



(12) 发明专利

(10) 授权公告号 CN 101248435 B

(45) 授权公告日 2013. 08. 28

(21) 申请号 200680030148. 3

US 20020152190 A1, 2002. 10. 17, 说明书第

(22) 申请日 2006. 06. 27

2-11 页及附图 1-6.

(30) 优先权数据

US 20050076003 A1, 2005. 04. 07, 全文.

11/169, 285 2005. 06. 29 US

US 20050060311 A1, 2005. 03. 17, 全文.

US 20030220913 A1, 2003. 11. 27, 全文.

(85) PCT申请进入国家阶段日

Mark H. Hansen

2008. 02. 18

Elizabeth Shriver. Using navigation data
to improve IR functions in the context of
Web search. Conference on Information and
Knowledge Management. 2001, 135-142.

(86) PCT申请的申请数据

审查员 李楠

PCT/US2006/025040 2006. 06. 27

(87) PCT申请的公布数据

WO2007/005431 EN 2007. 01. 11

(73) 专利权人 谷歌公司

地址 美国加利福尼亚

(72) 发明人 M·安格罗 D·布拉金斯基

J·金斯伯格 S·童

(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

代理人 李颖

(51) Int. Cl.

G06F 17/30 (2006. 01)

(56) 对比文件

US 6546388 B1, 2003. 04. 08, 全文.

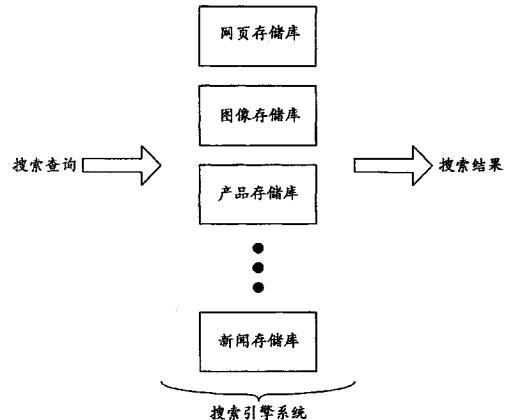
权利要求书5页 说明书9页 附图10页

(54) 发明名称

期望存储库的确定

(57) 摘要

一种系统从用户接收搜索查询，根据该搜索查询搜索一组存储库以为每个存储库识别一个搜索结果集合。该系统还根据用户期望来自所识别存储库的信息的可能性来识别存储库之一，并呈现与所识别存储库关联的搜索结果集合。



1. 一种由一个或多个计算机设备执行的搜索多个存储库的方法,该方法包括:

存储与多个用户的多个在先搜索相关的日志数据,其中使用所述日志数据生成的模型确定与多个存储库中的一个特定存储库包括响应于特定用户提供的特定搜索查询的信息的概率相关联的分数;

从用户接收搜索查询;

根据搜索查询搜索所述多个存储库,以为所述多个存储库中的每个存储库识别一搜索结果集合;

对于所述多个存储库中的每个存储库,通过将与搜索查询相关的信息、与用户相关的信息以及与存储库相关的信息输入模型,确定识别存储库包括响应于搜索查询的信息的概率的分数;

接收所述多个存储库的分别识别对应存储库包括响应于搜索查询的信息的概率的分数作为所述模型的输出;

根据所述多个存储库的所述分数选择所述多个存储库之一;

根据与所选择的所述多个存储库之一关联的搜索结果集合生成搜索结果文档;以及将搜索结果文档提供给与用户相关的客户设备。

2. 权利要求 1 的方法,其中模型是由表示为三位数据 (u, q, r) 的日志数据生成的,其中 u 是指提供搜索查询的用户, q 是指搜索查询,和 r 是指响应于搜索查询从中提供搜索结果的存储库。

3. 权利要求 2 的方法,其中日志数据包括以百万计的三位数据 (u, q, r)。

4. 权利要求 2 的方法,其中模型是通过识别每个三位数据 (u, q, r) 的标签而生成的,其中标签包括与当用户 u 提供搜索查询 q 时用户是否期望来自存储库 r 的信息相关的信息。

5. 权利要求 4 的方法,其中模型是通过根据三位数据 (u, q, r) 和相关标签的训练过程生成的。

6. 权利要求 1 的方法,其中选择所述多个存储库之一包括:

在所述多个存储库的分别识别对应存储库包括响应于搜索查询的信息的概率的分数中选择与最高分数相关联的所述多个存储库之一。

7. 权利要求 1 的方法,其中生成搜索结果文档包括:

根据所述多个存储库中的两个存储库的分别识别对应存储库包括响应于搜索查询的信息的概率的分数选择将要包含在所述搜索结果文档中的与所述多个存储库中的所述两个存储库关联的两个搜索结果集合。

8. 权利要求 7 的方法,其中根据所述多个存储库中的两个存储库的分别识别对应存储库包括响应于搜索查询的信息的概率的分数选择将要包含在所述搜索结果文档中的与所述多个存储库中的两个存储库关联的两个搜索结果集合包括:

根据与所述多个存储库中的所述两个存储库关联的所述分数,将所述两个搜索结果集合设置在搜索结果文档中。

9. 权利要求 8 的方法,其中根据与所述多个存储库中的所述两个存储库关联的分别识别对应存储库包括响应于搜索查询的信息的概率的分数将所述两个搜索结果集合设置在搜索结果文档中包括:

当与所述多个存储库中的所述两个存储库中的第一存储库关联的识别第一存储库包

括响应于搜索查询的信息的概率的第一分数高于与所述多个存储库中的所述两个存储库中的第二存储库关联的识别第二存储库包括响应于搜索查询的信息的概率的第二分数时，在搜索结果文档中将与第一存储库关联的第一搜索结果集合放置在与第二存储库关联的第二搜索结果集合相比更突出的位置上。

10. 权利要求 8 的方法，其中根据与所述多个存储库中的所述两个存储库关联的分别识别对应存储库包括响应于搜索查询的信息的概率的分数在搜索结果文档中放置所述两个搜索结果集合包括：

在搜索结果文档中提供到与所述多个存储库中的所述两个存储库关联的两个搜索结果集合中的至少之一的链接。

11. 权利要求 1 的方法，还包括：

根据所述多个存储库的分别识别对应存储库包括响应于搜索查询的信息的概率的分数选择所述多个存储库中的将要搜索的一组存储库；和

其中搜索多个存储库包括：

对该组存储库执行搜索。

12. 一种在一个或多个计算机设备中实施的搜索多个存储库的系统，所述系统包括：

用于存储与多个用户的多个在先搜索相关的日志数据的装置，其中使用所述日志数据生成的模型确定与多个存储库中的一个特定存储库包括响应于特定用户提供的特定搜索查询的信息的概率相关联的分数；

用于从用户接收搜索查询的装置；

用于根据所述搜索查询对所述多个存储库搜索、以为所述多个存储库中的每个存储库识别一搜索结果集合的装置；

用于对于所述多个存储库中的每个存储库，将与搜索查询相关的信息、与用户相关的信息以及与存储库相关的信息输入模型，确定识别对应存储库包括响应于搜索查询的信息的概率的分数的装置；

用于接收所述多个存储库的分别识别对应存储库包括响应于搜索查询的信息的概率的分数作为所述模型的输出的装置；

用于基于所述多个存储库的所述分数选择所述多个存储库之一的装置；

用于根据与所选择的所述多个存储库之一关联的搜索结果集合生成搜索结果文档的装置；和

用于将搜索结果文档提供给与用户相关的客户设备的装置。

13. 权利要求 12 的系统，还包括：

用于根据所述多个存储库的分别识别对应存储库包括响应于搜索查询的信息的概率的分数选择将要搜索的一组多个存储库的装置。

14. 一种在一个或多个计算机设备中实施的搜索多个存储库的系统，该系统包括：

一个或多个存储装置，被包含在模型生成系统中，用于根据多个用户提供的搜索查询存储与所述多个存储库搜索关联的日志数据，其中由所述模型生成系统使用所述日志数据生成的模型确定与多个存储库中的一个特定存储库包括响应于特定用户提供的特定搜索查询的信息的概率相关联的分数；以及

