容概况之间的相似性分数，以及，

将所述相似性分数与各自点击通过率和各自出价相结合，以确定指定给每个放置内容的分数;

按照放置内容的分数排位该组放置内容; 以及

将与所述排位相应的排位的放置内容组的至少一个小组发送给用户。

2. 根据权利要求 1 所述的方法，其中，用户概况的至少一部分基于有关用户的信息，包括从一组文档中导出的信息，该组文档包含从由如下组成的一个组中选择的多个文档: 通过来自搜索引擎的搜索结果识别的文档、链接到通过来自搜索引擎的搜索结果识别的文档的文档、链接到用户访问的文档的文档、和用户浏览的文档。

3. 根据权利要求 1 所述的方法，其中，相似性分数的确定包括:

确定用户概况的用户概况向量与放置内容的放置内容概况向量之间的数学距离，用户概况向量包括第一类别和各自权重对，放置内容概况向量包括第二类别和各自权重对。

4. 根据权利要求 1 所述的方法，进一步包括将相似性分数与比例因子相联系，其中，比例因子是从多个亚因子中选择的一个，每个亚因子与相似性分数值的相应范围相联系。

5. 根据权利要求 1 所述的方法，进一步包括将相似性分数与比例因子相联系，其中，将分数指定给该组放置内容的每个放置内容包括乘以比例因子、各自点击通过率和各自出价。

6. 根据权利要求 5 所述的方法，其中，与最大相似性分数相联系的比例因子小于与中点相似性分数相联系的比例因子。

7. 根据权利要求 6 所述的方法，进一步包括提供作为广告的放置内容。

8. 一种使与搜索查询相联系的放置内容个性化的系统，包含:

用于存储用户概况的装置，该用户概况的至少一部分基于多个以前提交的搜索查询中的查询术语；和

用于存储多个放置内容的装置;

放置内容服务器，用于识别与搜索查询匹配的多个放置内容的一个小组，按照用户概况、放置内容的各自出价、和放置内容的各自点击通过率将分数指定给该小组中的每个放置内容，并且根据放置内容的各自分数排位该小组，其中该分数基于用户概况和与每个放置内容相联系的放置内容概况之间的相似性分数，并且，其中所述相似性分数被与各自点击通过率和各自出价相结合，以确定指定给每一个放置内容的分数；以及

用于将与所述排位相应的排位的放置内容组的至少一个小组发送给用户的装置。

9. 根据权利要求 8 所述的系统，其中，用户概况的至少一部分基于有关用户的信息，包括从一组文档中导出的信息，该组文档包含从由如下组成的一个组中选择的多个文档：通过来自搜索引擎的搜索结果识别的文档、链接到通过来自搜索引擎的搜索结果识别的文档的文档、链接到用户访问的文档的文档、和用户浏览的文档。

10. 根据权利要求 8 所述的系统，其中，相似性分数基于用户概况的用户概况向量与放置内容的放置内容概况向量之间的数学距离，用户概况向量包括第一类别和各自权重对，放置内容概况向量包括第二类别和各自权重对。

11. 根据权利要求 8 所述的系统，进一步包括与相似性分数相联系的比例因子，其中，比例因子是多个亚因子之一，每个亚因子与相似性分数值的相应范围相联系。

12. 根据权利要求 8 所述的系统，其中，该组放置内容中的每个放置内容的分数对应于放置内容的各自比例因子、各自点击通过率和各自出价的乘积。

13. 根据权利要求 12 所述的系统，其中，与最大相似性分数相联系的比例因子小于与中点相似性分数相联系的比例因子。

14. 根据权利要求 13 所述的系统，其中，放置内容是广告。

15. 一种使与搜索查询相联系的放置内容个性化的系统，包含:

至少一个计算机，该计算机包括:

接收来自用户的搜索查询的装置;

访问用户的用户概况的装置，其中该用户概况的至少一部分基于多个以前提交的搜索查询中的查询术语;

识别与搜索查询匹配的一组放置内容的装置;

按照用户概况、放置内容的各自出价、和放置内容的各自点击通过率将分数指定给该组放置内容的每个放置内容的装置，其中将分数指定给该组放置内容的每个放置内容的装置包括:

确定用户概况和与每个放置内容相联系的放置内容概况之间的相似性分数的装置，以及，

将所述相似性分数与各自点击通过率和各自出价相结合，以确定指定给每个放置内容的分数的装置;

根据放置内容的分数排位该组放置内容的装置; 以及

将与所述排位相应的排位的放置内容组的至少一个小组发送给用户的装置。

搜索结果中放置内容排序的个性化

相关申请

本申请是2003年9月30日提出的美国专利申请第10/676,711号的部分继续申请，特此全文引用以供参考。

技术领域

本发明一般涉及计算机网络系统中的搜索引擎领域，尤其涉及响应用户提交的搜索查询，创建和利用用户概况(profile)定制放置内容（placed content）的排序的系统和方法。

背景技术

搜索引擎提供了可以响应用户提交的搜索查询迅速扫描的来自因特网（或内联网）的带索引文档的强大来源。这样的查询通常非常短（平均约两三个字）。随着可通过因特网访问的文档的数量不断增加，与查询匹配的文档的数量可能也不断增加。但是，从用户的角度来看，并非与查询匹配的每个文档都同等重要。其结果是，如果搜索引擎不根据文档与用户查询的相关性排序搜索结果，用户很容易被搜索引擎返回的大量文档淹没。

改善搜索结果与搜索查询的相关性的一种途径是利用不同网页的链接结构计算可以用于影响搜索结果的排行的全局“重要”分。这种途径有时被称为PageRank（页面排位）算法。PageRank算法的更详细描述可以从S. Brin和L. Page的文章“大型超文本搜索引擎的解剖”（“The Anatomy of a Large-Scale Hypertextual Search Engine”, S. Brin and L. Page, 7th International World Wide Web Conference, Brisbane, Australia）和美国专利第6,285,999号中找到，特此引用这两个文件作为背景信息，以供参考。

PageRank算法的重要假设是，存在从随机挑选的网页开始其万维网冲浪旅程和继续点击嵌在网页中的链接，而决不会击中“返回”按钮的“随机冲浪者”。最终，当这个随机冲浪者厌烦了该旅程时，其可以通过随机挑选另一

页网页重新开始新的旅程。随机冲浪者访问（即，观看或下载）一页网页的概率取决于网页的页面排位。

从最终用户的角度来看，由于搜索引擎不要求用户提供可以唯一标识用户的任何信息，无论谁提交查询，利用 PageRank 算法的搜索引擎都以相同的方式对待搜索查询。影响搜索结果的唯一因素是搜索查询本身，例如，在查询中有多少个术语和按什么顺序排列。搜索结果最适合抽象用户，即，“随机冲浪者”的兴趣，而不会被调整成适合特定用户的偏爱或兴趣。

实际上，像随机冲浪者那样的用户从不会存在。当用户向搜索引擎提交查询时，每个用户都有他自己的偏爱。搜索引擎返回的搜索结果的质量必须通过其用户满意度来评价。当用户的偏爱可以通过查询本身适当定义时，或当对于特定查询，用户的偏爱与随机冲浪者的偏爱类似时，用户更有可能对搜索结果感到满意。但是，如果用户的偏爱因未清楚反映在搜索查询本身中的一些个人因素而严重偏离，或如果用户的偏爱与随机用户的偏爱的差异相当大，来自相同搜索引擎的搜索结果即使不是完全无用，也可能几乎不可用于用户。

正如上面提出的那样，随机冲浪者的旅程往往是随机的和中性的，不会明显倾向于特定方向。当搜索引擎只返回与查询匹配的少量搜索结果时，返回结果的顺序较不重要，因为提出请求的用户花得起浏览它们的每一个以发现与自己最有关的项目的时间。但是，随着几十亿的网页与因特网连接，搜索引擎往往返回数百，甚至数千与搜索查询匹配的文档。在这种情况下，搜索结果的排序非常重要。偏爱与随机冲浪者不同的用户可能无法在列在搜索结果中的前五个到前十个文档中找到他正在寻找的文档。当发生这种情况时，通常留给用户两种选择:（1）花费审阅更多列出文档以定位相关文档所需的时间；或（2）改进搜索查询，以便减少与查询匹配的文档的数量。改进查询往往不是无足轻重的任务，有时需要比用户所拥有更多的有关主题的知识或更多的有关搜索引擎的专门知识，和有时需要比用户愿意花费的时间和努力更多的时间和努力。

例如，假设用户向搜索引擎提交只含有一个术语“blackberry”的搜索查询。在没有任何其它上下文的情况下，在基于 PageRank 搜索引擎返回的一系列文档的顶端可以是到 www.blackberry.net 的链接，因为这个页面具有最高页面排位。但是，如果查询请求者是对食物和烹饪感兴趣的人士，将搜索

结果排序成在返回结果的顶端包括含有食谱或其它食物相关文本、图片等的页面也许更有用。最好拥有以下搜索引擎：其能够重新排序其搜索结果，或定制搜索结果，以便强调提交搜索查询的人士最有可能感兴趣的页面。并且，最好使这样的系统只需要来自各个用户的最少输入，其大部分或者全部无需与用户偏爱和兴趣有关的来自用户的显式输入就可以操作。最后，最好使这样的系统可以满足用户在安全和隐私方面的要求。

发明内容

在使放置内容个性化的方法中，确定用户的兴趣，和访问与用户相联系的用户概况。识别与用户兴趣匹配的一组放置内容，和按照用户概况排序该组放置内容。

在本发明的一个方面中，搜索引擎利用用户概况定制可以包括放置内容，以及其它或一般内容的搜索结果。用户概况包含表征用户兴趣或偏爱的多个项目。这些项目是从包括用户提交的先前搜索查询、去向或来自先前查询识别的文档的链接、来自识别文档的取样内容，以及用户隐式或显式提供的个人信息的各种信息源中提取出来的。

当搜索引擎接收到来自用户的搜索查询时，它识别与搜索查询匹配的一组放置内容。将每个放置内容与至少部分基于放置内容与用户概况的相似性的排位相联系。然后，根据它们的排位排序放置内容项。

包括用户概况构建和搜索结果重新排序和/或评分的本发明可以在客户机-服务器网络环境的客户机方或服务器方实现。

附图说明

通过结合附图对本发明的优选实施例进行如下详细描述，本发明的上述特征和优点，以及本发明的其它特征和优点将更加清楚，在附图中：

图 1 例示了客户机-服务器网络环境；

图 2 例示了多个用户信息源和它们与用户概况的关系；

图 3 是可以用于为数个用户存储基于术语概况的示范性数据结构；

图 4A 是可以用于分类用户过去搜索经历的示范性类别图；

图 4B 是可以用于为数个用户存储基于类别概况的示范性数据结构；

图 5 是可以用于为数个用户存储基于链接概况的示范性数据结构；

图 6 是例示段落取样的流程图;

图 7A 是例示上下文分析的流程图;

图 7B 描绘了利用上下文分析识别重要术语的过程;

图 8 例示了可以分别用于在基于术语、基于类别和/或基于链接分析之后存储有关文档的信息的数种示范性数据结构;

图 9A 是例示根据一个实施例的个性化万维网搜索过程的流程图;

图 9B 是例示根据另一个实施例的个性化万维网搜索过程的流程图;

图 10 是个性化搜索引擎的方块图; 和

图 11 是例示根据本发明一个实施例的个性化放置内容处理过程的流程图。

在所有附图中，相同的标号自始至终表示相应的部件。

优选实施例详述

下面讨论的实施例包括根据用户过去使用搜索引擎的经历创建用户概况，然后，响应用户提供的搜索查询，利用用户概况排行搜索结果的系统和方法。

图 1 提供了可以实现本发明的典型客户机-服务器网络环境 100 的概貌。数个客户机 102 通过网络 105，例如，因特网与搜索引擎系统 107 连接。搜索引擎系统 107 包含一个或多个搜索引擎 104。搜索引擎 104 负责处理客户机 102 提交的搜索查询，按照搜索查询生成搜索结果，和将结果返回给客户机。搜索引擎系统 107 还可以包含一个或多个内容服务器 106、一个或多个用户概况服务器 108、和一个或多个放置内容服务器 111。内容服务器 106 存储从不同网站中检索的大量带索引文档。可替代地，或另外，内容服务器 106 存储在各种网站上存储的文档的索引。在一个实施例中，根据文档的链接结构将页面排位指定给每个带索引文档。页面排位用作文档重要性的查询无关度量。搜索引擎 104 与一个或多个内容服务器 106 通信，响应特定搜索查询选择数个文档。搜索引擎根据文档的页面排位、与文档相联系的文本、和搜索查询将分数指定给每个文档。搜索引擎 104 可以与一个或多个放置内容服务器 111 通信，以与搜索结果一起提供广告或其它类型的放置内容。放置内容服务器 111 可以与一个或多个用户概况服务器 108 通信。放置内容将在下面作更全面描述。

用户概况服务器 108 存储数个用户概况。每个概况包括唯一标识用户的信息，以及他的以前搜索经历和个人信息，这些信息可以用于响应这个用户提交的搜索查询改进搜索结果。构建用户概况可使用不同的途径。例如，可以通过要求首次用户填表或回答问卷来创建用户概况。这种途径可用在像银行开户那样的某些应用中。但是，在搜索引擎的情况下，这难以成为讨人喜欢的一种。首先，用户与搜索引擎的交互通常是动态过程。随着时间流逝，用户的兴趣可能发生变化。这种变化可以从用户提交的搜索查询、用户对搜索结果的管理，或两者中反映出来。除非用户选择定期更新他的回答，否则用户对表格上的问题的回答往往随时间变成越来越没用。与在联机银行帐户的情况中偶尔更新电话号码不同，在搜索引擎的情况中频繁更新用户概况会严重影响它的用户友好性，当用户从当前可用的搜索引擎中选择搜索引擎时，这是重点要考虑的。并且，众所周知，用户一般不情愿提供像填满表格那样的显式反馈，因为许多用户发现这太繁重了。因此，虽然一些用户可能提供有关他们兴趣的显式反馈，但最好拥有无需要求用户作出任何显式的或新的动作，就可以隐式地获得有关用户兴趣的信息的过程。

