[19] 中华人民共和国国家知识产权局

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

[12] 发明专利申请公开说明书

[21] 申请号 200610058126.X

[43] 公开日 2006 年 9 月 20 日

[11] 公开号 CN 1834965A

[22] 申请日 2006.3.6

[21] 申请号 200610058126.X

[30] 优先权

[32] 2005.3.17 [33] US [31] 11/083,204

[71] 申请人 国际商业机器公司

地址 美国纽约

[72] 发明人 E·阿米泰 A·达洛 U·韦斯

[74] 专利代理机构 北京市中咨律师事务所

代理人 于 静 李 峥

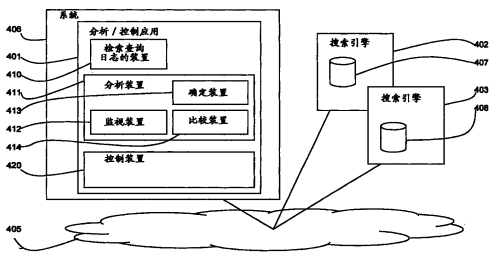
权利要求书 4 页 说明书 13 页 附图 5 页

[54] 发明名称

用于评估搜索引擎的质量的方法和系统

[57] 摘要

一种用于评估一个或多个搜索引擎的质量的方法和系统，通过从查询日志(307，407，408)检索数据来监视搜索引擎(308，402，403)的用户(201)的重构会话，其中一重构会话是用户(201)发出以满足单个信息需求的至少两个对搜索引擎(308)的查询的系列。 然后确定用于该搜索引擎(308，402，403)的重构会话参数，并分析该重构会话参数。 该重构会话参数可以是在重构会话中的查询重构的比率或重构会话持续时间。 分析单个搜索引擎的重构会话参数可确定参数是否随时间改变或根据单个搜索引擎内的不同设置确定参数。 分析两个或更多个搜索引擎的重构会话参数包括比较该两个或更多个搜索引擎的参数以度量搜索质量。 可使用该分析来控制一个或多个搜索引擎的操作。



1. 一种用于评估一个或多个搜索引擎的质量的方法，该方法包括：

监视（502）搜索引擎的用户的重构会话，其中一重构会话是用户发出以满足单个信息需求的至少两个对搜索引擎的查询的系列；

确定（503）用于该搜索引擎的重构会话参数；以及

分析（504）该重构会话参数。

2. 根据权利要求 1 的方法，包括根据所述分析控制（505）所述搜索引擎的操作。

3. 根据权利要求 1 或 2 的方法，其中，所述重构会话参数是以下组中的一个：重构会话中的查询重构的比率；重构会话持续时间；被重构的查询的内容；或被重构的查询的语法。