搜索引擎系统，被配置用于：

从用户接收搜索查询；

对于所述多个存储库中的每个存储库，将与搜索查询相关的信息、与用户相关的信息以及与存储库相关的信息输入模型，确定识别存储库包括响应于搜索查询的信息的概率的分数；

接收所述多个存储库的分别识别对应存储库包括响应于搜索查询的信息的概率的分数作为所述模型的输出；

根据搜索查询对所述多个存储库中的一个或多个存储库执行搜索，以为所述多个存储库中的一个或多个存储库中的每个存储库识别一搜索结果集合；和

根据所述多个存储库的所述分数向用户提供搜索结果集合中的至少一个。

15. 权利要求 14 的系统，其中当对一个或多个存储库执行搜索时，搜索引擎系统被配置为：

根据所述多个存储库的分别识别对应存储库包括响应于搜索查询的信息的概率的分数识别所述多个存储库中的将要搜索的一组存储库；和

搜索所述多个存储库中的该组存储库以为所述多个存储库中的该组存储库中的每个存储库识别一搜索结果集合。

16. 权利要求 14 的系统，其中当对所述多个存储库中的一个或多个存储库执行搜索时，搜索引擎系统被配置为：

根据搜索查询搜索所述多个存储库中的每个存储库。

17. 权利要求 14 的系统，其中该模型是查找表，所述多个存储库的分数中的识别对应存储库包括响应于搜索查询的信息的概率的每一个对应于与对应存储库关联的点击率。

18. 权利要求 14 的系统，其中所述模型还被配置为：

将日志数据表示为三位数据 (u, q, r) ，其中 u 是指提供搜索查询的用户， q 是指搜索查询，和 r 是指响应于搜索查询从中提供搜索结果的存储库。

19. 权利要求 18 的系统，其中日志数据包括以百万计的三位数据 (u, q, r) 。

20. 权利要求 18 的系统，其中所述模型被配置为：

识别每个三位数据 (u, q, r) 的标签，其中标签包括与当用户 u 提供搜索查询 q 时用户是否期望来自存储库 r 的信息相关的信息。

21. 权利要求 20 的系统，其中所述模型被配置为：

根据三位数据 (u, q, r) 和相关标签被训练。

22. 权利要求 14 的系统，其中当提供搜索结果集合中的一个或多个时，搜索引擎系统被配置为：

在所述多个存储库的分别识别对应存储库包括响应于搜索查询的信息的概率的分数中选择具有最高分数的所述多个存储库之一；

呈现与所选择的具有最高分数的所述多个存储库之一关联的搜索结果集合。

23. 权利要求 14 的系统，其中当提供搜索结果集合中的一个或多个时，搜索引擎系统被配置为：

根据与所述多个存储库中的一个或多个存储库关联的分别识别相应的所述一个或多个存储库中的一个存储库包括响应于搜索查询的信息的概率的一个或多个分数，将搜索结果集合中的一个或多个设置在搜索结果文档中；和

向用户提供搜索结果文档。

24. 权利要求 23 的系统,其中当将搜索结果集合中的一个或多个设置在搜索结果文档中时,搜索引擎系统被配置为 :

当与所述多个存储库中的一个或多个存储库中的第一存储库关联的第一分数高于与所述多个存储库中的一个或多个存储库中的第二存储库关联的第二分数时,在搜索结果文档中将与第一存储库关联的第一搜索结果集合放置在与第二存储库关联的第二搜索结果集合相比更突出的位置上,

所述第一分数识别第一存储库包括响应于搜索查询的信息的概率,以及

所述第二分数识别第二存储库包括响应于搜索查询的信息的概率。

25. 权利要求 23 的系统,其中当将搜索结果集合中的一个或多个设置在搜索结果文档中时,搜索引擎系统被配置为 :

在搜索结果文档中提供到与所述多个存储库中的一个或多个存储库关联的一个或多个搜索结果集合中的至少之一的链接。

26. 一种在一个或多个计算机设备中实施的搜索多个存储库的系统,所述系统包括 :

模型生成系统,用于生成模型,所述模型确定反映与所述多个存储库中的一个特定存储库包括响应于特定用户提供的特定搜索查询的信息的概率相关联的分数;和

搜索引擎系统,用于 :

从用户接收搜索查询;

根据该模型确定多个存储库的分别识别对应存储库包括响应于搜索查询的信息的概率的分数;

根据所述多个存储库的所述分数执行所述多个存储库中的一个或多个存储库的搜索;和

根据所述多个存储库的所述分数向与用户相关的客户装置呈现来自一个或多个存储库的搜索结果。

27. 权利要求 26 的系统,其中所述模型是查找表,所述多个存储库的所述分数中的每一分数对应于当用户提供搜索查询时与对应存储库关联的点击率。

28. 一种由一个或多个计算机设备执行的搜索多个存储库的方法,该方法包括 :

从用户接收搜索查询;

根据与在先用户相关的信息、与在先用户提供的在先搜索查询相关的信息以及与从其获得在先搜索查询关联的搜索结果的存储库相关的信息,确定所述多个存储库中每个存储库的识别存储库包括响应于搜索查询的信息的概率的分数;

根据搜索查询和所确定的分数对所述多个存储库中的至少一个存储库执行搜索,以为所述多个存储库中的至少一个存储库中的每个存储库识别一搜索结果集合;和

向与用户相关联的客户装置提供搜索结果集合中的一个或多个。

29. 一种在一个或多个计算机设备中实施的搜索多个存储库的系统,所述系统包括 :

模型生成系统,用于生成第一和第二模型,其中用于生成第二模型的至少一个因素与在生成第一模型时是不同的;和

搜索引擎系统,用于 :

从用户接收搜索查询;

根据第一模型确定所述多个存储库中每个存储库的识别存储库包括响应于搜索查询的信息的概率的第一分数；

根据搜索查询和第一分数对多个存储库中的一个或多个执行搜索；

根据第二模型确定所述多个存储库中的一个或多个存储库中每个存储库的识别存储库包括响应于搜索查询的信息的概率的第二分数；和

根据第二分数组呈现来自所述多个存储库中的一个或多个存储库中至少之一的搜索结果。

30. 权利要求 29 的系统，其中使用第一模型的输出作为第二模型的输入。

期望存储库的确定

技术领域

[0001] 在此描述的实施方式一般地涉及信息检索,更具体地,涉及确定用于进行搜索的期望存储库。

背景技术

[0002] 万维网(“web”)包含大量信息。然而,定位期望的信息部分可能是富有挑战性的。因为万维网上的信息量和缺乏万维网搜索经验的新用户数量在快速地增加,导致这个问题更复杂。

[0003] 搜索引擎系统试图返回到用户感兴趣网页的超链接。通常,搜索引擎系统将它们的用户兴趣确定基于用户输入的搜索项目(称作搜索查询)。搜索引擎系统的目的是根据搜索查询向用户提供到高质量的相关搜索结果(例如网页)的链接。典型地,搜索引擎系统通过匹配搜索查询内的项目与预存储网页的语料库来实现这一目的。包含用户搜索项目的网页为“命中”,其作为链接返回给用户。

[0004] 一些搜索引擎系统可以提供各种信息作为搜索结果。例如,搜索引擎系统可能能够提供与网页、新闻文章、图像、商品、usenet 页面、黄页条目、扫描书籍和 / 或其它类型信息相关的搜索结果。典型地,搜索引擎系统提供到这些不同类型信息的分离界面。

[0005] 当用户将搜索查询提供给标准搜索引擎系统时,通常向用户提供到网页的链接。如果用户期望另一类型的信息(例如图像或新闻文章),用户通常需要访问由搜索引擎系统提供的分离界面。

发明内容

[0006] 根据一个方面,一种方法可以包括:从用户接收搜索查询;根据搜索查询搜索多个存储库,以为每个存储库识别一搜索结果集合;根据用户期望来自所识别存储库的信息的可能性来识别存储库之一;和呈现与所识别存储库关联的搜索结果集合从用户接收搜索查询。

[0007] 根据另一个方面,一种系统可以包括:搜索引擎系统,用于:从用户接收搜索查询;为多个存储库中的每个存储库确定分数,所述存储库之一的分数基于用户期望来自所述一个存储库的信息的可能性。所述搜索引擎系统还根据搜索查询对一个或多个存储库执行搜索,以为一个或多个存储库中的每个存储库识别一搜索结果集合;和根据分数据提供搜索结果集合中的一个或多个。

[0008] 根据又一个方面,提供一种存储数据和计算机可执行指令的计算机可读介质,包括:基于用户提供的搜索查询的与多个存储库搜索关联的日志数据;用于将日志数据表示为三位数据(u,q,r)的指令,其中u是指与提供搜索查询的用户相关的信息,q是指与搜索查询相关的信息,和r是指与响应于搜索查询从中提供搜索结果的存储库相关的信息;用于为每个三位数据(u,q,r)确定标签的指令,其中标签包括与当用户u提供搜索查询q时用户是否期望来自存储库r的信息相关的信息;和用于根据三位数据(u,q,r)和相关标签

训练模型的指令，其中所述模型预测当特定用户提供特定搜索查询时该用户是否期望来自存储库的信息。

[0009] 根据又一个方面，一种系统可以包括：存储第一类型数据的第一存储库；存储第二类型数据的第二存储库；和搜索引擎系统。所述搜索引擎系统从用户接收搜索查询；和根据关于用户、搜索查询和第一或第二存储库的信息，确定用户期望来自第一或第二存储库的信息的可能性。

[0010] 根据另一个方面，一种系统可以包括：模型生成系统和搜索引擎系统。所述模型生成系统用于生成模型，所述模型确定与当特定用户提供特定搜索查询时，该用户期望来自存储库的信息的可能性关联的分数。所述搜索引擎系统从用户接收搜索查询；根据该模型确定多个存储库中每个存储库的分数；和根据分数呈现来自一个或多个存储库的搜索结果。

[0011] 根据又一个方面，一种方法可以包括：从用户接收搜索查询；确定多个存储库中每个存储库的分数，所述存储库之一的分数基于用户期望来自所述一个存储库的信息的可能性；根据搜索查询和所确定的分数对至少一个存储库上执行搜索，以为至少一个存储库中的每个存储库识别一搜索结果集合；和提供搜索结果集合中的一个或多个。

[0012] 根据另一个方面，一种系统可以包括：模型生成系统，用于生成第一和第二模型，其中用于生成第二模型的至少一个因素与在生成第一模型时是不同的或者不存在。该系统还包括搜索引擎系统，用于：从用户接收搜索查询；根据第一模型确定多个存储库中每个存储库的第一分数；根据搜索查询和第一分数对一个或多个存储库执行搜索；根据第二模型确定一个或多个存储库中每个存储库的第二分数；和根据第二分数呈现来自一个或多个存储库中至少之一的搜索结果。