人们已经观察到，搜索引擎用户的过去搜索活动提供了有关用户个人搜索偏爱的有用暗示。图 2 提供了有利于构建用户概况的一系列用户信息源。例如，以前提交的搜索查询 201 非常有助于概括用户的兴趣。如果用户提交了多个与糖尿病有关的搜索查询，那么，该用户更有可能对这个话题感兴趣。如果用户随后提交包括术语“organic food”的查询，那么，可以合理地推断，他可能更对有助于抗糖尿病的那些有机食物感兴趣。类似地，与响应先前搜索查询的搜索结果相联系的统一资源定位符（URL）203 和它们的相应锚定文本 203 有助于确定用户的偏爱，尤其对于用户已经选择或“访问”过（例如，用户下载或观看过）的搜索结果项。当第 1 页包含到第 2 页的链接，和该链接含有与它相联系的文本（例如，与链接相邻的文本）时，与该链接相联系的文本被称为与第 2 页有关的“锚定文本”（anchor text）。锚定文本建立起文档中与 URL 链接相联系的文本与 URL 链接指向的另一个文档之间的关系。锚定文本的优点包括，它往往可以提供 URL 链接指向的文档的精确描述，和可以用于对像图像或数据库那样，不能对通过基于文本搜索引擎加索引的文档加索引。

在接收到搜索结果之后，用户可以点击一些 URL 链接，从而下载那些链

接引用的文档，以便掌握更多有关那些文档的细节。可以将某些类型的一般信息 207 与一组用户所选或用户识别的文档相联系。为了形成用户概况，从中导出包括在用户概况的信息的识别文档可以包括：通过来自搜索引擎的搜索结果识别的文档、用户访问（例如，利用（例如）浏览器应用程序观看或下载）的文档（包括未在以前搜索结果中识别出的文档）、链接到通过来自搜索引擎的搜索结果识别的文档的文档、和与用户访问的文档，或链接到这样文档的任何一个小组的文档。

有关识别文档的一般信息 207 可以回答像“文档的格式是什么？”、“是超文本标记语言（HTML）、简单文本、可移植文档格式（PDF）、还是 Microsoft Word？”、“文档的主题是什么？”、“是科学、健康、还是商业方面？”那样的问题。这个信息也有助于概括用户兴趣。另外，像用户花费了多长时间观看文档、文档上的滚动活动量、和用户打印文档、保存文档还是给文档加上书签那样，有关用户针对用户所选文档（本文有时称为识别文档）的活动的信息 209 也暗示着文档对用户的重要性，以及用户的偏爱。在一些实施例中，有关用户活动的信息 209 用在加权从用户识别文档中提取或导出的信息的重要性的时候。在一些实施例中，有关用户活动的信息 209 用于确定哪个用户识别文档用作导出用户概况的基础。例如，信息 209 可以用于只选择经历过重要用户活动（按照预定准则）的文档来生成用户概况，或信息 209 可以用于从概括过程（profiling process）中排除用户观看时间少于预定阈时间量的文档。

来自以前搜索活动的识别文档的内容是有关用户兴趣和偏爱的丰富信息源。出现在识别文档中的关键术语和它们出现在识别文档中的频率不仅可用于给文档加索引，而且也是用户个人兴趣的强烈指示，尤其当它们与上述其它类型用户信息组合在一起时。在一个实施例中，为了构建用户概况，取代整个文档，从识别文档中提取取样内容 211，以便节省存储空间和计算成本。在另一个实施例中，可以分类与识别文档有关的各种信息，以构成有关识别文档的类别信息 213。所述各种信息可以包括以前访问过页面的个体的类型或可以描述文档的其它元数据。下面将提供有关内容取样、在识别文档中识别关键术语的过程、和类别信息的使用的更多讨论。

用户概况的另一个潜在信息源是用户浏览模式 217。用户浏览模式可以用用户在一段时间，譬如，前 N 天（例如，60 天）内访问的 URL 表示。

在一些实施例中，按照其年龄加权用户概况信息，越新的信息权重越大，越老的信息权重越小。这有助于用户概况更好地跟踪用户兴趣的变化，和减小过去兴趣对用户的影响或下降兴趣主题对用户的影响。各种各样的数据结构可以用于支持时间加权用户概况，通常包括保存与一系列时间间隔相联系的用户信息的许多箱子（bin）或叠层(tier)。

可选地，用户可以选择提供个人信息 215，包括与用户相联系的人口和地理信息，譬如，用户年龄或年龄范围、教育程度或范围、收入水平或范围、语言偏爱、婚姻状况、地理位置（例如，用户居住的国家、州和城市，和可能还包括像街道地址、邮政编码、和电话区号那样的附加信息）、文化背景或偏爱、或这些信息的任何一个小组。与像用户喜欢的运动或电影那样往往随时间而变的其它类型个人信息相比，这种个人信息更加静态和更难以从用户的搜索查询和搜索结果中推测，但在正确解释用户提交的某些查询时也许是至关重要的。例如，如果用户提交了包含“Japanese restaurant”（日本餐馆）的查询，那么，他非常有可能在搜索本地日本餐馆用餐。如果不知道用户的地理位置，就难以将搜索结果排序成使与用户真正意图最有关的那些项目放在顶端。但是，在某些情况下，可以推测这个信息。例如，用户往往选择与与他们生活的地方相对应的特定区域相联系的结果。

从各种用户信息源中创建用户概况 230 是既动态又复杂的过程。在一些实施例中，将该过程划分成子过程。每个子过程从特定角度生成表征用户兴趣或偏爱的一种用户概况。它们是:

•基于术语概况 231 – 这个概况代表用户对数个术语的搜索偏爱，其中，给予每个术语以指示该术语对用户的重要性的权重;

•基于类别概况 233 – 这个概况将用户搜索偏爱与可以以分层方式组织的一组类别相关联，其中，给予每个类别以指示用户搜索偏爱与该类别之间的关联程度的权重; 和

•基于链接概况 235 – 这个概况识别直接或间接与用户搜索偏爱有关的数个链接，其中，给予每个链接以指示用户搜索偏爱与该链接之间的相关性的权重。

在一些实施例中，用户概况 230 只包括这些概况 231、233、235 的一个小组，例如，仅仅这些概况的一个或两个。在一个实施例中，用户概况 230 包括基于术语概况 231 和基于类别概况 233，但不包括基于链接概况 235。

例如，通过将一组搜索术语（例如，来自每个独立查询）或识别内容术语（来自特定识别文档）映射到类别，然后，汇集所得的一组类别，针对它们的出现频率和搜索术语或识别内容术语与类别的相关性两者加权类别，可以构建出基于类别概况 233。可替代地，为了映射到加权类别，可以将在一段时间内累积的所有搜索术语或识别内容术语当作一个组来对待。此外，也可以将用户提供个人信息 215 映射到加权类别，并且，可以将那些类别与利用上述的任何技术生成的加权类别组合或汇集在一起。也可以使用将用户相关信息映射到类别的其它适当方式。

在一些实施例中，用户概况 230 是基于与多个用户相联系的信息的汇集概况。汇集了概况信息的用户可以以许多方式选择或识别。例如，作为俱乐部或其它机构的成员，或特定公司的雇员的所有用户可以使其概况信息汇集在一起。在另一个例子中，汇集前用户概况相似的用户可以使其概况信息汇集在一起。可替代地，一个机构或网站可以拥有与之相联系、可以根据机构成员的活动自动生成或可以由机构或为机构定制的“用户概况”。当执行搜索查询时，或当与任何其它适当信息服务结合在一起提供放置内容或其它内容，以便帮助选择请求者或订户感兴趣的内容时，搜索引擎或其它服务可以利用机构的用户概况。

在一个实施例中，在与搜索引擎相联系的服务器（例如，用户概况服务器 108）上创建和存储用户概况。这样部署的优点在于，多个计算机可以容易地访问用户概况，并且，由于概况存储在与搜索引擎 104（或它的一部分）相联系的服务器上，搜索引擎 104 可以容易地利用它，以便使搜索结果个性化。在另一个实施例中，可以在在网络环境下有时称为客户机的用户计算机上创建和存储用户概况。在用户计算机上（例如，在 cookie 中）创建和存储用户概况不仅降低搜索引擎服务器的计算和存储成本，而且满足一些用户的隐私要求。在又一个实施例中，可以在客户机上创建和更新用户概况，但存储在服务器上。这样的实施例结合了例示在其它两个实施例中的一些好处。这种布置的缺点在于，它可能增加了客户机和服务器之间的网络流量。本领域的普通技术人员应该明白，本发明的用户概况可以利用客户机计算机、服务器计算机、或两者来实现。

图 3 例示了可以用于为数个用户存储基于术语概况的示范性数据结构 - 基于术语概况表 300。表 300 包括数个记录 310，每个记录对应于用户基于术

语概况。基于术语概况记录 310 包括数个列，这些列包括 USER_ID 列 320 和（TERM，WEIGHT）（术语，权重）对的数个列 340。USER_ID 列存储唯一标识用户或共享同一组(TERM,WEIGHT)对的一群用户的值，和每个(TERM,WEIGHT)对 340 包括通常对用户或一群用户来说重要、典型地长 1－3 个字的术语、和量化术语的重要性的与术语相联系的权重。在一个实施例中，可以将术语表示成一个或多个 n-gram（n 元组）。n-gram 被定义成 n 个记号的序列，其中，记号可以是字。例如，短语“search engine”是长度为 2 的 n-gram，和字“search”是长度为 1 的 n-gram。

n-gram 可以用于将文本对象表示成向量。这样就可以应用为向量适当定义、但不是为一般对象适当定义的几何、统计和其它数学技术。在本发明中，n-gram 可以用于根据数学函数对术语的向量表示的应用来定义两个术语之间的相似性度量。

术语的权重未必是正值。如果术语具有负权重，这可能暗示着用户偏向于他的搜索结果不包括这个术语，和负权重的幅度表示用户偏向于在搜索结果中避开这个术语的程度。举例来说，对于在加州圣克鲁斯（Santa Cruz，California）的一群冲浪爱好者来说，基于术语概况可能包括具有正权重的像“surfing club”、“surfing event”和“Santa Cruz”那样的术语。像“Internet surfing”或“web surfing”那样的术语也可能包括在概况中。但是，这些术语更有可能拥有负的权重，因为它们与共享这个基于术语概况的用户的真正偏爱无关和混淆了共享这个基于术语概况的用户的真正偏爱。

基于术语概况利用特定术语逐项列出用户偏爱，每个术语具有一定权重。如果一个文档与用户基于术语概况中的术语匹配，即，它的内容明确包括这个术语，则将术语的权重指定给该文档；但是，如果一个文档没有与术语确切匹配，则它将不拥有与这个术语相联系的任何权重。当对付在用户偏爱与文档之间存在模糊相关性的各种情形时，文档与用户概况之间这样的相关性要求有时可能缺乏灵活性。例如，如果用户基于术语概况包括像“Mozilla”和“browser”那样的术语，不包含这样的术语，但包含像“Galeon”或“Opera”那样的其它术语的文档不拥有任何权重，因为，尽管它们实际上是因特网浏览器，但它们不与该概况中的任何现有术语匹配。为了解决没有确切术语匹配地匹配用户兴趣的需要，用户概况可以包括基于类别概况。

图 4A 例示了根据开放目录计划（http:/dmoz.org/）的分层类别图 400。

从图 400 的根层开始，在诸如“Art”、“News”、“Sports”等的几个主题下组织文档。这些主题往往太宽泛，难以描绘用户特定兴趣。因此，进一步将它们划分成更特定的子主题。例如，主题“Art”可以包含像“Movie”、“Music”和“Literature”那样的子主题，和子主题“Music”可以进一步包含像“Lyrics”、“News”和“Reviews”那样的孙主题。注意，每个主题都与像，对于“Art”，1.1，对于“Talk Show”，1.4.2.3，和对于“Basketball”，1.6.1 那样的唯一 CATEGORY_ID 相联系。

尽管图 4A 例示了利用开放目录计划的示范性类别，但也可以使用其它类型的类别。例如，可以通过分析文档或其它信息的各种内容生成围绕概念组织的相关信息的类别来确定类别。换句话说，可以将字或短语映射到与各种概念有关的群集。本领域的普通技术人员应该认识到，有许多不同的方式可以用于将信息分类成有助于确定文档与不同概念的关系的群集。

可以将用户特定兴趣与各种层上的多个类别相联系，每个类别可具有指示该类别与用户兴趣之间的相关程度的权重。类别和权重可以通过分析与用户有关的前面讨论的任何或所有信息来确定。在一些实施例中，通过分析如下数组信息的任何一组或多组确定类别：用户提交的先前搜索查询 201、通过先前搜索查询识别的 URL 203、有关识别文档的一般信息 207（例如，嵌在识别文档中或以其他方式与识别文档相联系的元数据）、用户针对识别文档的活动 209（例如，用户点击一般内容和/或放置内容）、来自识别文档的取样内容 211、有关识别文档的类别信息 213、用户个人信息 215、或它们的任何组合。在一个实施例中，基于类别概况可以利用如图 4B 所示的散列（Hash）表数据结构实现。基于类别概况表 450 包括包含数个记录 460 的表格 455，每个记录包括 USER_ ID 和指向像表格 460-1 那样的另一个数据结构的指针。表格 460-1 可以包括两个列，即，CATEGORY_ID 列 470 和 WEIGHT 列 480。CATEGORY_ID 列 470 包含如图 4A 所示的类别标识号，暗示这个类别与用户兴趣有关，和 WEIGHT 列 480 中的值指示该类别与用户兴趣的相关程度。

基于类别图 400 的用户概况是面向主题的实现。基于类别概况中的项目也可以以其它方式组织。在一个实施例中，可以根据用户识别的文档的格式，譬如，HTML、简单文本、PDF、Microsoft Word 等分类用户偏爱。不同的格式可以具有不同的权重。在另一个实施例中，可以按照识别文档的类型，例如，机构主页、个人主页、研究论文、或新闻组公告分类用户偏爱，每种类

型具有相关权重。可以用于表征用户搜索偏爱的另一种类别是文档来源，例如，与每个文档主机相联系的国家。在又一个实施例中，上述基于类别概况可以共存，每一个反映用户偏爱的一个方面。

除了基于术语和基于类别概况之外，另一种类型的用户概况被称为基于链接概况。如上所述，PageRank 算法基于连接因特网上的各种文档的链接结构。拥有更多指向它的链接的文档往往被指定更高的页面排位，因此，吸引更多搜索引擎的注意。与用户识别的文档有关的链接信息也可以用于推测用户偏爱。在一个实施例中，可以通过分析用户对那些 URL 的访问频率，为用户识别一系列优选 URL。可以根据访问 URL 上的文档时用户花费的时间和用户在 URL 上的滚动活动、和/或其它用户活动（209，图 2）进一步加权每个优选 URL。在另一个实施例中，可以通过分析用户访问不同主机的网页的频率，为用户识别一系列优选主机。当两个优选 URL 与同一个主机有关时，可以结合两个 URL 的权重确定主机的权重。在另一个实施例中，可以通过分析用户访问不同域的网页的频率，为用户识别一系列优选域。例如，对于 finance.yahoo.com，主机是 “finance.yahoo.com”，而域是 “yahoo.com”。