4. 根据权利要求 1－3 中的任何一个的方法，其中，所述监视（502）重构会话的步骤包括识别在阈值时间内的重构查询，并将这些查询归组为重构会话。

5. 根据前面任何一个权利要求的方法，其中，所述监视（502）重构会话的步骤包括识别在阈值类似性内的重构查询，并将这些查询归组为重构会话。

6. 根据前面任何一个权利要求的方法，其中，所述分析（504）重构会话参数包括确定对于单个搜索引擎该参数是否随时间改变。

7. 根据前面任何一个权利要求的方法，其中，所述分析（504）重构会话参数包括根据单个搜索引擎内的不同设置确定该参数。

8. 根据权利要求 2－7 中的任何一个的方法，其中，所述控制（505）搜索引擎的操作控制单个搜索引擎的操作参数。

9. 根据前面任何一个权利要求的方法，其中，所述分析（504）

重构会话参数包括比较两个或更多个搜索引擎的参数。

10. 根据权利要求 9 的方法，其中，所述控制（505）搜索引擎的操作从两个或更多个搜索引擎中选择供使用的搜索引擎。

11. 根据权利要求 2－10 中的任何一个的方法，其中，如果重构会话参数改变到预定的阈值之外，则所述控制（505）搜索引擎的操作提供警报。

12. 根据权利要求 2－11 中的任何一个的方法，其中，所述控制（505）搜索引擎的操作为该搜索引擎启动爬虫操作。

13. 根据权利要求 2－12 中的任何一个的方法，其中，所述控制（505）搜索引擎的操作向查询细化过程添加输入查询项。

14. 根据权利要求 2－13 中的任何一个的方法，其中，所述控制（505）搜索引擎的操作确定用户输入指令。

15. 根据权利要求 2－14 中的任何一个的方法，其中，所述控制（505）搜索引擎的操作启动搜索引擎内的索引改变。

16. 根据前面任何一个权利要求的方法，其中，所述监视（504）是在被搜索的数据集合的更新之后执行的。

17. 一种用于评估一个或多个搜索引擎（402，403）的质量的系统，该系统包括：

搜索引擎（402，403）的用户提交的查询的查询日志（407，408）；

用于监视搜索引擎的用户的重构会话的装置（412），其中一重构会话是用户发出以满足单个信息需求的至少两个对搜索引擎的查询的系列；

用于确定搜索引擎的重构会话参数的装置（413）；以及

用于分析该重构会话参数的装置（411）。

18. 根据权利要求 17 的系统，其中，该系统包括用于根据所述分析控制搜索引擎（402，403）的操作的装置（420）。

19. 根据权利要求 17 或 18 的系统，其中，所述重构会话参数

是以下组中的一个：重构会话中的查询重构的比率；重构会话持续时间；被重构的查询的内容；或被重构的查询的语法。

20. 根据权利要求 17－19 中的任何一个的系统，其中，所述查询日志（407，408）被设置在搜索引擎（402，403）内。

21. 根据权利要求 17－19 中的任何一个的系统，其中，所述查询日志在所述搜索引擎（402，403）的外部。

22. 根据权利要求 17－21 中的任何一个的系统，其中，该系统包括用于从所述查询日志（407，408）检索数据的装置（410）。

23. 根据权利要求 17－22 中的任何一个的系统，其中，所述用于分析重构会话参数的装置（411）包括确定对于单个搜索引擎该参数是否随时间改变。

24. 根据权利要求 17－23 中的任何一个的系统，其中，所述用于分析重构会话参数的装置（411）包括根据单个搜索引擎内的不同设置确定该参数。

25. 根据权利要求 17－24 中的任何一个的系统，其中，该系统包括两个或更多个搜索引擎（402，403），并且所述用于分析重构会话参数的装置（411）包括比较该两个或更多个搜索引擎的参数。

26. 根据权利要求 17－25 中的任何一个的系统，其中，所述搜索引擎（402，403）是因特网搜索引擎、内联网搜索引擎、网站搜索引擎、或专用于文件的任何集合的搜索引擎。

27. 一种存储在计算机可读存储介质上的计算机程序产品，其包括用于执行以下步骤的计算机可读程序代码装置：

监视（502）搜索引擎的用户的重构会话，其中一重构会话是用户发出以满足单个信息需求的至少两个对搜索引擎的查询的系列；

确定（503）用于该搜索引擎的重构会话参数；以及

分析（504）该重构会话参数。

28. 根据权利要求 27 的计算机程序产品，包括根据所述分析控制（505）搜索引擎的操作。

29. 一种用于控制一个或多个搜索引擎的操作的系统，该系统包括：

用于接收搜索引擎的用户对重构会话的分析的装置，其中一重构会话是用户发出以满足单个信息需求的至少两个对搜索引擎的查询的系列；以及

用于根据所述分析控制搜索引擎的操作的装置（420）。

30. 根据权利要求 29 的系统，其中，所述用于控制搜索引擎的操作的装置（420）从两个或更多个搜索引擎（402，403）中选择供使用的搜索引擎。

31. 根据权利要求 29 或 30 的系统，其中，如果重构会话参数改变到预定的阈值之外，则所述用于控制搜索引擎的操作的装置（420）提供警报。

32. 根据权利要求 29－31 中的任何一个的系统，其中，所述用于控制搜索引擎的操作的装置（420）包括用于为该搜索引擎启动爬虫操作的装置。

33. 根据权利要求 29－32 中的任何一个的系统，其中，所述用于控制搜索引擎的操作的装置（420）包括用于向查询细化过程添加输入查询项的装置。

34. 根据权利要求 29－33 中的任何一个的系统，其中，所述用于控制搜索引擎的操作的装置（420）包括用于确定用户输入指令的装置。

35. 根据权利要求 29－34 中的任何一个的系统，其中，所述用于控制搜索引擎的操作的装置（402）包括用于在搜索引擎内提供索引改变的装置。

用于评估搜索引擎的质量的方法和系统

技术领域

本发明涉及信息搜索和检索领域。具体地，本发明涉及使用从查询日志中提取的信息评估搜索引擎的质量。

背景技术

搜索万维网所涉及的人中有三个群体。有提供 Web 的所有内容的作者。有使用搜索引擎查找其感兴趣的内容的搜索者。最后，有创建和维护搜索引擎的开发者。这三个群体有时会重叠，人们根据他们的需要常常属于几个群体。