附图说明

[0013] 包含并构成此说明书一部分的附图图示本发明的实施例，并与说明书一起解释本发明。在附图中：

- [0014] 图 1 图示符合本发明原理的概念；
- [0015] 图 2 图示根据符合本发明原理的实施方式的示例模型生成系统；
- [0016] 图 3 是根据符合本发明原理的实施方式的图 2 设备的示例图；
- [0017] 图 4 是根据符合本发明原理的实施方式的用于生成模型的示例处理流程图；
- [0018] 图 5 图示其中可以实施符合本发明原理的系统和方法的示例信息搜索网络；
- [0019] 图 6 是根据符合本发明原理的实施方式的用于提供搜索结果的示例处理流程图；和
- [0020] 图 7-10 图示符合本发明原理的示例实施方式。

具体实施方式

[0021] 下面对本发明的详细描述参考附图。在不同附图中的相同参考标记可以标识相同或类似的单元。而且，下文的详细描述并不限制本发明。

[0022] 概述

[0023] 图 1 图示符合本发明原理的概念。搜索引擎系统可以维护用户可能期望的不同类

型的信息。搜索引擎系统可以维护与不同类型信息相关的一组存储库 (repository)。如图 1 所示,搜索引擎系统可以和与诸如网页、图像、产品和新闻相关的存储库关联。网页存储库可以包括网页相关信息。图像存储库可以包括图像相关信息。产品存储库可以包括商品相关信息。新闻存储库可以包括新闻文档相关信息。搜索引擎系统可以对涉及特定存储库的搜索提供分离界面。

[0024] 在下文的描述中,将术语“文档”广义解释为包括任何机器可读和机器可存储工程产品。文档可以包括例如网页、新闻事件相关信息、图像文件、商品相关信息、usenet 页面相关信息、黄页条目、扫描书籍、文件、文件组合、内嵌有到其它文件的链接的一个或多个文件、博客、网页广告、电子邮件等。文档通常包括文本信息,和可以包括内嵌信息(例如元信息、超链接等)和 / 或内嵌指令(例如 Javascript 等)。如在此使用的术语,将“链接”广义地解释为包括从 / 到一个文档到 / 从另一文档或同一文档的另一部分的任意引用。

[0025] 如图 1 所示,用户可以将搜索查询提供给搜索引擎系统。搜索引擎系统可以确定用户可能期望哪个或哪些存储库。搜索引擎可以执行搜索,和根据用户可能期望哪个或哪些存储库的确定结果呈现包括来自一个或多个存储库的信息的搜索结果。

[0026] 例如,如果用户将项目“日落 (sunset)”作为搜索查询提供给搜索引擎系统,则搜索引擎系统可以确定用户更关心日落图片而不是日落相关的网页。因此,搜索引擎系统可以向用户提供来自图像存储库的搜索结果而不是来自其它存储库的搜索结果,或者作为其补充。

[0027] 类似地,如果用户将短语“伊拉克战争”作为搜索查询提供给搜索引擎系统,则搜索引擎可以确定用户更关心涉及伊拉克战争相关的新闻文档而不是伊拉克战争相关的网页。因此,搜索引擎系统可以向用户提供来自新闻存储库的搜索结果而不是来自其它存储库的搜索结果,或者作为其补充。

[0028] 符合本发明原理的实施方式可以在用户提供搜索查询时生成预测用户关注哪个或哪些存储库的模型,并使用此模型将相关搜索结果提供给用户。

[0029] 示例的模型生成系统

[0030] 图 2 是符合本发明原理的模型生成系统 200 的示例图。系统 200 可以包括一个或多个设备 210 和日志数据存储器 220。存储器 220 可以包括一个或多个逻辑或物理存储设备,其可以存储如下文更详细描述的可能使用的大型数据集合(例如成百万的实例和数以万计的特征)以建立和训练模型。该数据可以包括涉及在先搜索的日志数据,例如用户信息、查询信息和存储库信息,其可以用于建立可用于识别用户可能期望的一个或多个存储库的模型。在一种实施方式中,该模型可以当用户提供特定查询时预测用户是否期望来自特定存储库的信息。

[0031] 用户信息可以包括与用户相关的因特网协议 (IP) 地址、cookie 信息、语言和 / 或地理信息、用户提供的在前查询和 / 或用户提供当前或在前查询的当天时间和 / 或日期。查询信息可以包括与提供的查询项目相关的信息。存储库信息可以包括与用于搜索的存储库界面、显示的文档和从中获取它们的存储库和 / 或选择的文档(例如点击)相关的信息。在其它的示例实施方式中,可以替代的或者附加地由存储器 320 保存其它类型的数据。

[0032] 一个或多个设备 210 可以包括能够通过任意类型的连接机制访问存储器 220 的任意类型的计算设备。根据符合本发明原理的一种实施方式,系统 200 可以包括多个设备

210。根据另一种设施方式,系统 200 可以包括单个设备 210。

[0033] 图 3 是根据符合本发明原理的实施方式的设备 210 的示例图。设备 210 可以包括总线 310、处理器 320、主存储器 330、只读存储器 (ROM) 340、存储设备 350、输入设备 360、输出设备 370 和通信接口 380。总线 310 可以包括允许在设备 210 的单元之间通信的路径。

[0034] 处理器 320 可以包括可以解释和执行指令的处理器、微处理器或者处理逻辑。主处理器 330 可以包括可存储信息和用于由处理器 320 执行的指令的随机访问存储器 (RAM) 或另一类型的动态存储设备。ROM 340 可以包括可存储静态信息和由处理器 320 使用的指令的 ROM 设备或另一类型的静态存储设备。存储设备 350 可以包括磁和 / 或光记录介质及其相应驱动器。

[0035] 输入设备 360 可以包括允许操作者将信息输入给设备 210 的机械装置,例如键盘、鼠标、笔、语音识别和 / 或生物测定机械装置等。输出设备 370 可以包括将信息输出给操作者的机械装置,包括显示器、打印机、扬声器等。通信接口 380 可以包括支持设备 210 与其它设备和 / 或系统通信的任意收发信机类似的机械装置。例如,通信接口 380 可以包括用于与另一设备 210 或存储器 220 通信的机械装置。

[0036] 如将在下文中详细描述的,符合本发明原理的设备 210 可以执行某些模型生成相关操作。响应于处理器 320 执行在诸如存储器 330 等计算机可读介质内包含的软件指令,设备 210 可以执行这些操作。可以将计算机可读介质定义为物理或逻辑存储设备和 / 或载波。

[0037] 可以从诸如数据存储设备 350 等另一个计算机可读介质或者通过通信接口 380 从另一个设备将软件指令读入存储器 330。在存储器 330 内包含的软件指令可以致使处理器 320 执行随后将要描述的处理。可替代地,可以使用硬连线电路替代软件指令或者与其组合以实现符合本发明原理的处理。因而,符合本发明原理的实施方式并不限制于硬件电路和软件的任意特定组合。

[0038] 示例的模型生成处理

[0039] 为了下文讨论的目的,在存储器 220 内的数据组 (图 2) 可以包括多个单元,称作实例。存储器 220 可以包括以百万计的实例。每个实例可以包括三位数据 (triple of data) : (u, q, r) , 其中“u”是指用户信息,“q”是指用户 u 提供的查询,和“r”是指响应于查询 q 从中提供搜索结果的存储库。存储器 220 还可以存储与当用户 u 提供查询 q 时用户 u 是否期望来自存储库 r 的信息相关的信息,其中例如可以通过确定用户是否从存储库选择文档来测量用户的期望。此信息将称作该实例的“标签”。

[0040] 可以从任意给定的 (u, q, r) 提取若干特征。存储器 220 可以包括数以万计的不同特征。在一种实施方式中,这些特征中的一些特征可以包括一个或多个下述内容:用户 u 位于的国家、用户 u 位于的国家的语言、与用户 u 相关的 cookie 标识符、查询 q 的语言、查询 q 中的每个项目、用户 u 提供查询 q 的当天时间、提供给用户 u 的存储库 r 的文档、提供给用户 u 的存储库 r 中文档内的每个项目和 / 或提供给用户 u 的存储库 r 中文档标题中的每个项目。也可以替代地或者附加地使用其它特征。

[0041] 在另一种实施方式中,附加地或者替代上面识别的一些特征,一些特征可以包括一个或多个下述内容:提供给存储库 r 的界面的查询片断 (fraction)、提供给存储库 r 的界面对其它存储库的界面的查询片断、包含提供给存储库 r 的界面对其它存储库的界面的

查询 q 内项目的查询片断、提供给存储库 r 的界面的查询的整体点击率、为用户 u 提供给存储库 r 的界面的查询点击率、为与用户 u 同一国家内的用户提供给存储库 r 界面的查询点击率和 / 或提供给存储库 r 界面的查询 q 的点击率。

[0042] 在又一种实施方式中,还可以包括下述两个特征:为用户 u 提供给存储库 r 的界面的查询 q 的点击率和为用户 u 提供给存储库 r 界面的查询 q 的片断。不是直接确定这些特征,而是可以生成模型以使用常规技术预测这些特征并可以将模型输出用作特征。