图 5 例示了利用散列表数据结构的基于链接概况。基于链接概况表 500 包括包含数个记录 520 的表格 510，每个记录包括 USER_ID 和指向像表格 510-1 那样的另一个数据结构的指针。表格 510-1 可以包括两个列，即，LINK_ID 列 530 和 WEIGHT 列 540。存储在 LINK_ID 列 530 中的标识号可以与优选 URL 或主机相联系。取代 LINK_ID，可以将实际 URL/主机/域存储在表格中，但是，最好存储 LINK_ID 以便节省存储空间。

一系列优选 URL 和/或主机包括用户直接识别的 URL 和/或主机。一系列优选 URL 和/或主机可以进一步推广到利用本领域的普通技术人员熟知、像合作过滤或文献计量分析那样的方法间接识别的 URL 和/或主机。在一个实施例中，间接识别的 URL 和/或主机包括含有去向/来自直接识别的 URL 和/或主机的链接的 URL 或主机。这些间接识别 URL 和/或主机按它们与用户直接识别的相关 URL 或主机之间的距离加权。例如，当直接识别的 URL 或主机具有 1 的权重时，相距一条链接的 URL 或主机具有 0.5 的权重，相距两条链接的 URL 或主机可具有 0.25 的权重等等。这个过程可以通过降低与最初 URL 或主机的主题无关的链接，例如，到可以用于观看与用户所选 URL 或主机相联系的文档的版权页或万维网浏览器软件的链接的权重得到进一步改进。无关链接可

以根据它们的前后关系或它们的分布来识别。例如，版权链接往往使用特定术语（例如，版权或“保留版权”是版权链接的锚定文本中的常用术语）；和从许多无关网站到一个网站的链接可能暗示这个网站是主题无关的（例如，到 Internet Eeplorer 网站的链接往往包括在无关网站中）。间接链接也可以根据一组主题分类，和可以排除或将低权重指定给具有差别极大主题的链接。

上面讨论的三种用户概况一般彼此互补，因为不同的概况从不同的有利位置描绘用户兴趣和偏爱。但是，这并不意味着一种用户概况，例如，基于类别概况不能扮演通常由另一种用户概况扮演的角色。举例来说，基于链接概况中的优选 URL 或主机往往与特定主题相联系，例如，finance.yahoo.com 是关注金融新闻的 URL。因此，表征用户偏爱的由包含一系列优选 URL 或主机的基于链接概况所取得的也可以至少部分由含有涵盖优选 URL 或主机涵盖的相同主题的一系列类别的基于类别概况取得。

根据列在图 2 中的用户信息构建可以存储在如图 3－5 所示的数据结构中的各种用户概况不是无足轻重的操作。给定用户识别（例如，观看）的文档，文档中的不同术语在揭示文档的主题方面可能具有不同的重要性。一些术语，例如，文档标题可能极其重要，而其它术语可能较不重要。例如，许多文档都包含导航链接、版权声明、弃权声明和可能与文档的主题无关的其它文本。如何有效地选择适当文档、从那些文档中选择内容和从内容中选择术语是计算语言学中富有挑战性的课题。另外，最好使处理的用户信息量最小，以便使构建用户概况的过程在计算上更有效。跳过文档中的较不重要术语有助于准确地将文档与用户兴趣匹配。

段落取样（下面参照图 6 所述）是从可能与用户有关的文档中自动提取内容的过程。这个过程背后的重要观察是，文档中像导航链接、版权声明、弃权声明等那样的较不相关内容往往是相对较短的文本段。在一个实施例中，段落取样寻找文档中长度最长的段落，按长度缩短的顺序处理段落，直到段落的长度是预定阈值以下。可选地，段落取样过程从每个所处理的段落中选择多达某个最大量的内容。如果在文档中没有找到几个适当长度的段落，该过程后退到从文档的其它部分中提取文本，譬如，锚定文本和 ALT 标签。

图 6 是例示段落取样的主要步骤的流程图。段落取样从从文档中消除诸如注释、JavaScript 和风格表单（style sheet）等的预定项目的步骤 610 开始。消除这些项目是因为，当在浏览器上再现时它们通常与文档的可视方

面有关，并且不可能与文档的主题有关。接着，该过程可以在步骤 620 中从长度大于阈值 MinParagraphLength 的每个段落中选择前 N 个字（或 M 个句子），作为取样内容。在一个实施例中，N 和 M 的值被分别选择成 100 和 5。在其它实施例中可以使用其它值。

为了减轻与段落取样过程相联系的计算和存储负担，该过程可以将最大极限，例如，1000 个字强加在来自每个文档的取样内容上。在一个实施例中，段落取样过程首先按长度缩短顺序组织文档中的所有段落，然后，从长度最长的段落开始取样过程。应该注意到，段落的开关和结尾取决于段落在浏览器中的表现，而不是取决于段落的 HTML 表示中不间断文本串的存在。由于这个原因，当确定段落边界时，忽略像有关内嵌链接的命令和有关粗体文本的命令那样的某些 HTML 命令。在一些实施例中，段落取样过程筛选前 N 个字(或 M 个句子)，以便过滤除包括像 “Terms of Service” 或 “Best viewed” 那样的样板术语的那些句子，因为这样的句子通常被认为与文档主题无关。

在取样长度在阈值之上的段落之前，如果取样内容中的字数已经达到最大字数极限，该过程可以停止从文档中取样内容。如果在处理了长度大于阈值的所有段落之后还没有达到最大字数极限，执行可选步骤 630、640、650 和 670。具体地说，该过程将文档标题（630）、非内嵌 HREF 链接（640）、ALT 标签（650）和元标签（670）加入取样内容中，直到它达到最大字数极限。

一旦用户识别的文档得到扫描，可以将取样内容用于通过上下文分析识别一系列最重要（或不重要）术语。上下文分析试图查明预测一组识别文档中的最重要（或不重要）术语的上下文术语。具体地，其寻找前置模式、后置模式、和两者的组合。例如，表达式 “x′s home page” 可以将术语 “x” 识别成用户的重要术语，因此，后置模式 “* home page” 可以用于预测重要术语在文档中的位置，其中，星号 “*” 代表与这个后置模式相配的任何术语。一般说来，通过上下文分析识别的模式通常由重要（或不重要）术语之前的 m 个术语和重要（或不重要）术语之后的 n 个术语组成，其中，m 和 n 两者都大于等于 0，和它们的至少一个大于 0。典型地，m 和 n 小于 5，和当非零时，最好在 1 和 3 之间。取决于其出现频率，模式可以具有指示通过模式识别的术语预期有多重要（或不重要）的相关权重。

根据本发明的一个实施例（图 7A），上下文分析分两个不同阶段，即，训练阶段 701 和操作阶段 703。训练阶段 701 接收和利用一系列预定重要术

语 712、可选一系列预定不重要术语 714、和一组训练文档（步骤 710）。在一些实施例中，不使用一系列预定不重要术语。列表 712 和 714 的来源并不至关重要。在一些实施例中，这些列表 712 和 714 通过如下步骤生成，按照一组规则从一组文档（例如，高页面排位的一组（几千页）网页）中提取字或术语，然后编辑它们，以消除在编辑者看来不属于这些列表的术语。训练文档的来源也不是至关重要的。在一些实施例中，训练文档包含随机或伪随机选择的搜索引擎已知的一组文档。在其它实施例中，按照预定准则从搜索引擎中的文档数据库中选择训练文档。

在训练阶段 701 期间，利用一系列预定重要和不重要术语处理训练文档（步骤 720），以便识别数种上下文模式（例如，前置模式、后置模式、和前置-后置模式），和将权重与每种识别上下文模式相联系。在操作阶段 703 期间，将上下文模式应用于用户识别的文档（步骤 730），以便识别表征用户特定兴趣和偏爱的一组重要术语（步骤 740）。查明和描绘用户兴趣和偏爱通常是持续进行的过程。因此，可以重复操作阶段 703，以便更新以前捕获的一组重要术语。这可以每当用户访问文档时、根据预定时间表、在按照指定准则确定的时间、或不时地完成。类似地，也可以重复训练阶段 701，以便发现新的一组上下文模式和重校与识别上下文模式相联系的权重。

下面是例示训练阶段的一段伪码:

```
for each document in the set{
 for each important term in the document{
  for m=0 to MaxPrefix{
   for n=0 to MaxPostfix{
    Extract the m words before the important
    term and the n words after the important
    term as s;
    Add 1 to ImportantContext(m,n,s);
   }
  }
 }
 for each unimportant term in the document{
  for m=0 to MaxPrefix{
```

```
    for n=0 to MaxPostfix{
     Extract the m words before the
     unimportant term and the n words after
     the unimportant term as s;
     Add 1 to UnimportantContext(m,n,s);
    }
   }
  }
 }
 for m=0 to MaxPrefix{
  for n=0 to MaxPostfix{
   for each value of s{
    Set the weight for s to a function of
    ImportantContext(m,n,s),and
    UnimportantContext(m,n,s);
   }
  }
 }
```

在上面的伪码中，表达式引用前置模式（n=0）、后置模式（m=0）或两者的组合（m>0 和 n>0）。将特定模式的每次出现登记在两个多维数组之一，即，ImportantContext(m,n,s)或 UnimportantContext(m,n,s)上。如果前置、后置或组合模式识别出较多重要术语和较少不重要术语，将这种模式的权重设置成较高的，反之亦然。注意，可以将相同的模式与重要和不重要术语两者相联系。例如，后置表达式“* operating system”可以与一系列预定重要术语 712 中的术语结合在一起用在训练文档 716 中，也可以与一系列预定不重要术语 714 中的术语结合在一起使用。在这种状况下，与后置模式“* operating system”相联系的权重(用表达式 Weight(1,0,“operating system”)表示）将考虑后置表达式与一系列预定重要术语中的术语结合在一起使用的次数，以及后置表达式与一系列预定不重要术语中的术语结合在一起使用的次数。确定上下文模式权重的一个可能公式是:

$$Weight(m,n,s)=Log(ImportantContext(m,n,s)+1)-Log(UnimportantContext(m,n,s)+1)$$

在其它实施例中可以使用其它权重确定公式。

在上下文分析过程的第二阶段中，将加权上下文模式用于识别用户识别的一个或多个文档中的重要术语。参照图 7B，在第一阶段中，计算机系统接收训练数据 750 和创建一组上下文模式 760，每种上下文模式具有相关权重。然后，计算机系统将该组上下文模式 760 应用于文档 780。在图 7B 中，在文档 780 内找到的以前识别上下文模式被加亮了。识别与上下文模式相联系的术语 790，和每个这样的术语获得基于与上下文模式相联系的权重的权重。例如，术语“Foobar”在文档中出现过两次，与两种不同模式，即，前置模式“Welcome to *”和后置模式“* builds”相联系，和指定给“Foobar”的权重 1.2 是两种模式的权重 0.7 和 0.5 之和。其它识别术语“cars”具有 0.8 的权重，因为匹配前置模式“world's best *”具有 0.8 的权重。在一些实施例中，利用对数变换计算每个术语的权重，其中，最后权重等于 log(initial weight+1)。两个术语“Foobar”和“cars”可能未在训练数据 750 中和可能用户以前从未遇到过。不过，上述的上下文分析方法识别这些术语和将它们加入用户基于术语概况中。因此，即使那些术语未包括在预定术语数据库中，上下文分析也可以用于发现与用户兴趣和偏爱相联系的术语。

注意，上下文分析的输出可以直接用于构建用户基于术语概况。另外，它也可以用于构建其它类型的用户概况，譬如，用户基于类别概况。例如，可以分析一组加权术语和将它们分类成涵盖不同主题的数个类别，并且，可以将那些类别加入用户基于类别概况中。

在对由用户识别或为用户识别的一组文档进行上下文分析之后，所得的一组术语和权重可能占据比分配给每个用户基于术语概况大的存储量。此外，该组术语和相应权重可能包括权重比该组中的其它术语小得多的一些术语。因此，在一些实施例中，在上下文分析结束时，通过消除权重最低的术语整理该组术语和权重，（A）以便基于术语概况占据的总存储量满足预定极限，和/或（B）以便消除权重太低的术语，或被认为不表示用户搜索偏爱和兴趣、按预定准则定义、与较老术语相对应的术语。在一些实施例中，也可以将类似的整理准则和技术应用于基于类别概况和/或基于链接概况。

如上所述，可以根据参照图 2 所述的信息创建基于类别概况。例如，可以将以前提交的查询术语与特定信息类别相联系。用户概况引擎可以分析用户提交的先前搜索查询，以确定用户可能感兴趣的特定信息类别和它们各自的权重。这样的用户概况引擎可以分析参照图 2 所述的任何信息源。

在一些实施例中，每当用户进行搜索和从搜索结果中选择至少一个文档进行下载或观看时就更新用户概况。在一些实施例中，搜索引擎随时间构建用户识别（例如，通过从搜索结果中选择文档）的文档的列表，和在预定时间（例如，当列表达到预定长度，或经过了预定时间时），进行概况更新。当进行更新时，生成新的概况数据，和将新的概况数据与用户的以前生成概况数据合并。在一些实施例中，将新概况数据指定成比以前生成概况数据更重要，从而使系统能够随用户搜索偏爱和兴趣的变化迅速调整用户概况。例如，可以在与新概况数据合并之前自动按比例降低以前生成概况数据中的项目的权重。在一个实施例中，存在与概况中的每个项目相联系的日期，和根据其年龄加权概况中的信息，较老的项目获得比它们较新时更低的权重。在其它实施例中，不将新的概况数据指定成比以前生成概况数据更重要。

段落取样和上下文分析方法可以独立使用或结合在一起使用。当结合在一起使用时，段落取样的输出用作上下文分析方法的输入。

并且，应该注意到，上述用于创建用户概况的方法，例如，段落取样和上下文分析也可以作为确定候选文档与用户偏爱的相关性的手段。的确，搜索引擎的首要使命是根据用户提交的搜索查询，以及用户的用户概况识别与用户偏爱最相关的一系列文档。图 8 例示了可以用于从多个角度存储有关文档与用户概况的相关性的信息的几个示范性数据结构。对于通过各自 DOC_ID 识别的每个候选文档，基于术语文档信息表 810 包括多对术语和它们的权重，基于类别文档信息表 830 包括数个类别和相关权重，和基于链接文档信息表 850 包括一组链接和相应权重。