[0043] 可以根据此数据建立模型。在一种实施方式中,给定新的 (u, q, r),可以使用模型预测如果用户 u 提供了查询 q,用户 u 是否期望来自存储库 r 的信息。如在下文中将更详细描述的,可以使用模型输出确定是否搜索存储库,是否在搜索结果文档中包含来自存储库的搜索结果和 / 或在搜索结果文档中呈现搜索结果的方式。

[0044] 图 4 是根据符合本发明原理的实施方式的用于生成模型的示例处理流程图。该处理可以由单个设备 210 或多个设备 210 的组合执行。

[0045] 为了便于生成模型,可以将存储器 220 内的日志数据表示为实例集合(方框 410)。例如,可以与用户的先前搜索相关地识别信息,例如关于用户、用户提供的查询和从中获取和 / 或选择搜索结果的存储库的信息。如上文所述,可以将此信息形成为三位数据 (u, q, r)。

[0046] 随后,可以确定每个实例的标签(方框 420)。例如,可以为每个三位数据 (u, q, r) 确定当用户 u 提供了查询 q 时用户 u 是否期望存储库 r 内的信息(例如,选择文档)。标签可以与存储器 220 内的它们的相应实例关联。还可以确定与每个实例相关的特征(方框 430)。

[0047] 随后,可以根据实例、标签和特征生成模型(方框 440)。例如,可以使用标准机器学习或统计技术确定当用户 u 提供查询 q 时用户 u 期望来自存储库 r 的信息的概率:

[0048] $P(\text{desire} | u, q, \text{show_r})$,

[0049] 其中“show_r”表示提供来自存储库 r 的文档。可以使用若干公知技术中的任一种技术生成模型,例如逻辑回归、增强判决树、随机森林、支持向量机器、感知器和辨别学习器。该模型可以输出反映当用户 u 提供查询 q 时用户 u 期望来自存储库 r 的信息的信任的值,而不是生成概率。在下文中通常将模型输出称作“分数”(score),其可以包括概率输出和 / 或输出值。

[0050] 如下文解释的,可以使用模型输出确定是否搜索存储库,是否将来自存储库的搜索结果包括在搜索结果文档中和 / 或用于在搜索结果文档中呈现搜索结果的方式。

[0051] 示例的信息提取网络

[0052] 图 5 是其中可以实施符合本发明原理的系统和方法的网络示例图。网络 500 可以包括经网络 550 连接到多个服务器 520-540 的多个客户机 510。为了简化,已经图示了两个客户机 510 和三个服务器 520-540 连接到网络 550。实际上,可能存在更多或更少的客户机和服务器。而且,在一些实例中,客户机可以执行服务器功能,服务器可以执行客户机功能。

[0053] 客户机 510 可以包括客户机实体。可以将实体定义为设备,例如个人计算机、无线电话机、个人数字助理 (PDA)、便携式或另一类型的计算或通信设备、在这些设备之一上运行的线程或过程和 / 或由这些设备之一可执行的对象。服务器 520-540 可以包括以符合本发明原理的方式收集、处理、搜索和 / 或保存文档的服务器实体。

[0054] 在符合本发明原理的实施方式，服务器 520 可以包括可由客户机 510 使用的搜索引擎系统 525。搜索引擎系统 525 可以与多个文档存储库（未图示）关联，例如网页存储库、新闻存储库、图像存储库、产品存储库、usenet 存储库、黄页存储库、扫描书籍存储库和 / 或其它类型的存储库。这些存储库可以物理驻留于服务器 520 内的一个或多个存储设备内或者在服务器 520 外部。服务器 530 和 540 可以存储或保存可与一个或多个存储库关联的文档。

[0055] 虽然将服务器 520-540 图示为分离实体，但是也可以由一个或多个服务器 520-540 执行另外一个或多个服务器 520-540 的一个或多个功能。例如，可以将两个或更多服务器 520-540 实施为单个服务器。也可以将单个服务器 520-540 实施为两个或更多分离（并且可能是分布式）的设备。

[0056] 网络 550 可以包括局域网（LAN）、广域网（WAN）、诸如公用交换电话网（PSTN）的电话网络、内联网、互联网或者网络组合。客户机 510 和服务器 520-540 可以通过有线、无线和 / 或光连接连到网络 550。

[0057] 提供搜索结果的示例过程

[0058] 图 6 是根据符合本发明原理的实施方式的用于提供搜索结果的示例处理流程图。处理可以开始于接收搜索查询（方框 610）。例如，用户可以使用在诸如客户机 510（图 5）等客户机上的 web 浏览器软件访问搜索引擎界面。用户可以将搜索查询提供给搜索引擎界面。

[0059] 可以获取用户相关信息（方框 620）。例如，可以使用诸如与用户相关的 IP 地址、cookie 信息、语言和 / 或地理信息识别用户。可以使用常规技术收集用户信息。

[0060] 在一种实施方式中，可以根据搜索查询对每个存储库执行搜索（方框 430）。可以获取与每个存储库对应的一个搜索结果集合。可以使用任意信息检索技术识别将包括在检索结果集合内的相关文档。

[0061] 随后，可以根据模型确定如何提供搜索结果（方框 640）。例如，可以使用关于用户、用户提供的搜索查询和每个存储库的信息作为模型输入。可以将该模型应用于每个存储库并可以使用模型输出（“分数”）以确定是否提供与该存储库相关的搜索结果。例如，可以确定应当提供来自具有最高相关分数的两个存储库的搜索结果。可替代地，可以确定应当始终提供来自一个特定存储库的搜索结果，并且如果与其他一个或多个存储库相关的分数大于与该特定存储库关联的分数，则还应当提供来自另外一一个或多个存储库的搜索结果。可替代地，可以确定应当提供来自具有高于某个阈值的相关分数的存储库的搜索结果，如果没有分数高于该阈值，则提供来自具有最高相关分数的存储库的搜索结果。可以替代地或附加地使用用于确定是否提供与存储库关联的搜索结果的其它规则。

[0062] 可以替代地或者附加地使用模型输出确定提供来自不同存储库的搜索结果的方式。例如，可以确定如果与存储库关联的分数低于某个阈值，则可以将与存储库相关的搜索结果提供在向用户呈现的搜索结果文档的底部，而不是搜索结果文档的顶部。可替代地或者附加地，可以确定如果与存储库关联的分数低于某个阈值，则呈现到与该存储库相关的搜索结果的链接，而不是搜索结果本身。可以替代地或者附加地使用用于确定提供与存储库相关的搜索结果的方式的其它规则。

[0063] 随后，可以将搜索结果设置在搜索结果文档中并提供给用户。每个搜索结果例如

可以包括到来自对应存储库的文档的链接和可能的对该文档的简要描述或摘录。

[0064] 在另一种实施方式中,可以根据模型识别将要搜索的一个或多个存储库(方框650)。例如,可以使用关于用户、用户提供的搜索查询和每个存储库的信息作为模型的输入。可以将该模型应用于每个存储库,和可以使用模型的输出(“分数”)确定将要搜索哪个存储库。例如,可以确定应当搜索具有最高相关分数的两个存储库。可替代地,可以确定应当始终搜索存储库中的一个特定存储库,并且如果与另外一个或多个存储库相关的分数高于与该特定存储库关联的分数,则还应当搜索另外一个或多个存储库。可替代地,可以确定应当搜索具有高于某个阈值的相关分数的存储库,如果没有分数高于该阈值,则搜索具有最高相关分数的存储库。可以可替代地或者附加地使用用于确定将要搜索哪个存储库的其它规则。

[0065] 可以执行搜索以获得来自每个所识别存储库的搜索结果集合(方框660)。可以使用任意的常规信息检索技术识别相关文档以包括在搜索结果集合内。

[0066] 随后,可以根据模型提供搜索结果(方框670)。例如,可以使用模型输出确定提供来自不同存储库的搜索结果的方式。例如,可以确定如果与存储库关联的分数低于某个阈值,则可以将与该存储库关联的搜索结果呈现在呈现给用户的搜索结果文档的底部而不是在搜索结果文档的顶部。可替代地,或者附加地,可以确定如果与存储库关联的分数低于某个阈值,可以提供到与该存储库关联的搜索结果的链接,而不是搜索结果本身。可以替代地或者附加地使用用于确定提供与存储库关联的搜索结果的方式的其它规则。

[0067] 随后,可以将搜索结果设置在搜索结果文档中和提供给用户。每个搜索结果可以包括例如到来自相应存储库的文档的链接和可能的对该文档的简要描述或摘录。

[0068] 在另一种实施例中,可以使用两个或更多模型。例如,可以使用第一模型确定是否搜索存储库;可以使用第二模型确定是否在搜索结果文档中包括来自搜索存储库之一的搜索结果;以及可以使用第二模型、可能还有第三模型确定用于在搜索结果文档中呈现搜索结果的方式。可以根据彼此不同的一个或多个因素生成第一、第二和/或第三模型。例如,在一种实施方式中,可以使用第一模型的输出作为第二模型的输入和/或可以使用第一和/或第二模型的输出作为第三模型的输入。

[0069] 可以将与此搜索相关的信息作为日志数据提供给存储器220。例如,可以使用此信息作为用于训练或优化该模型的训练数据。

[0070] 例子

[0071] 图7至图10图示符合本发明原理的示例实施方式。如图7所示,假设搜索引擎系统710具有三个相关存储库,包括网页存储库720、图像存储库730和新闻存储库740。网页存储库720可以存储网页相关信息。图像存储库730可以存储图像相关信息。新闻存储库740可以存储新闻文档相关信息。搜索引擎系统710可以接收来自用户的搜索查询,和提供来自一个或多个存储库720-740的相关搜索结果。

[0072] 如图8所示,假设用户访问与搜索引擎系统710关联的界面。该界面可以与存储库之一关联或者不与任何存储库关联。如图8所示,假设用户将搜索查询“sunset”提供给搜索引擎系统710。除了搜索查询之外,搜索引擎系统710可以获取用户相关信息,例如与用户相关的IP地址、cookie信息、语言和/或地理信息。

[0073] 在一种实施方式,如上文所述,搜索引擎系统710可以对每个存储库720-740执行

搜索以获取每个存储库 720–740 的搜索结果集合。假设搜索引擎系统 710 识别出来自网页存储库 720 的 10 个网页结果、来自图像存储库 730 的 10 个图像结果和来自新闻存储库 740 的 10 个新闻文档结果作为用于搜索查询“sunset”的相关搜索结果。

[0074] 搜索引擎系统 710 可以输入与用户、用户提供的搜索查询和每个存储库 720–740 相关的信息作为模型的输入。可以使用该模型确定当用户提供搜索查询“sunset”时用户期望来自每个存储库 720–740 的信息的概率。

[0075] 例如，假设通过模型生成下述输出：

[0076] $P(\text{desire} | u, q, \text{show_web page repository}) = 0.45$

[0077] $P(\text{desire} | u, q, \text{show_image repository}) = 0.91$

[0078] $P(\text{desire} | u, q, \text{show_news repository}) = 0.23$

[0079] 其中“u”是指与提供搜索查询的用户对应的用户信息，“q”是指与用户提供的搜索查询对应的信息（即“sunset”），和“show_xrepository”（其中 x 对应于“web page”、“image”或“news”）是指与所识别存储库对应的信息。在这种情况下，当用户提供搜索查询“sunset”时用户期望来自网页存储库 720 的信息的概率是 45%；当用户提供搜索查询“sunset”时用户期望来自图像存储库 730 的信息的概率是 91%；和当用户提供搜索查询“sunset”时用户期望来自新闻存储库 740 的信息的概率是 23%。

[0080] 随后，搜索引擎系统 710 可以使用与每个存储库 720–740 相关的模型输出确定是否提供与该存储库关联的搜索结果。例如，假设规则指示搜索引擎系统 710 将仅提供来自具有最高分数的存储库的搜索结果。在这种情况下，搜索引擎系统 710 可以根据从图像存储库 730（即具有最高分数 0.91 的存储库）识别出的 10 个图像结果形成搜索结果文档，如图 9 所示。

[0081] 可替代地，假设规则指示搜索引擎系统 710 始终提供来自网页存储库 720 的搜索结果，并且如果另一个存储库具有高于与网页存储库 720 关联的分数的关联分数，则提供来自该存储库（或多个存储库）的搜索结果。在这种情况下，搜索引擎系统 710 可以确定它提供来自网页存储库 720 和图像存储库 730 的搜索结果，因为与图像存储库 730 关联的分数（0.91）大于与网页存储库 720 关联的分数（0.45）。

[0082] 随后，搜索引擎系统 710 可以根据来自网页存储库 720 的 10 个网页结果和来自图像存储库 730 的 10 个图像结果形成搜索结果文档，如图 10 所示。因为与图像存储库 730 关联的分数高于与网页存储库 720 关联的分数（或者一定程度高于或大于阈值），可以在搜索结果文档中在与 10 个网页结果相比更突出的位置上提供与 10 个图像结果相关的信息，同样如图 10 所示。类似于图 9 所示，用户可以选择将与 10 个图像结果关联的链接与图像结果相关的附加信息（例如“SEE10 IMAGE RESULTS FOR SUNSET”）一起呈现。

[0083] 结论

[0084] 符合本发明原理的实施方式可以生成可用于预测当用户提供搜索查询时用户可能对哪个或哪些存储库感兴趣的模型，并使用此模型向用户提供相关搜索结果。

[0085] 本发明优选实施例的上述描述提供说明和描述，但是将不是穷尽的或者不将本发明限制于所公开的具体形式。鉴于上述教导可以进行修改和变化，或者可以通过实施本发明获得。

[0086] 例如，虽然已经参考图 4 和图 6 描述动作序列，但是可以在符合本发明原理的其它

实施方式中修改动作顺序。此外,可以并行执行非从属动作。

[0087] 而且,已经参考图 8-10 描述示例的用户界面。在符合本发明原理的其它实施方式中,用户界面可以包括更多、更少或者不同的信息。

[0088] 前面的描述提到用户。“用户”将是指客户机,例如客户机 510(图 5)或者客户机的操作者。

[0089] 此外,已经描述了可以使用模型输出(“分数”)确定是否搜索存储库,是否在搜索结果文档中包括来自存储库的搜索结果,和 / 或用于在搜索结果文档中呈现搜索结果的方式。在另一种实施方式中,可以使用分数作为对确定是否搜索存储库、是否在搜索结果文档中包括来自存储库的搜索结果、和 / 或用于在搜索结果文档中呈现搜索结果的方式的函数的一个输入或者多个输入。

[0090] 此外,确定上面描述的一些特征比确定其它特征需要更大计算量。例如,在存储库内基于文档的特征可能需要查询这些存储库和提取文档。为了计算效率,可以根据较低计算量(例如更廉价)的特征建立近似主模型,可以使用该近似主模型确定将要搜索哪些存储库。一旦已经提取来自这些存储库的文档,则可以使用完全主模型确定从哪些存储库提供搜索结果。

[0091] 而且,能够根据“探测”(exploration)策略使用该模型以收集关于不同存储库的信息。例如,可能希望提供与次佳存储库相关的搜索结果(例如提供新闻文档而不是图像)。一种探测策略可以指示将来自随机存储库的文档呈现给一小部分用户。另一种探测策略可以指示与分数成比例地呈现来自存储库的文档(例如如果确定图像分数两倍于新闻文章分数,则随后可以以两倍于新闻文章的频度提供图像)。

[0092] 已经描述可以生成模型以根据用户期望来自所识别存储库的信息的可能性来识别存储库(或一组存储库)。在一种实施方式中,可以将该模型构建为查找表,其具有根据诸如与查询相关的一个或多个特征(例如查询项)等一个或多个特征确定的关键字(key)。查找表的输出可以包括用于每个存储库的点击率(或者估计点击率)。在这种情况下,用户期望来自存储库之一的信息的可能性可以是该存储库的点击率的函数。例如,可以根据存储库的点击率确定是否搜索存储库、是否在搜索结果文档中包括来自存储库的搜索结果、和 / 或呈现搜索结果的方式。

[0093] 对于本领域的普通技术人员来说,显然可以将上面描述的本发明的各个方面实施为多种不同形式的如图所示实施方式中的软件、固件和硬件。用于实施符合本发明原理各个方面实际软件编码或专用控制硬件并不限制于本发明。因而,在不参考特定软件代码的情况下描述各个方面的操作和行为 - 将理解本领域的普通技术人员将能够根据在此的描述设计软件和控制硬件以实现各个方面。

[0094] 不应当将本申请中使用的单元、动作或指令解释为本发明必需的,除非明确如此描述。而且,如在此使用的,“一”将包括一个或多个项目。在仅指一个项目时,使用术语“一个”或类似用词。此外,短语“基于”将指“至少部分地基于”,除非明确陈述。