三个表格（810、830 和 850）每一个的最右列存储利用一种特定用户概况评估文档时文档的排位（即，计算分）。用户概况排位可以通过组合与文档相联系的项目的权重来确定。例如，基于类别或基于主题概况排位可以按如下计算。用户可能偏爱权重为 0.6 的有关科学的文档，而不喜欢权重为 0.2 的有关商业的文档。因此，当科学文档与搜索查询匹配时，使它的权重比商业文档高。一般说来，文档主题分类可能不是排它性的。候选文档可能被分

类成概率为 0.8 的科学文档和概率为 0.4 的商业文档。基于链接概况排位可以根据分配给用户 URL、主机、域等的相对权重、和基于链接概况的偏爱计算。在一个实施例中，基于术语概况排位可以利用已知技术，譬如，术语频率－逆文档频率（TF-IDF）确定。术语的术语频率是术语出现在文档中的次数的函数。逆文档频率是一批文档内出现术语的文档的数量的逆函数。例如，像“the”那样的非常普通术语出现在许多文档中，因此，被指定成相应低逆文档频率。

当搜索引擎响应搜索查询生成搜索结果时，按照搜索查询，将查询分 QueryScore 指定给满足查询的候选文档 D。然后，通过文档 D 的页面排位 PageRank 调整这个查询分，生成表达成如下的通用分 GenericScore:

GenericScore=QueryScore*PageRank。

如果用户的兴趣或偏爱显著偏离随机冲浪者的兴趣或偏爱，这个通用分可能未适当反映文档 D 对特定用户 U 的重要性。根据这里称为 TermScore 的文档 D 的内容与用户 U 的基于术语概况之间的关联性、这里称为 Category-Score 的与文档 D 相联系的一个或多个类别与用户 U 的基于类别概况之间的关联性、和这里称为 LinkScore 的文档 D 的 URL 和/或主机与用户 U 的基于链接概况之间的关联性，可以通过一组概况排位准确表征文档 D 与用户 U 的相关性。因此，可以将是文档通用分和用户概况分两者函数的个性化排位指定给文档 D。在一个实施例中，这个个性化分可以表达成:

PersonalizedScore=GenericScore*(TermScore
+CategoryScore+LinkScore)

图 9A 和 9B 代表都在像如图 1 所示的网络环境 100 那样的客户机－服务器网络环境下实现的两个实施例。在如图 9A 所示的实施例中，搜索引擎 104 在步骤 910 中接收特定用户提交的来自客户机 102 的搜索查询。作出响应，搜索引擎 104 可以在步骤 915 中可选地生成查询策略（例如，将搜索查询标准化成用于进一步处理的适当形式，和/或可以按照预定准则修正搜索查询，以便自动扩大或缩小搜索查询的范围）。在步骤 920 中，搜索引擎 104 将搜索

查询（或查询策略，如果生成的话）提交给内容服务器 106。内容服务器在步骤 920 中识别与搜索查询匹配的一系列文档，每个文档具有取决于文档页面排位和搜索查询的通用分。一般说来，所有三个操作（步骤 910、915 和 920）都由处在网络环境 100 的服务器方的搜索引擎系统 107 执行，对于实现这前三个步骤之后的操作，存在两种选择。

在应用服务器方实现的一些实施例中，将用户的标识号嵌入搜索查询中。根据用户的标识号，用户概况服务器 108 在步骤 925 中识别用户的用户概况。从步骤 930 开始，用户概况服务器 108 或搜索引擎 104 分析在步骤 920 中识别的每个文档，以确定它与用户概况的相关性，在步骤 935 中为识别的文档创建概况分，然后，在步骤 940 中将作为文档通用分和概况分函数的个性化分指定给文档。在步骤 942 中，用户概况服务器 108 或搜索引擎 104 检验这是否是一系列识别文档中的最后一个。如果不是，该系统处理该列表中的下一个文档。否则，根据它们的个性化分重新排序一系列文档，然后，将它们发送到用户从那里提交搜索查询的相应客户机。

除了在步骤 920 之后，将识别文档发送到用户从那里提交查询的相应客户机之外，利用客户机方实现的实施例与服务器方实现类似。这个客户机存储用户的用户概况和负责根据用户概况重新排序文档。因此，这种客户机方实现可以减轻服务器的工作负担。并且，由于客户机方实现没有隐私方面的问题，用户可能更愿意提供隐私信息来定制搜索结果。但是，客户机方实现的严重局限性在于，由于有限的网络带宽，只能将有限数量的文档，例如，顶部 50 个文档（利用通用排位确定）发送到客户机进行重新排序。相反，服务器方实现能够将用户概况应用于与搜索查询匹配的数量大得多的文档，例如，1000 个。因此，客户机方实现可能剥夺了用户访问通用排位相对较低，但个性化排位很高的那些文档的权利。

图 9B 例示了另一个实施例。与在向搜索引擎 104 提交搜索查询之前不个性化搜索查询、描绘在图 9A 中的实施例不同，根据用户的用户概况调整通用查询策略（步骤 965），以便创建个性化查询策略。例如，可以将来自用户概况的相关术语以相关权重加入搜索查询中。个性化查询策略的创建可以在系统的客户机方或服务器方进行。这个实施例避免了前一个实施例所面临的网络带宽制约。最后，搜索引擎 104 将个性化查询策略提交给内容服务器 106（步骤 970），因此，已经按文档的个性化排位排序了内容服务器返回的搜索

结果（步骤 975）。

可以将拥有相关兴趣的一群用户的概况结合在一起形成群体概况，或者，可以根据群体中的用户识别的文档形成个体概况。例如，几个家庭成员可能使用相同的计算机将搜索查询提交给搜索引擎。如果搜索引擎用单个用户标识符标记计算机，“用户”将是用户的整个家庭，和用户概况将代表各个家庭成员的搜索偏爱的组合或混合。群体中的各个用户可以有选择地拥有将这个用户与其它群体成员区分开的分立用户概况。在操作过程中，根据群体概况，或当用户也拥有分立用户概况时，根据群体概况和用户的用户概况排行群体中的用户的搜索结果。

用户可能剧烈地转移他的兴趣，使他的新兴趣和偏爱几乎不像他的用户概况，或用户可能临时对新的主题感兴趣。在这种情况下，根据描绘在图 9A 和 9B 中的实施例生成的个性化搜索结果可能比按照搜索结果中的文档的通用排位排行的搜索结果更不利。另外，提供给用户的搜索结果可能在顶部列出文档中不包括新网站，因为用户的概况倾向于提高用户过去访问过的老网站（即，用户观看或下载过网页的老网站）的权重。

为了降低用户偏爱和兴趣的变化造成的影响，可以将个性化搜索结果与通用搜索结果合并。在一个实施例中，交织通用搜索结果和个性化搜索结果，为通用搜索结果保留搜索结果列表的奇位置（例如，1、3，5 等），和为个性化搜索结果保留搜索结果列表的偶位置（例如，2、4、6 等），或反过来。最好，通用搜索结果中项目不重复列在个性化搜索结果中的项目，反之亦然。更一般地说，使通用搜索结果与个性化搜索结果混合或交织在一起，以便向用户展示的搜索结果中的项目包括通用搜索结果和个性化搜索结果两者。

在另一个实施例中，通过用户概况的置信水平进一步加权个性化排位和通用排位。置信水平考虑诸如获取了多少有关用户的信息、当前搜索查询与用户概况有多匹配、用户概况有多老等的因素。如果只可获得非常短的用户历史，可以将相应低的置信值指定给用户的概况。识别文档的最后分数可以确定成：

$$FinalScore=ProfileScore*ProfileConfidence+GenericScore*(1-ProfileConfidence)。$$

当通用结果和个性化结果混合时，可以根据概况置信度调整个性化结果部分，例如，当置信度低时，只利用一个个性化结果。

有时，多个用户可能共享，例如，公共图书馆中的机器。这些用户可能拥有不同的兴趣和偏爱。在一个实施例中，用户可能显式地登录到服务上，因此，系统知道他的身份。可替代地，可以根据他们访问的项目或他们访问模式的其它特性自动识别不同的用户。例如，不同的用户可能以不同方式移动鼠标，不同地打字，和使用不同的应用和那些应用的不同特征。根据客户机和/或服务器上的事件的资料库，可以创建识别用户，然后，利用那种识别选择适当“用户”概况的模型。在这样的环境下，“用户”实际上可以是拥有有些相似计算机使用模式、兴趣等的一群人。

参照图 10，个性化搜索引擎系统 1000 通常包括一个或多个处理单元（CPU）1002、一个或多个网络或其它通信接口 1010、存储器 1012、和互连这些部件的一条或多条通信总线 1014。可选地，系统 100 可以包括用户接口 1004，例如，显示器 1006 和键盘 1008。存储器 1012 可以包括高速随机访问存储器，也可以包括像一个或多个磁盘存储设备那样的非易失性存储器。存储器 1012 可以包括远离中央处理单元 1002 的大容量存储器。存储器 1012 最好存储:

- 操作系统 1016，包括管理各种基本系统服务和执行硬件相关任务的例程;
- 网络通信模块 1018，用于通过一个或多个通信网络（有线或无线），譬如，因特网、其它广域网、局域网、城域网等将系统 1000 与其它服务器或计算机连接;
- 系统初始化模块 1020，用于初始化系统 1000 的适当操作所需的存储在存储器 1012 中的其它模块和数据结构;
- 搜索引擎 1022，用于处理搜索查询，根据搜索查询和用户概况识别和排序搜索结果;
- 用户概况引擎 1030，用于收集和处理像在图 2 中识别的用户信息那样的用户信息，和创建和更新表征用户搜索偏爱和兴趣的用户的用户概况;
- 数据结构 1040、1060 和 1080，用于存储数个用户概况。

搜索引擎 1022 可以进一步包含:

- 通用排位模块（或指令）1024，用于处理用户提交的搜索查询，识别

与查询匹配的一系列文档，和无需参照用户特定信息地将通用排位指定给每个识别文档；

• 用户概况排位模块（或指令）1026，用于将通用排位模块 1024 识别的数个文档的每一个与用户的用户概况相关联，和将指示文档与用户搜索偏爱和兴趣的相关性的概况排位指定给文档；和

• 排位混合模块（或指令）1028，用于将识别文档的通用排位和概况排位组合成个性化排位，和根据它们的个性化排位重新排序一系列文档。

在一些实施例中，这些模块 1024、1026、1028 可以在单个例程内实现，或在驻留在单个软件模块内的一组例程中实现。

用户概况引擎 1030 可以进一步包含：

• 用户信息收集模块 1032，用于收集和分类列在图 2 中的各种用户信息；

• 文档内容提取模块 1034，用于从用户识别的文档中选择和提取内容，以便利用像段落取样（如上所述）那样的技术识别与用户兴趣有关的内容；和

• 上下文分析模块 1036，用于分析文档内容提取模块 1034 提取的内容，以便识别表征用户搜索偏爱的术语。

寄存用户概况的每个数据结构可以进一步包含：

• 数据结构 1042、1062 或 1082，用于存储基于术语用户概况；

• 数据结构 1044、1064 或 1084，用于存储基于类别用户概况；和

• 数据结构 1046、1066 或 1086，用于存储基于链接用户概况。

按照用户概况排序放置内容

可以向搜索服务、电子邮件服务、和通过因特网或其它广域网提供的各种其它服务的用户显示放置内容。下文将描述排序放置内容（例如，在浏览器窗口或用户观看的其它应用窗口内），以便（A）最大化或至少提高用户有兴趣观看放置内容的机会，或（B）最大化或至少提高到放置内容提供者的收益流，或（C）优化或至少提高与放置内容的传送和排序相联系的度量的系统和方法。首先，将针对向搜索引擎的用户传送放置内容，描述该系统和方法，然后，将描述该系统和方法用于其他因特网服务。

当响应搜索查询将搜索结果返还给用户时，常常也返还某些放置内容。放置内容通常是广告形式，但也可以是与搜索查询有关或与发送给用户的文档有关的任何类型的内容。尽管为了例示起见，如下的描述使用广告内容，

但通过本发明的一些实施例可以设想出内容提供者竞争位置或支付位置费的任何类型内容。在针对文档资源库运行用户的搜索查询的同时，可以针对广告（ad）资源库运行该搜索查询。搜索广告资源库返回的广告（例如，关键字与搜索查询的至少一个术语匹配的广告）通常按每个广告的分数排序。该分数基于点击通过率（click through rate, CTR）乘以出价（例如，标价）。向用户展示分数最高的广告。在一些实施例中，内容提供者可以提供与同一出价相联系的多个相似广告。在这种情况下，可以以随机方式，或按任何其它顺序向用户展示各种广告。例如，如果内容提供者提供了对术语“hat”实施单次投出价一群（三个）广告，每当该群广告具有可以包括在一组搜索结果中的足够高分数时，选择（例如，随机地，或按循环顺序）该群中的三个广告之一，并且，向用户展示。

广告商可以通过，例如，广告商将出价放在某些搜索术语或短语上的拍卖对不同关键字或概念投标。例如帆船的帆的制造者可以对关键字“spinnaker”投标，以便当那个术语出现在搜索查询中时，广告商的广告将出现在要向用户展示的一系列潜在广告中。如果该广告的分数足够高，就向用户展示该广告。如上所述，分数基于 CTR 乘以出价。然后，广告商根据它的出价和根据在特定结帐间隔内对广告的击中次数（例如，出价乘以击中次数）支付广告费。在一些实施例中，拍卖可能具有“荷兰式拍卖”的特点，在这种情况下，广告商为特定广告支付的费用可以是乘以特定结帐间隔内的击中次数的修正或降低出价。

提高广告的 CTR 是提高广告分数的一种方式。提高 CTR 可以通过，例如，展示比其它广告更吸引用户的广告来达到。可替代地，广告商可以选择提高他或她对与广告相联系的关键字或短语的出价，以便提高广告分数。并且，理所当然，广告商可以既提高广告的 CTR 又提高它对与广告相联系的关键字的出价。在一些实施例中，广告的 CTR 等于点击在广告上的次数除以留下印象的次数，即，向用户展示广告的次数。新的广告通常不存在有用的 CTR，因为广告的留下印象次数太低，难以使 CTR 的值成为广告对用户的吸引力的可靠指示。在这样的情况下（例如，当广告留下印象少于 1000 次时），由系统提供初始 CTR。广告的初始 CTR 可以是像平均 CTR 值那样的默认值。可替代地，初始 CTR 可以根据同一广告商的其它广告的 CTR 来选择，或可以基于与所涉及的广告存在既定关系的一些其它广告组的 CTR。

最好提高将用户感兴趣的广告展示给用户的可能性。于是，在某个方面与用户概况有关的广告是展示的较好候选者。做到这一点的一种方式是根据广告与用户概况的相似性修正广告分数。回头参看更广义的术语“放置内容”，图 11 例示了与搜索结果一起提供放置内容的一个实施例。

首先，在，例如，搜索引擎上接收搜索查询（1102）。搜索查询可以通过，例如，包括提交搜索查询的客户机计算机或客户机例程的标识符，识别提交搜索查询的用户。可替代地，用户的标识符可能由于过去登录到服务、cookie、或其它适用方法而已知。从用户概况的数据库或资源库中获取用户概况（1104）。在一个实施例中，用户的概况是类别概况。虽然如下的描述使用类别概况，但本领域的普通技术人员容易认识到，这里的概念可以应用于其它类型的概况。在搜索引擎处理搜索查询，以便获取搜索结果（1106）的同时（1106），放置内容服务器识别与搜索查询匹配或相关的一个或多个放置内容项（这里称为潜在放置内容）（1108）。在其它实施例中，放置内容服务器可以根据正在将什么文档（不管作为搜索结果还是特定请求文档）提供给用户来提供放置内容。在那个实施例中，放置内容服务器确定哪个放置内容与正在向用户展示的文档有关。在其它实施例中，放置内容服务器可以提供基于展示的一个或多个文档的内容的放置内容作为搜索结果。

每个潜在放置内容拥有与之相联系的概况。在一个实施例中，该概况具有包含类别和权重对的类别概况的形式。该概况可以通过，例如，从放置内容中提取关键术语，将它们与各种类别相联系和指定各自权重来创建。

对于每个潜在放置内容，将潜在放置内容的概况与用户的概况相比较（1110）。将用户的概况与放置内容概况相比较以便获得相似性分数。然后，将相似性分数用于修正放置内容的排行。如果将每个概况当作一个向量，那么，本领域的普通技术人员可以识别出各种比较概况的数学方法。例如，相似性分数可以通过取出用户概况中的每个类别，确定它与放置内容概况的每个类别之间的数学距离，然后乘以各自权重来确定。表示这种计算的一种方式是通过如下公式：

$$\text{similarity score}=\sum_{i=0}^{n-1}\sum_{j=0}^{m-1}\text{distance}(\text{category}(i),\text{category}(j))*\text{weight}(i)*\text{weight}(j),$$

其中，n 代表用户概况中的类别数和 m 代表放置内容概况中的类别数；distance(category(i),category(j))代表 category(i)与 category(j)之间的数学距离；和 weight(i)和 weight(j)分别代表与 category(i)和 category(j)相联系的权重。

更一般地，代表相似性分数的计算的另一种方式是：

similarity score=function(user profile, content profile),

其中，“function”是用户概况和特定放置内容项的内容概况的任何适用函数。当用户和内容概况是类别概况时，相似性分数的计算可以表示成：

similarity score=function(user profile categories, user profile weights, content profile categories, content profile weights),

其中，“function”是用户概况类别和权重的向量和内容概况类别和权重的向量的任何适用函数。与如上所示的双求和计算不同的计算相似性分数的稍微更具体例子是：

$$\text{similarity score}=\sum_{i}\text{Max}_{j}\,(\text{function}(\text{category}(i),\text{category}(j),\ \text{weight}(i),\text{weight}(j)),$$

其中，“Max_j”代表对于 j 的所有有效值，该函数的最大值，和“function”代表用户和内容概况类别和权重的任何适用函数。

在一些实施例中，将相似性分数归一化到特定范围，从而得出一个比例因子。例如，可以将相似性分数归一化成在 0 到 1 或 0 到 2 的闭区间内。较高的相似性分数表示该概况比其比较导致较低相似性分数的概况更密切相关。在一些实施例中，将归一化相似性分数用作比例因子。在其它实施例中，通过按照比例因子映射函数或比例因子查用表将相似性分数或归一化相似性分数映射到相应比例因子，确定比例因子。

在一个实施例中，将一组（N 个）预定比例因子（有时称为亚因子）存

储在比例因子查用表中，每个比例因子对应于各自范围的相似性分数值。在这个示范性实施例中，N 是大于 1 的整数，和最好大于 3。例如，通过将相似性分数乘以或除以预定数，将结果上舍入或下舍入成最接近整数生成箱号，然后，通过将箱号用作进入比例因子查用表的索引，将所得箱号映射到比例因子，将特定放置内容的相似性分数映射到“箱子”中。比例因子的范围可以随实现而不同。

比例因子映射函数或比例因子查用表的使用使联系相似性分数与比例因子变得非常灵活。例如，可以创建向下调整相似性分数非常低的放置内容，以及相似性分数非常高的放置内容的 CTR 的比例因子映射函数或比例因子查用表。在一些实施例中，与最大相似性分数相联系的比例因子小于与中点相似性分数相联系的比例因子，其中，中点可以是相似性分数的平均值或中值。可替代地，中点也可以是最小和最大相似性分数之间的任何识别点。在一些实施例中，与最大相似性分数相联系的比例因子大于与中点相似性分数相联系的比例因子，但小于与比例因子映射函数或比例因子查用表相联系的最大比例因子。当随着相似性分数的值从最小分数变到最大分数观查比例因子映射函数时，比例因子通常最初从与最小分数相联系的低值开始增大，直到达到比例因子峰值，然后减小，直到相似性分数达到最大值。

在一些实施例中，按照联系相似性分数与点击通过率的统计信息确定与相似性分数相对应的比例因子。具体地，用户的点击通过率可以统计地与用户和放置内容项的相似性分数相关联。例如，可以通过收集有关留下印象、击中和与每次留下印象和击中相联系的相似性分数的数据，为相似性分数的一组 N 个范围中的每个范围确定独立点击通过率。根据那些点击通过率，可以生成一组 N 个比例因子，以便存储在比例因子查用表中。可替代地，收集的统计信息可以用于通过，例如，曲线拟合技术生成比例因子映射函数。

在一些实施例中，将放置内容的 CTR 乘以每个识别放置内容的各自比例因子提供修正 CTR，以反映用户对放置内容感兴趣的可能性提高了（图 11 的 1112）。更具体地说，按如下计算与搜索查询匹配的每个放置内容（例如，至少含有一个与搜索查询的术语匹配的关键字）的分数:

$$score=scaling\ factor \times CTR \times bid。$$

然后，根据它们各自的分数排行或排序放置内容项（1114），和例如，通过发送给用户计算机上的浏览器应用程序，将具有最高分的放置内容项提供给用户（1116）。在一些实施例中，可以将具有最高分 H（其中，H 是大于 1 的整数）的放置内容项与对数据库进行搜索查询获得的搜索结果（有时称为基本搜索结果）合并（1118）。例如，当放置内容项包含广告时，可以在基本搜索结果的上面、下面和/或旁边显示具有最高分的一个或多个广告。

在一些实施例中，放置内容项的分数基于利用用户概况和出价得出的相似性分数，但不基于点击通过率。例如，在一些实施例中，可能得不到放置内容项的点击通过率。结果是，在这样的实施例中，不会出现动作 1112，或用不同的分数调整或分数计算动作取代动作 1112。

在一些其它实施例中，放置内容项的分数基于利用用户概况和点击通过率得出的相似性分数，但不基于出价。并且，在另一些其它实施例中，放置内容项的分数基于利用用户概况得出的相似性分数，但那些分数不基于出价或点击通过率。当放置内容分数考虑到用户概况，但未考虑到出价时，与放置内容项的其它排序的潜在经济效益无关地针对用户有可能感兴趣的放置内容优化或改善放置内容的排序。

上述的系统和方法也可以用在除搜索引擎系统之外的其它系统中。例如，在电子邮件系统中，或在通过因特网或其它广域网向用户或客户提供显示文档或其它内容的服务的几乎任何其它系统中，也可以选择和向用户显示放置内容。放置内容可以根据与匹配显示文档或一组文档的内容的放置内容相联系的关键字来选择，或可以基于其它选择准则。然后，如上所述，根据用户概况和所选放置内容项的概况排序所选放置内容项。

为了说明起见，上面参照特定实施例对本发明作了描述，但是，上面的例示性讨论是不完全的，也不打算使本发明局限于所公开的确切形式。可以按照上面讲述的内容作出许多修正和改变。选择和描述这些实施例是为了最佳地说明本发明的原理和它的实际应用，从而使本领域的其它普通技术人员能够最佳地利用本发明和为了适合特定应用而作出各种修正的各种实施例。

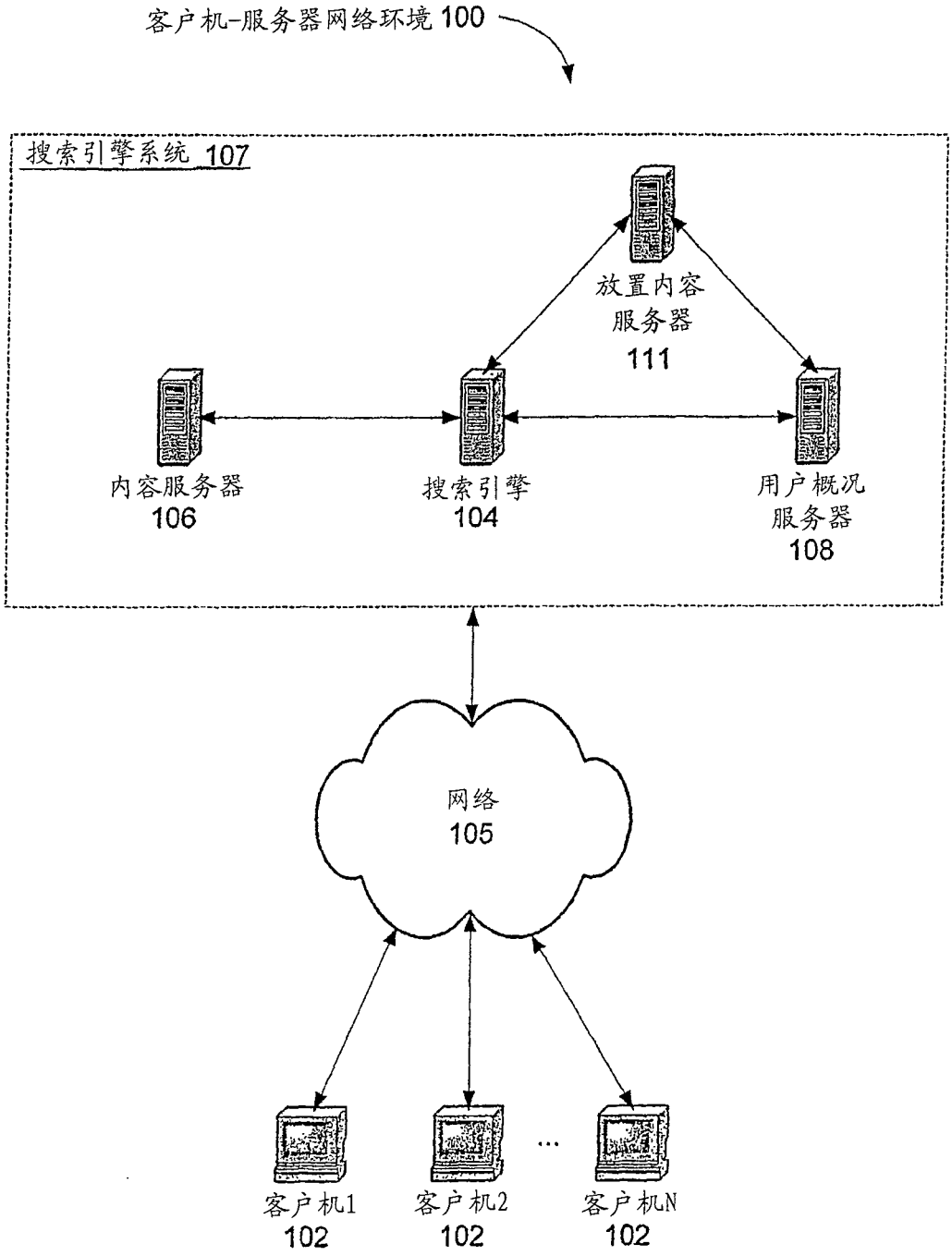

客户机-服务器网络环境 100
搜索引擎系统 107
放置内容
服务器
111
内容服务器
106
搜索引擎
104
用户概况
服务器
108
网络
105
客户机1
102
客户机2
102
...
客户机N
102

图 1

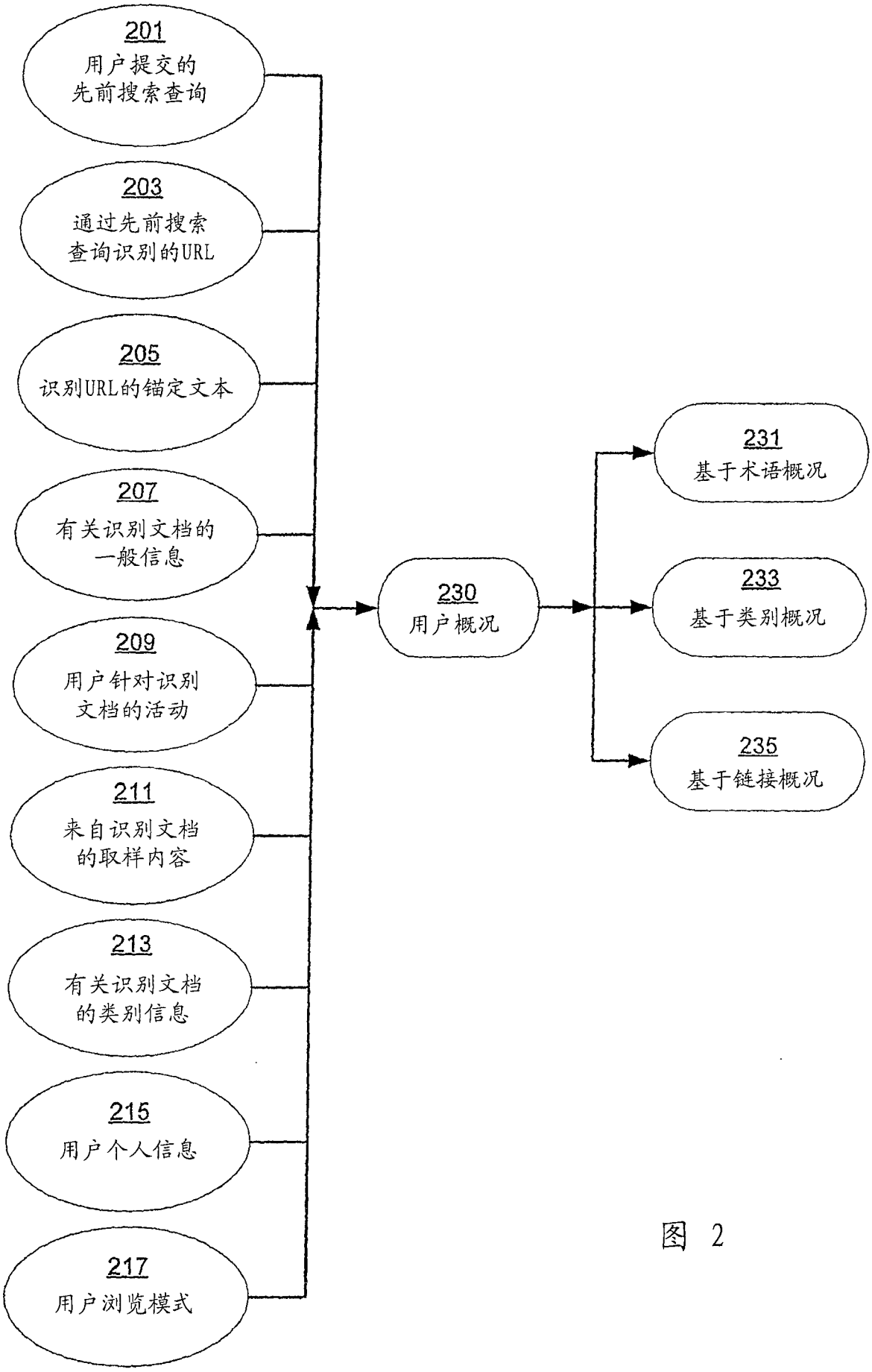

201
用户提交的
先前搜索查询
203
通过先前搜索
查询识别的URL
205
识别URL的锚定文本
207
有关识别文档的
一般信息
209
用户针对识别
文档的活动
211
来自识别文档
的取样内容
213
有关识别文档
的类别信息
215
用户个人信息
217
用户浏览模式
230
用户概况
231
基于术语概况
233
基于类别概况
235
基于链接概况