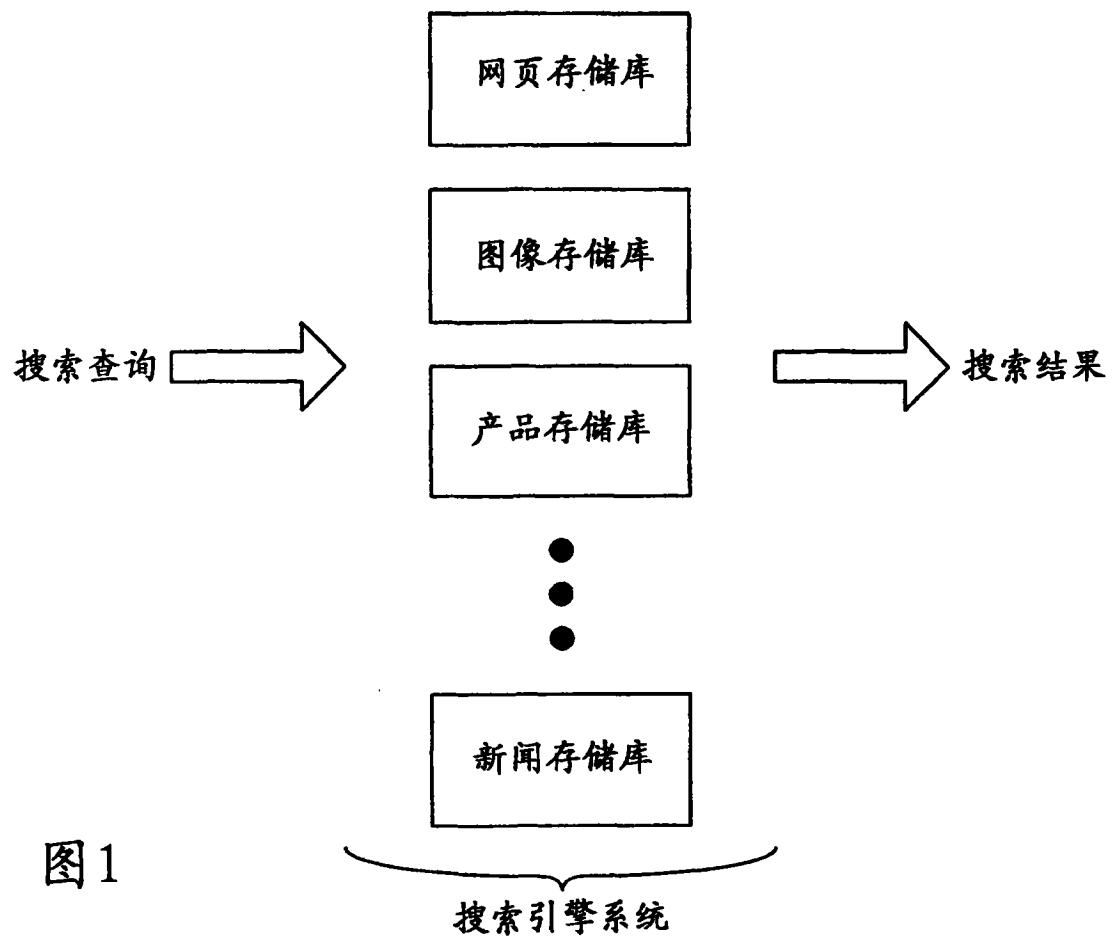


图 1

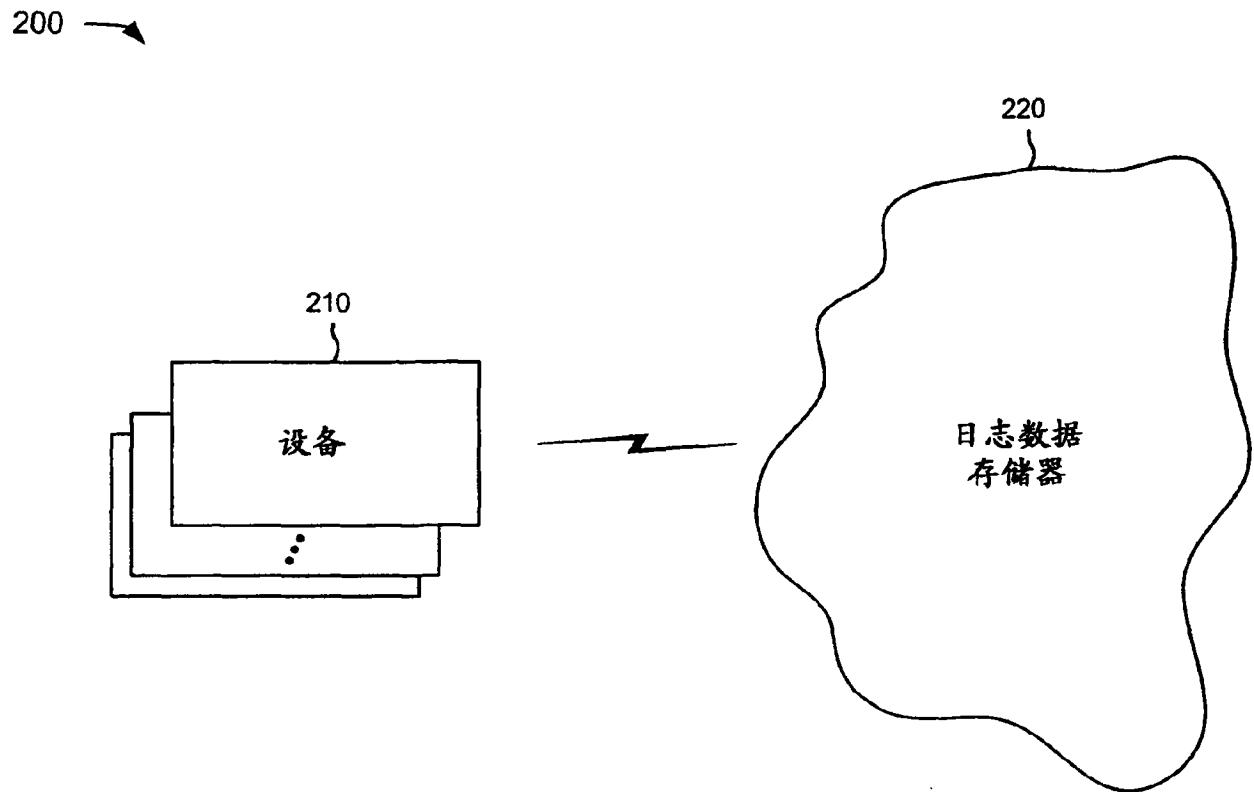


图 2

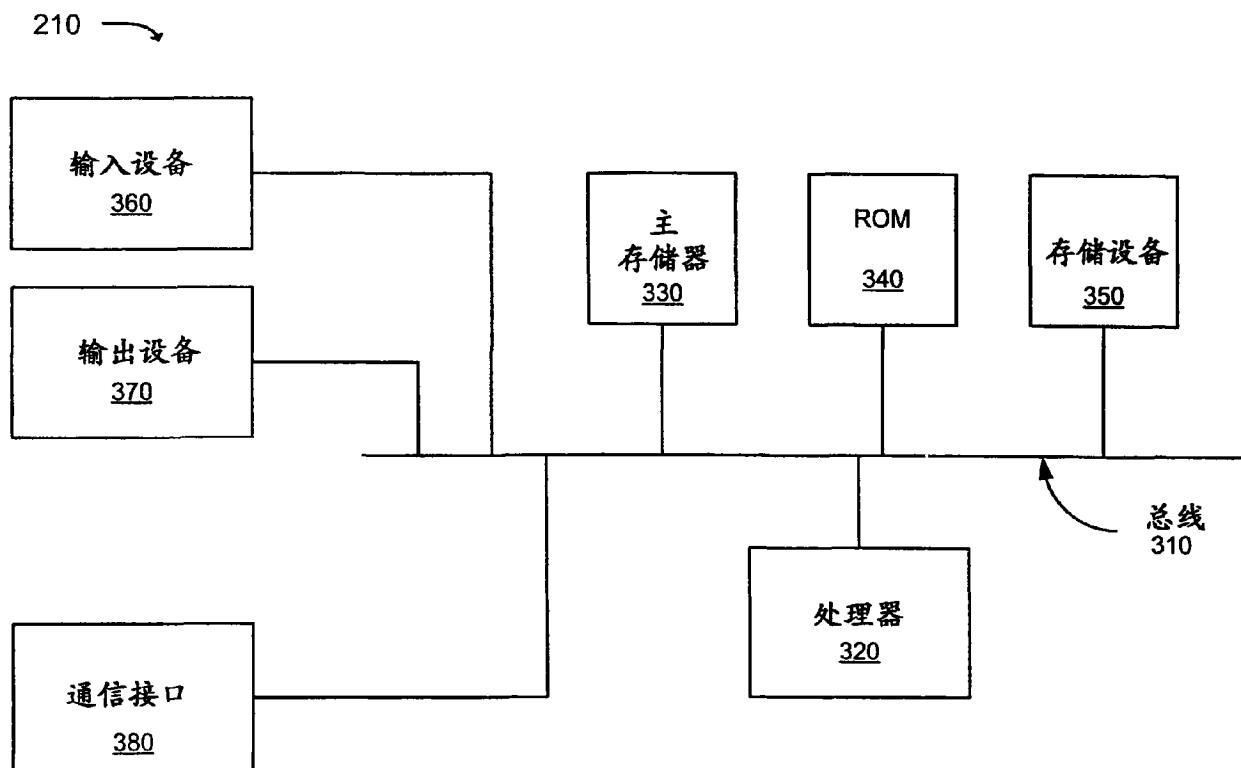


图 3