图 2

基于术语概况表 300

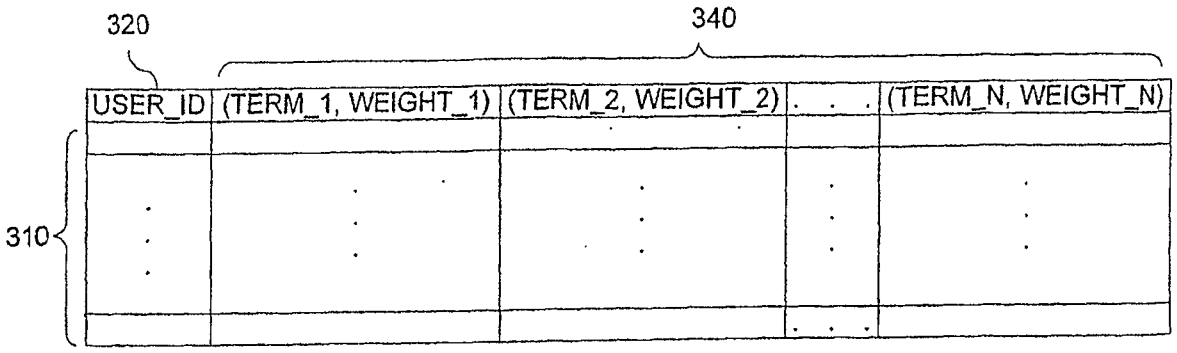


图 3

基于链接概况表 500

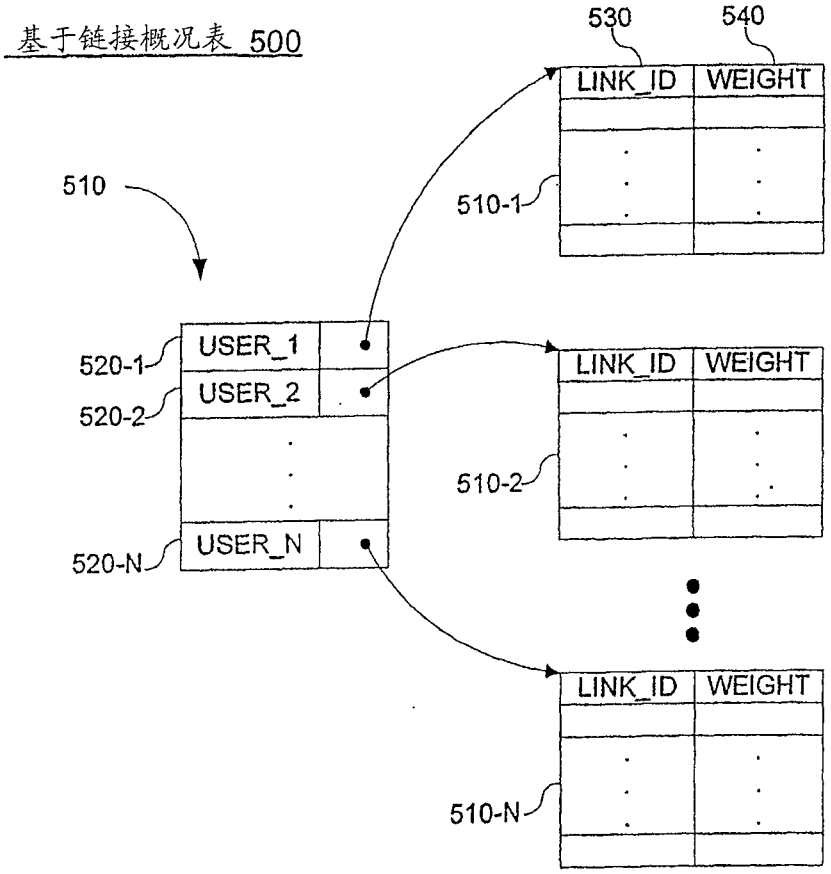


图 5

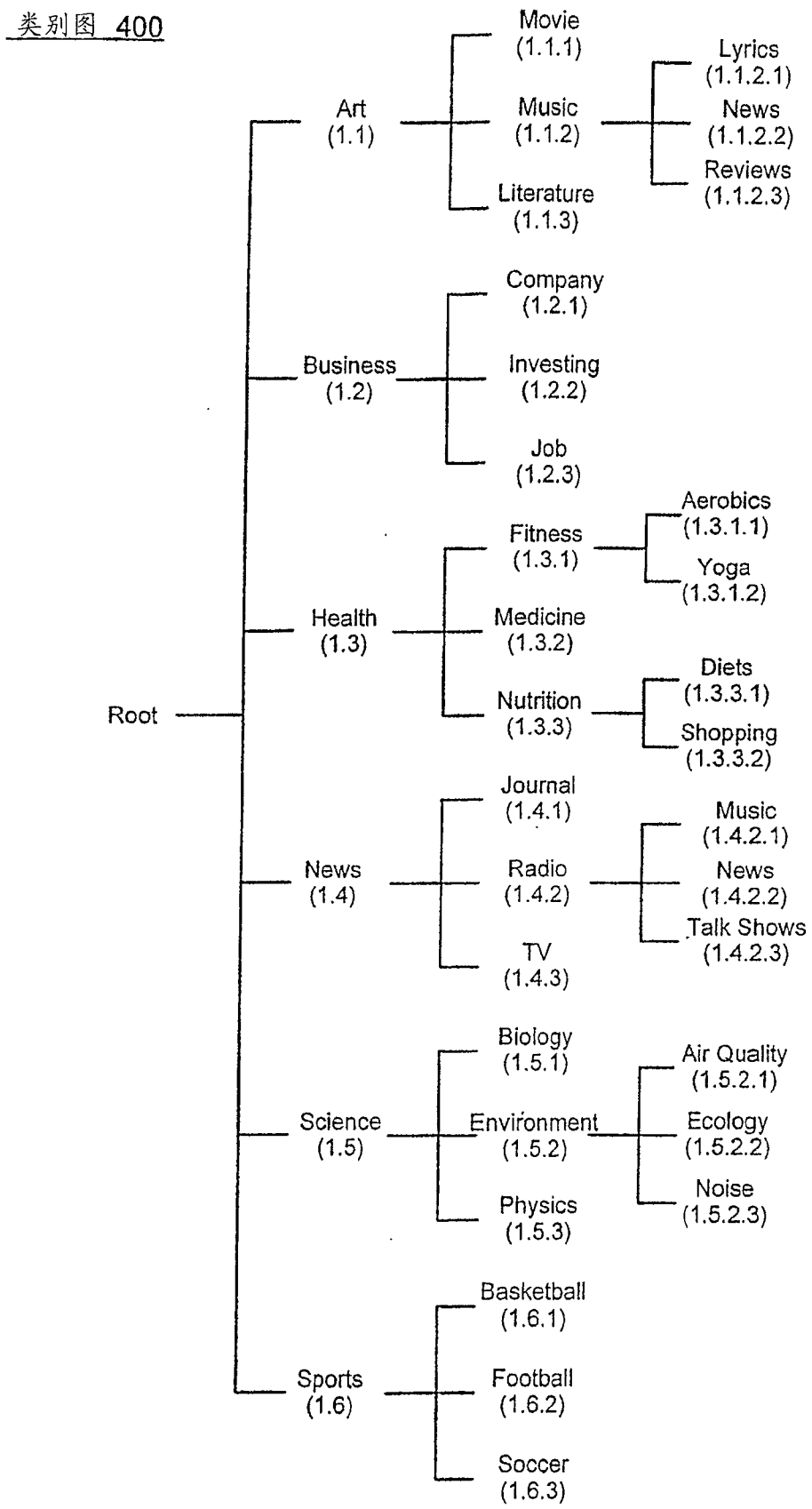

类别图 400
Root
Art (1.1)
Movie (1.1.1)
Music (1.1.2)
Lyrics (1.1.2.1)
News (1.1.2.2)
Reviews (1.1.2.3)
Literature (1.1.3)
Business (1.2)
Company (1.2.1)
Investing (1.2.2)
Job (1.2.3)
Health (1.3)
Fitness (1.3.1)
Aerobics (1.3.1.1)
Yoga (1.3.1.2)
Medicine (1.3.2)
Nutrition (1.3.3)
Diets (1.3.3.1)
Shopping (1.3.3.2)
News (1.4)
Journal (1.4.1)
Radio (1.4.2)
Music (1.4.2.1)
News (1.4.2.2)
Talk Shows (1.4.2.3)
TV (1.4.3)
Science (1.5)
Biology (1.5.1)
Environment (1.5.2)
Air Quality (1.5.2.1)
Ecology (1.5.2.2)
Noise (1.5.2.3)
Physics (1.5.3)
Sports (1.6)
Basketball (1.6.1)
Football (1.6.2)
Soccer (1.6.3)