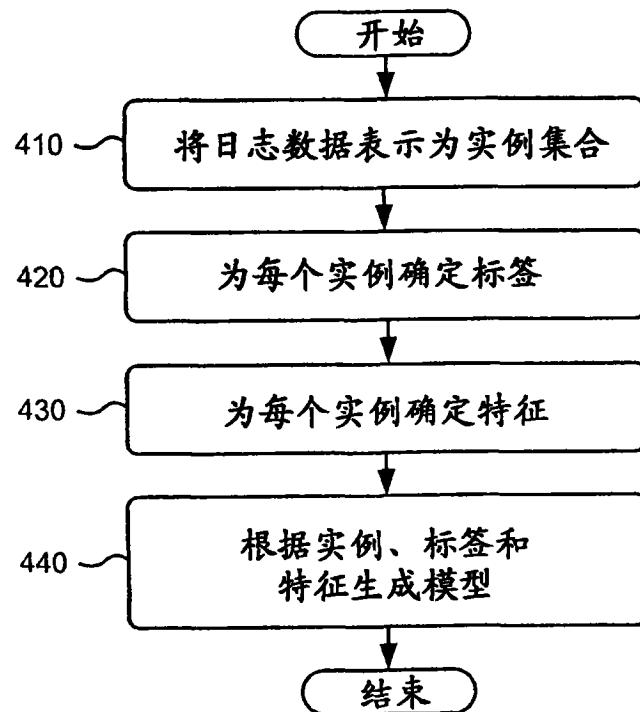


图 4

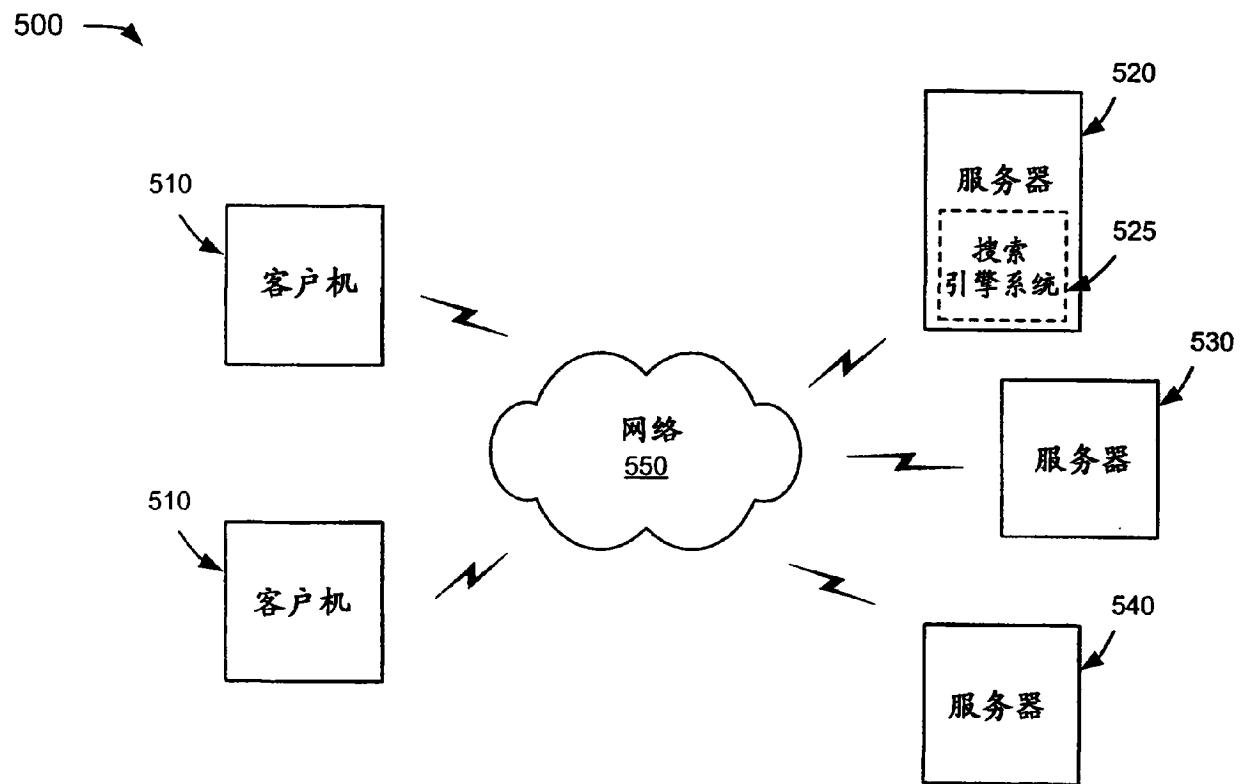


图 5

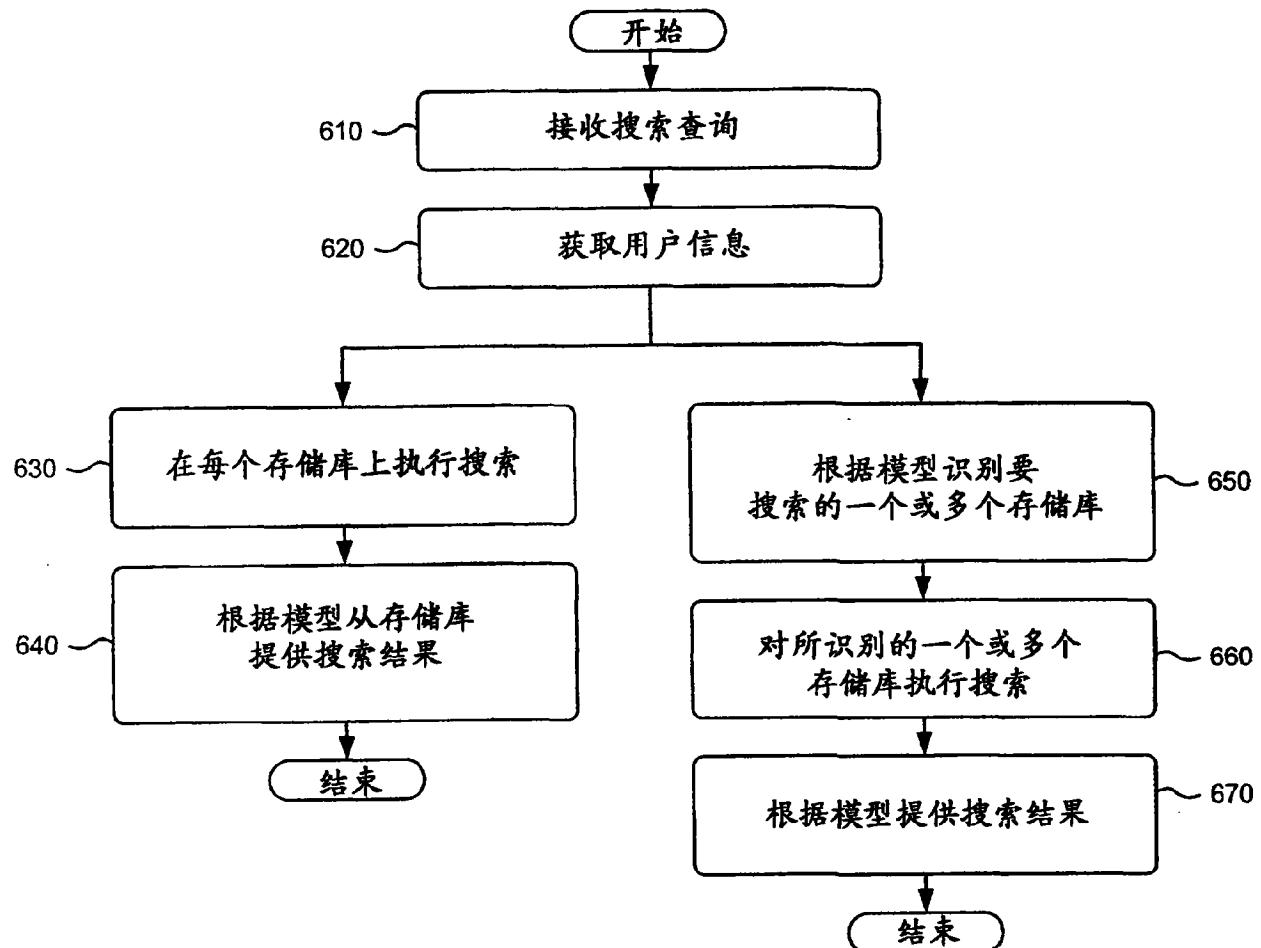


图 6

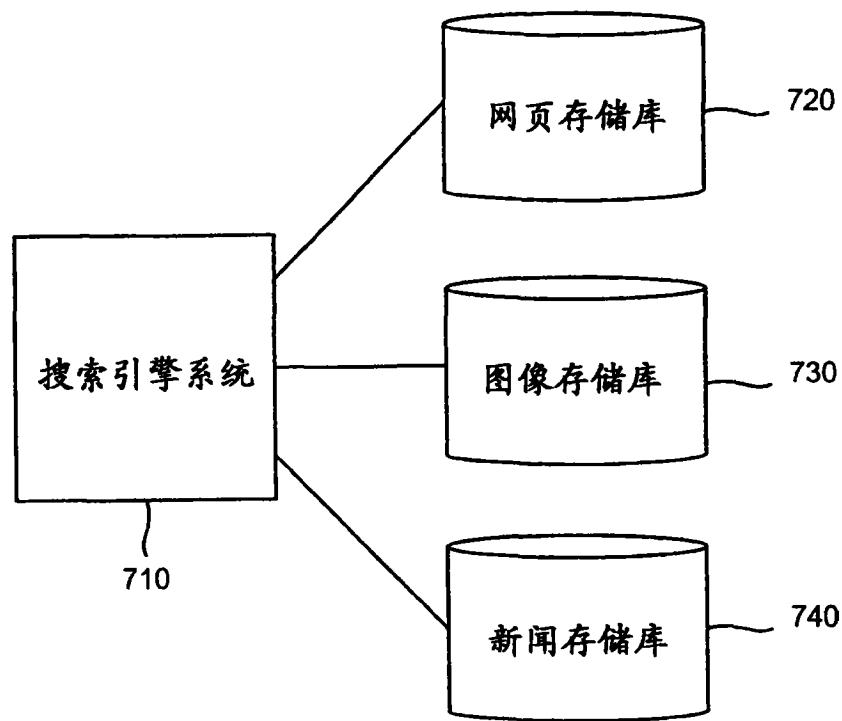