图 4A

基于类别概况表 450

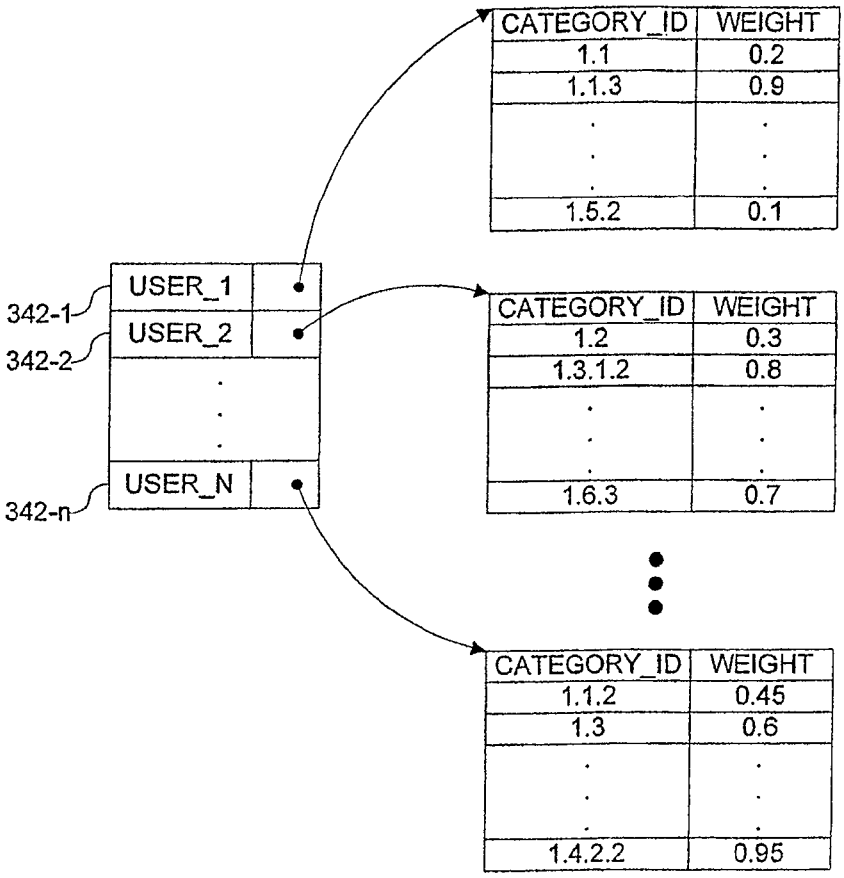


图 4B

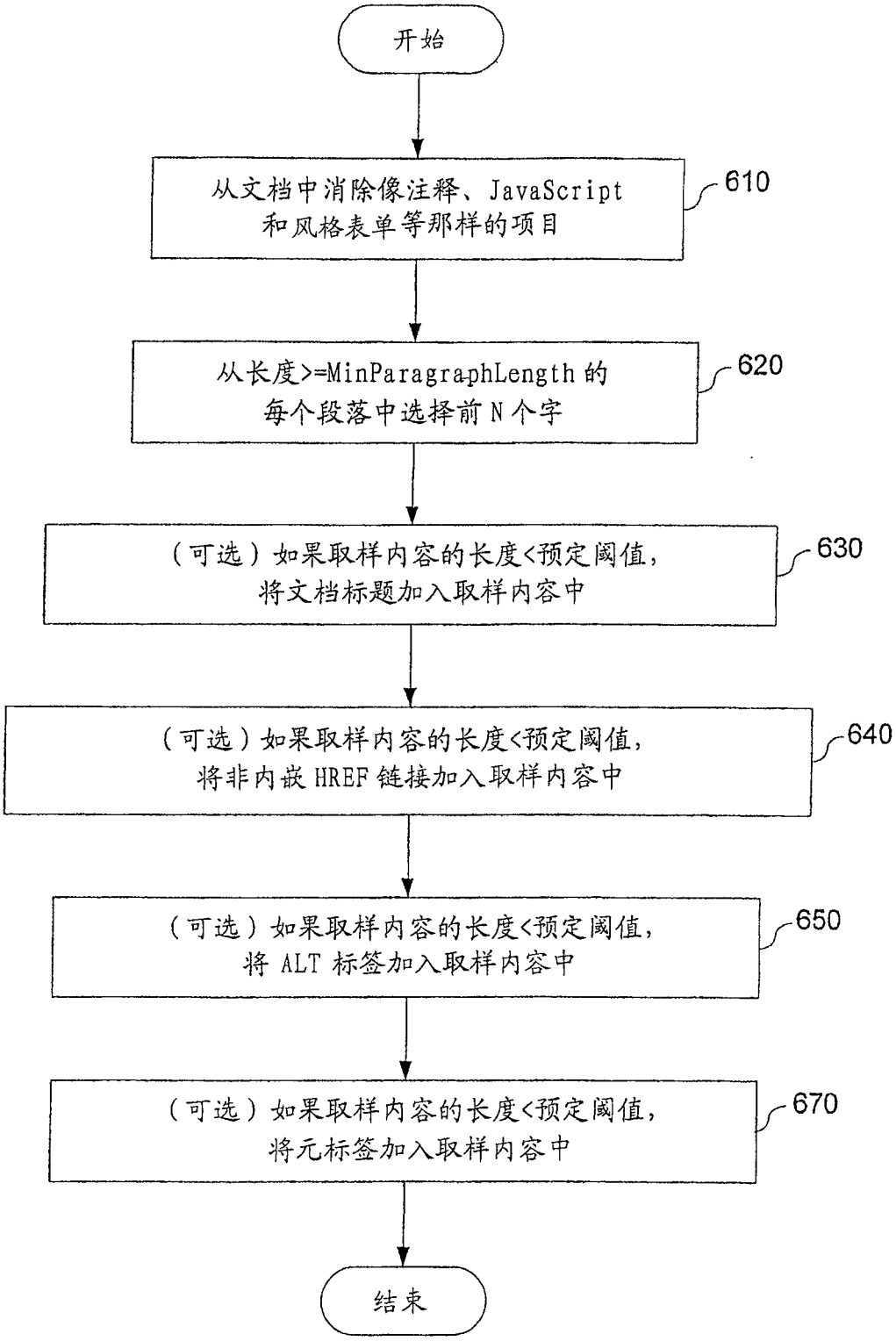

开始
从文档中消除像注释、JavaScript
和风格表单等那样的项目
610
从长度>=MinParagraphLength 的
每个段落中选择前 N 个字
620
（可选）如果取样内容的长度<预定阈值，
将文档标题加入取样内容中
630
（可选）如果取样内容的长度<预定阈值，
将非内嵌 HREF 链接加入取样内容中
640
（可选）如果取样内容的长度<预定阈值，
将 ALT 标签加入取样内容中
650
（可选）如果取样内容的长度<预定阈值，
将元标签加入取样内容中
670
结束

图 6

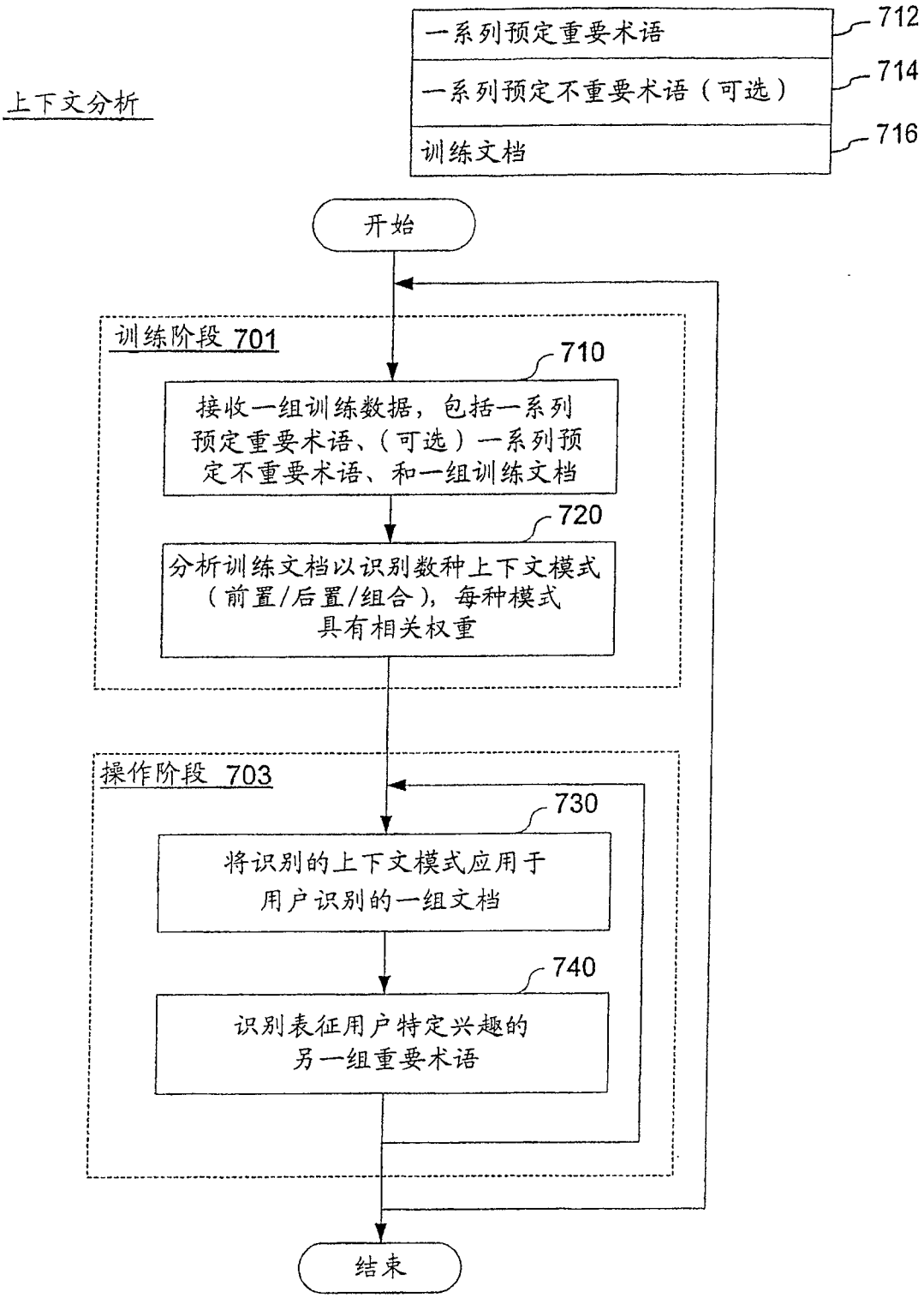

上下文分析
一系列预定重要术语
712
一系列预定不重要术语（可选）
714
训练文档
716
开始
训练阶段 701
710
接收一组训练数据，包括一系列预定重要术语、（可选）一系列预定不重要术语、和一组训练文档
720
分析训练文档以识别数种上下文模式（前置/后置/组合），每种模式具有相关权重
操作阶段 703
730
将识别的上下文模式应用于用户识别的一组文档
740
识别表征用户特定兴趣的另一组重要术语
结束

图 7A

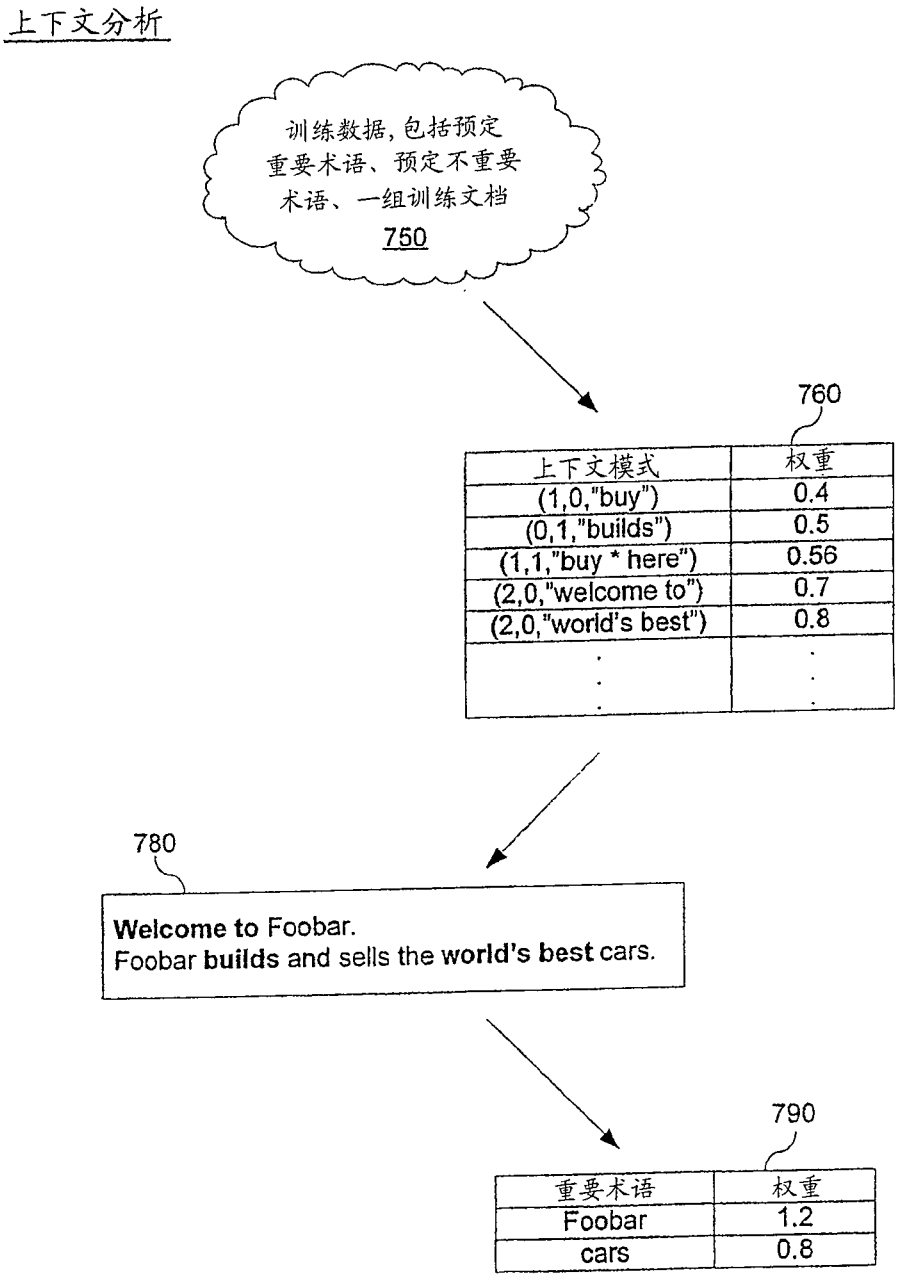

上下文分析
训练数据，包括预定重要术语、预定不重要术语、一组训练文档
750
760
上下文模式
权重
(1,0,"buy")
0.4
(0,1,"builds")
0.5
(1,1,"buy * here")
0.56
(2,0,"welcome to")
0.7
(2,0,"world's best")
0.8
780
Welcome to Foobar.
Foobar builds and sells the world's best cars.
790
重要术语
权重
Foobar
1.2
cars
0.8

图 7B

基于术语文档信息表 810

DOC_ID	(TERM_1, WEIGHT_1)	(TERM_2, WEIGHT_2)	. . .	(TERM_X, WEIGHT_X)	Term-based Ranking Score
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
			. . .		

基于类别文档信息表 830

DOC_ID	(CATEGORY_1, WEIGHT_1)	(CATEGORY_2, WEIGHT_2)	. . .	(CATEGORY_Y, WEIGHT_Y)	Category-based Ranking Score
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
			. . .		

基于链接文档信息表 850

DOC_ID	(LINK_1, WEIGHT_1)	(LINK_2, WEIGHT_2)	. . .	(LINK_Z, WEIGHT_Z)	Link-based Ranking Score
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
			. . .		

图 8

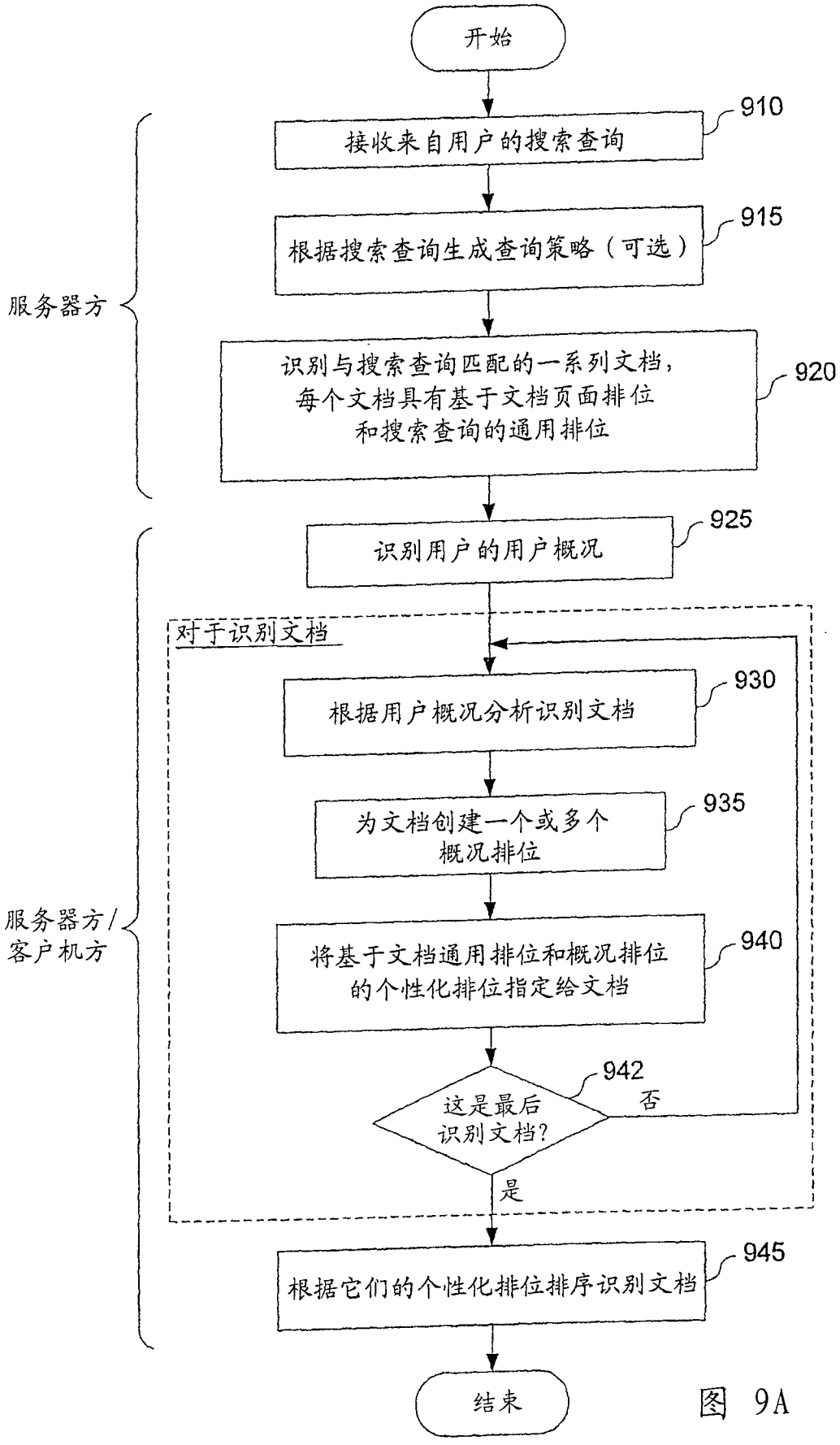

开始
接收来自用户的搜索查询
910
根据搜索查询生成查询策略（可选）
915
识别与搜索查询匹配的一系列文档，每个文档具有基于文档页面排位和搜索查询的通用排位
920
服务器方
识别用户的用户概况
925
对于识别文档
根据用户概况分析识别文档
930
为文档创建一个或多个概况排位
935
将基于文档通用排位和概况排位的个性化排位指定给文档
940
这是最后识别文档?
942
否
是
根据它们的个性化排位排序识别文档
945
服务器方/客户机方
结束