图 7

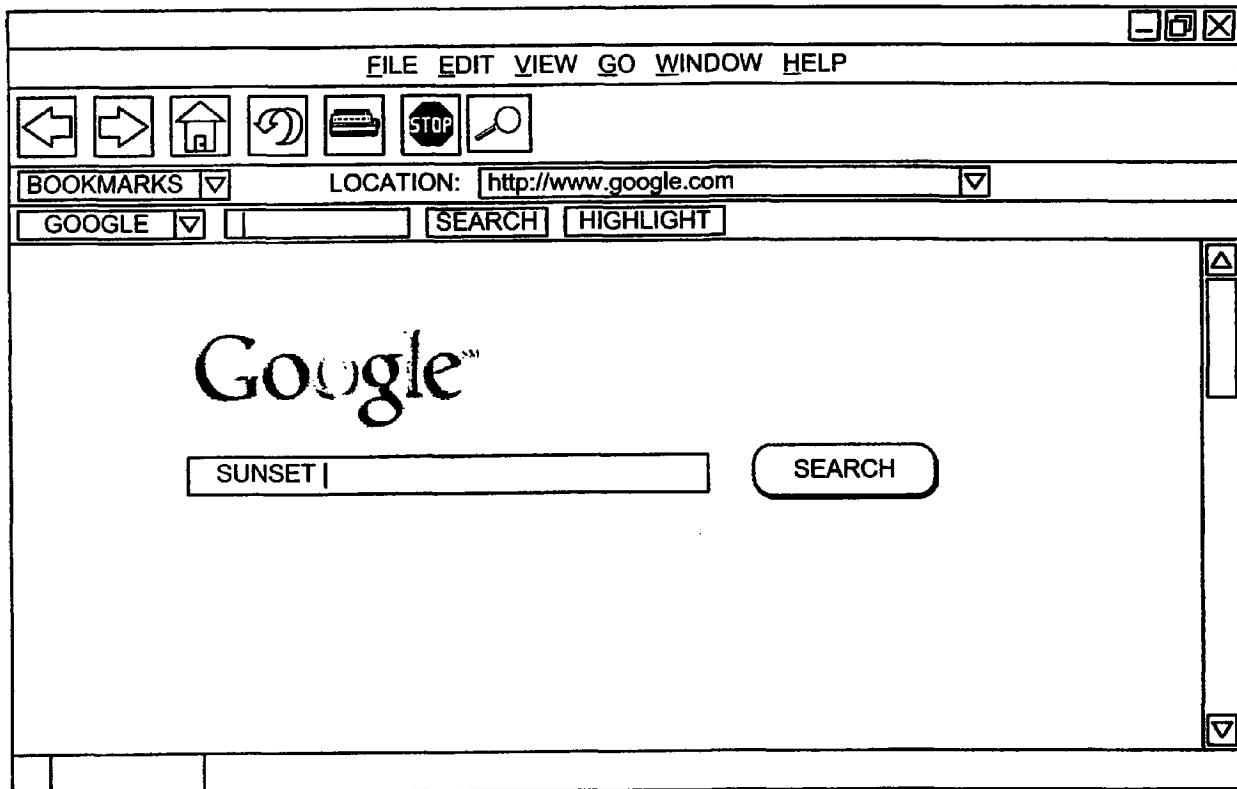


图 8

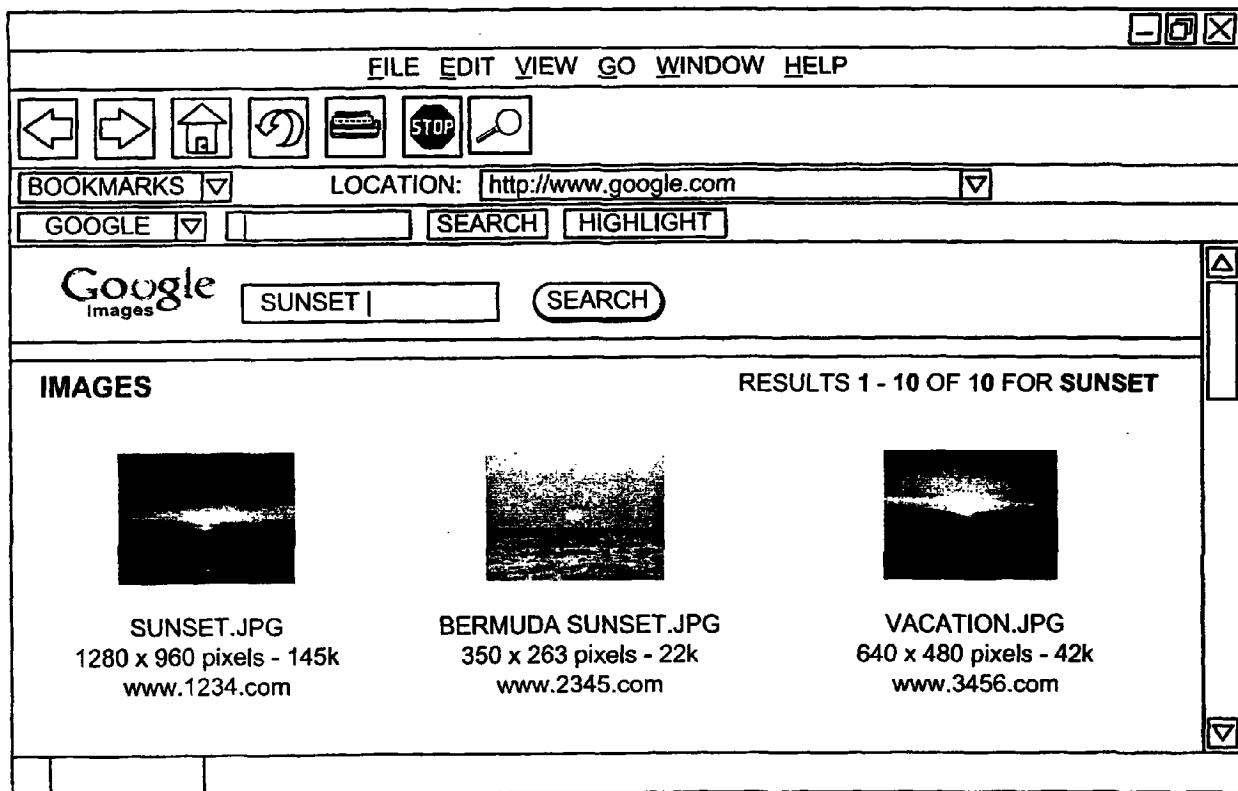


图 9

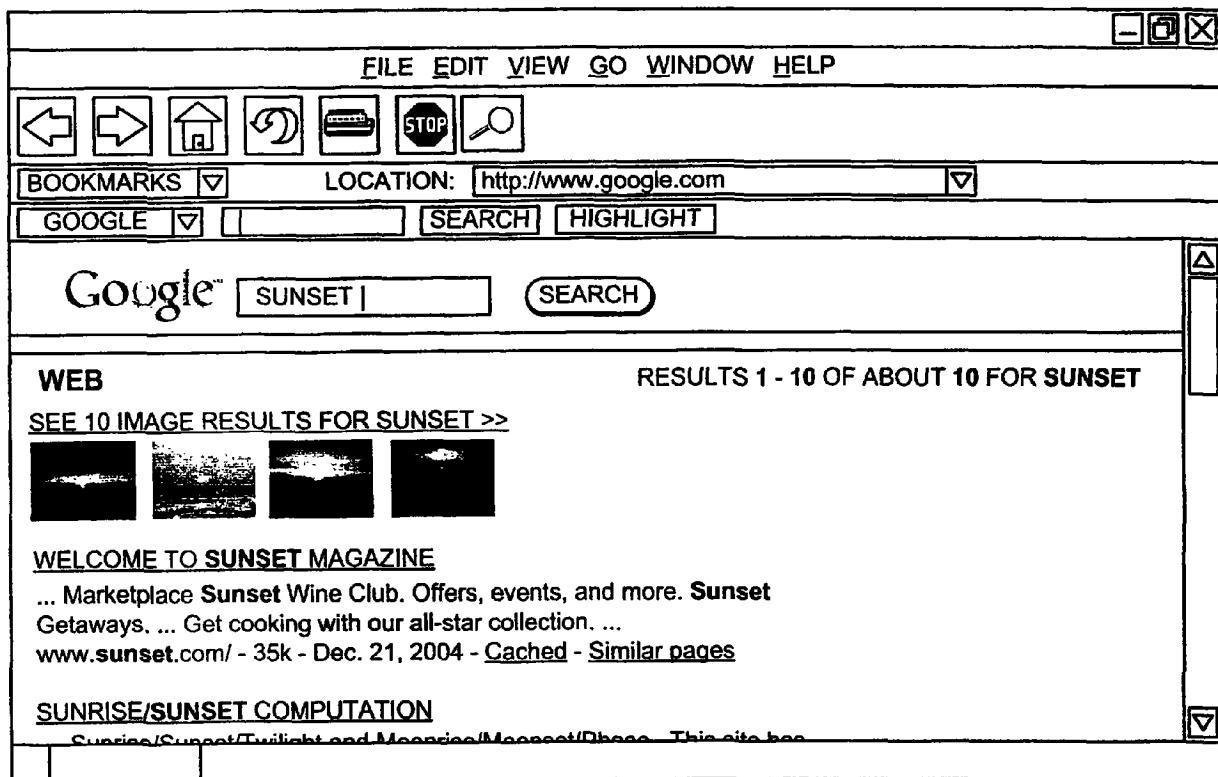


图 10