图 9A

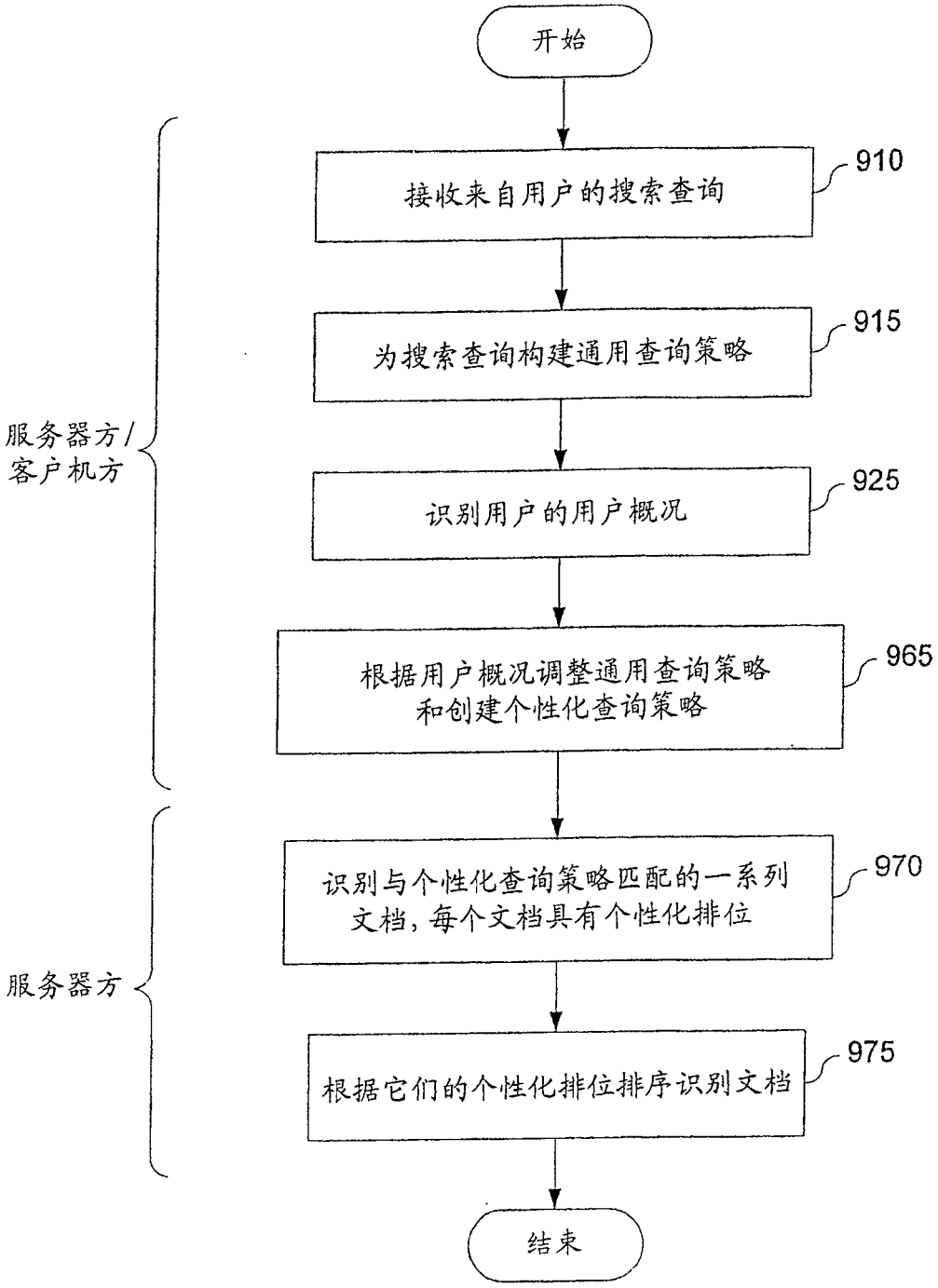

开始
接收来自用户的搜索查询
910
为搜索查询构建通用查询策略
915
识别用户的用户概况
925
根据用户概况调整通用查询策略
和创建个性化查询策略
965
识别与个性化查询策略匹配的一系列
文档，每个文档具有个性化排位
970
根据它们的个性化排位排序识别文档
975
结束
服务器方/
客户机方
服务器方

图 9B

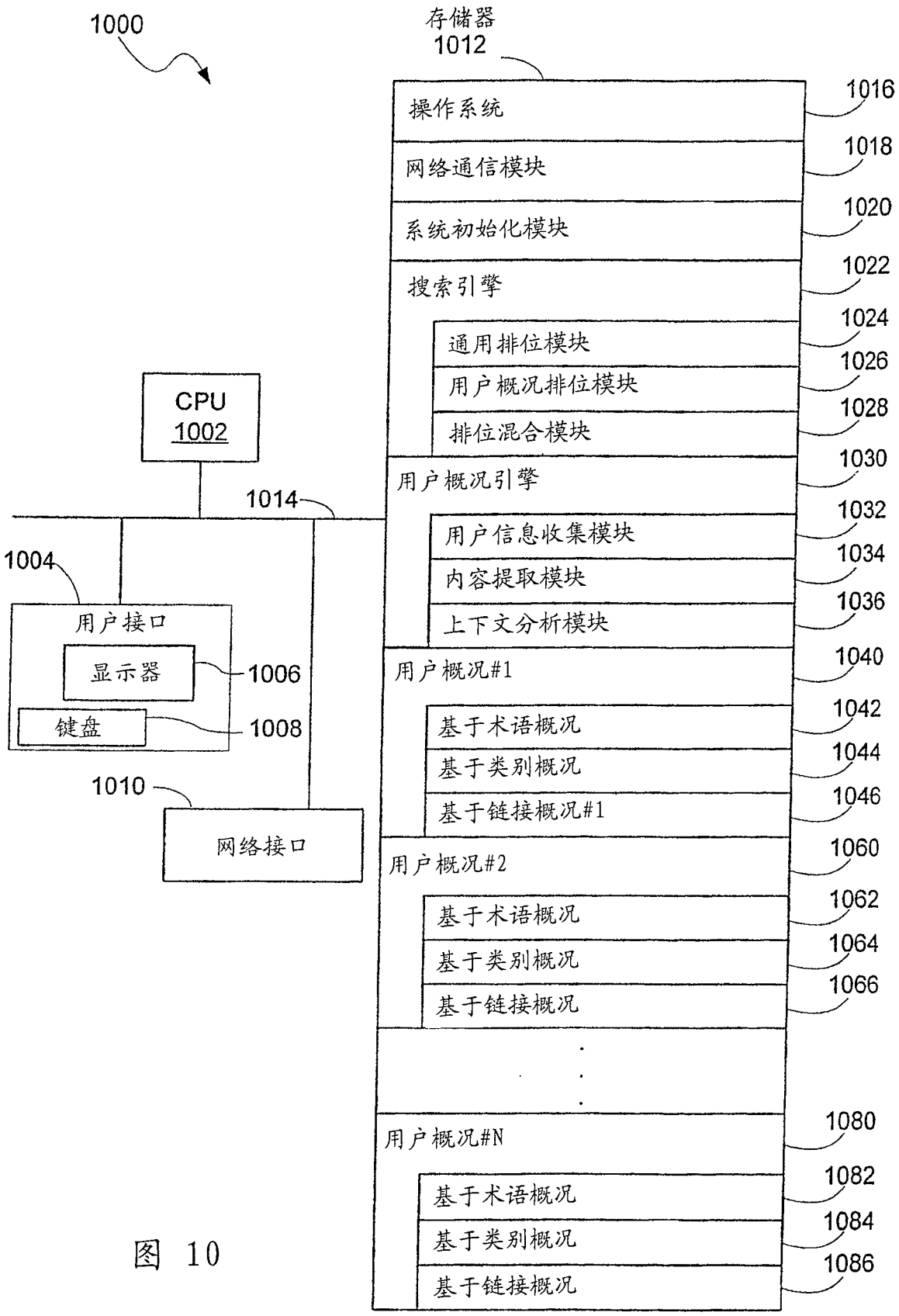

1000
存储器
1012
操作系统
1016
网络通信模块
1018
系统初始化模块
1020
搜索引擎
1022
通用排位模块
1024
用户概况排位模块
1026
排位混合模块
1028
CPU
1002
用户概况引擎
1030
1014
用户信息收集模块
1032
内容提取模块
1034
上下文分析模块
1036
1004
用户接口
显示器
1006
键盘
1008
用户概况#1
1040
基于术语概况
1042
基于类别概况
1044
基于链接概况#1
1046
1010
网络接口
用户概况#2
1060
基于术语概况
1062
基于类别概况
1064
基于链接概况
1066
用户概况#N
1080
基于术语概况
1082
基于类别概况
1084
基于链接概况
1086

图 10

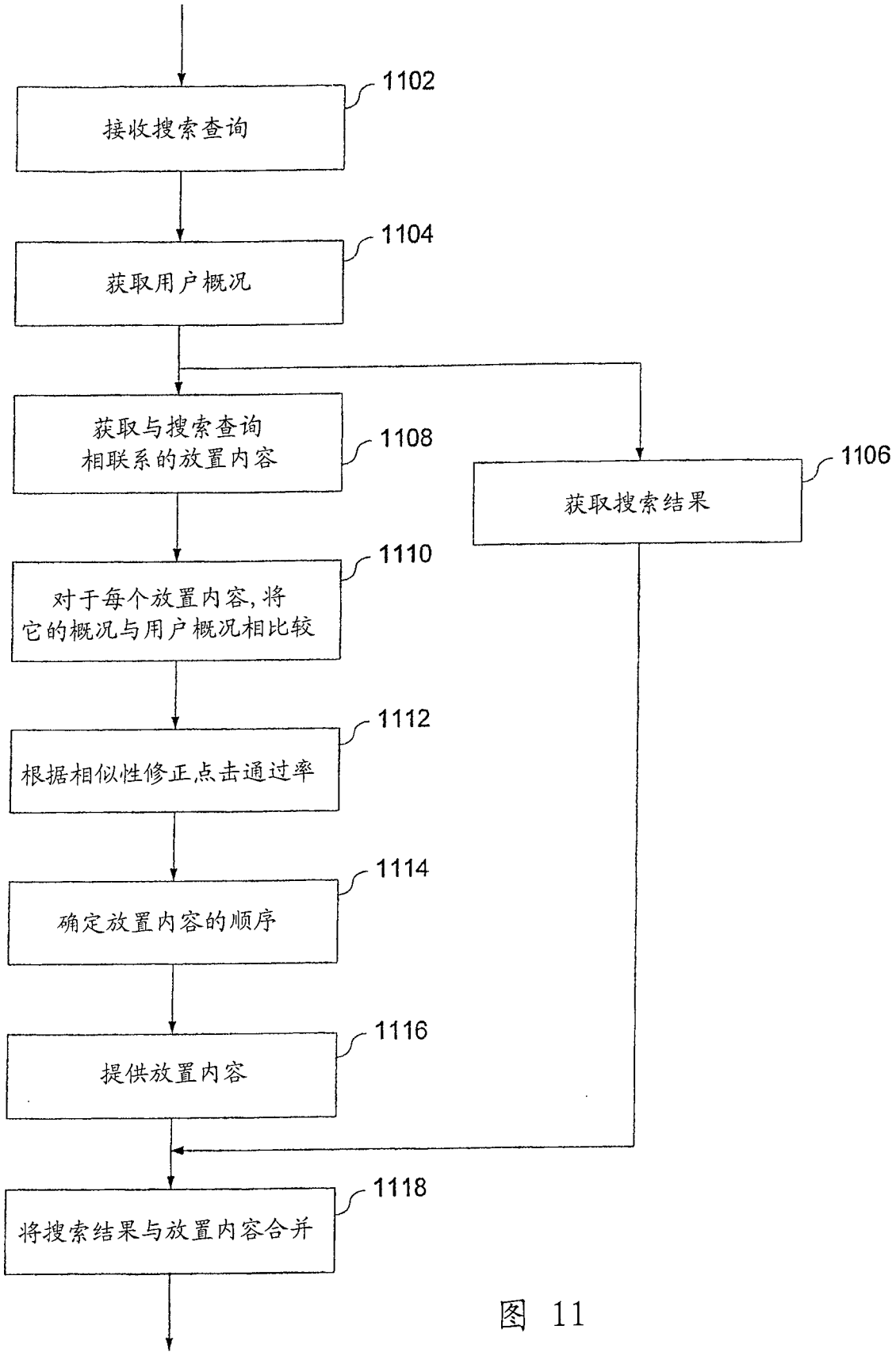

1102
接收搜索查询
1104
获取用户概况
获取与搜索查询
相联系的放置内容
1108
1106
获取搜索结果
1110
对于每个放置内容，将
它的概况与用户概况相比较
1112
根据相似性修正点击通过率
1114
确定放置内容的顺序
1116
提供放置内容
1118
将搜索结果与放置内容合并

图 11