搜索引擎用户将这样的知识带入搜索过程，该知识可能没有在集合（collection）中被记录，可能没有被开发者处理和在排序函数中被处理，且可被除了提交查询的人之外的所有其他搜索者认为是不相关的。如图 1 中所示，用户 102 的知识范围和搜索引擎 101 的通过其集合和搜索过程的单个视野之间的重叠从一个个别用户 102 到另一个用户各不相同。一些用户可能会在他们如何描述内容上达成一致，但是不能在哪个查询最好地捕获该描述上达成一致。其他用户会提出完全相同的查询并且会期望找到完全不同的事物。一些人会选择在他们的查询中使用非常有限制性的语法以要求搜索引擎符合他们的请求。其他人可能会对引擎发展出信任感并让其决定应该如何处理查询。

搜索引擎可信赖度的概念对于与搜索引擎的交互是必要的。它指示人们开始搜索过程的方式，以及他们愿意花费多长时间来探查可搜索的集合以找到答案。将搜索引擎理解为具有不同范围的视野的机器使得搜索引擎用户开始进行关于他们的信息需求的小的协商。用户会试图以不同的风味

和焦点询问相同的问题以得到这样的结论，即他们已做完所有可能的事情，并且已得到可搜索范围内的最大信息。

在因特网上存在很多搜索引擎，每个搜索引擎具有其自己的操作方式。通常，搜索引擎包括：在因特网上爬行以采集信息的至少一个蜘蛛（spider）或爬虫（crawler）应用；以索引或目录的形式包含爬虫采集的所有信息的数据库；以及用于用户搜寻该数据库的搜索工具。搜索引擎以不同方式提取和索引信息并也以不同方式返回结果。

因特网技术也被用于创建称为内联网的私有公司网。内联网网络和资源不能在因特网上公开地可用，并且通过防火墙与因特网的其余部分隔开，防火墙禁止未被授权的对内联网的访问。内联网也具有在内联网的界限内进行搜索的搜索引擎。

另外，在例如大公司的单独网站内设置了搜索引擎。使用搜索引擎仅索引和检索它所相关的网站的内容以及相关联的数据库和其他资源。

2003 年 12 月 23 日提交的美国专利申请 10/743158 认识到在用户查询中存在大量的关于用户如何看待他们搜索的项目的信息，并提供了一个系统，其中使查询字与搜索引擎的索引内的信息相结合从而增加可描述项目的方式。

搜索引擎的用户经常不能以他们提出的第一个查询找到所要查找的内容。一些用户然后以各种方式—可能是通过增加或除去项—来改变他们的最初的查询，并重新提交。

从搜索者的角度看，必须重构（reformulate）查询损害了用户的体验。另外，每当雇员必须花费额外的时间来在内联网搜索引擎中重构查询时，公司直接遭受经济损失。因此，在查询日志中找到的会话的数量和长度可以是搜索质量的有价值的量度。

搜索引擎用户使用一些不同的方法来协商他们通过信息失配的路径。此协商通常被称为查询重构，但是也可使用其他术语。

查询重构不同于查询细化。查询重构是专门由单个人类使用者采取以找到所需信息的行为。另一方面，查询细化是许多检索系统使用以便改进

用户查询以使其最好地匹配索引的信息的自动过程。有可能搜索引擎对用户隐瞒此事，或者它们要求用户选择最好的细化，但是查询细化在本质上仍是自动的。查询重构源于搜索引擎用户的对世界的感知，而查询细化源于搜索引擎的对世界的感知。

重构通常在已知的一段时间内并对单个搜索引擎发生。它们被分组成被称为重构会话的会话。重构会话的定义是由一用户发出以便满足单个信息需求的至少两个查询的系列。一个示例可包括查询“hershy park”，“hersky park pa”和最终“hershey park pa”。尽管在结果中翻页可被认为是一种重构，但是如果用户进行的重构的唯一类型是翻页，则在此上下文内不认为它是重构。

影响会话长度的因素有很多，包括搜索算法、集合的质量、用户的搜索技能以及用户的耐心。但是，当所有其他因素不变时，其查询日志分析显示较高的会话比率和/或较长的会话的搜索引擎应被认为质量较差。可针对可用于搜索的不同内容使用该相同的比较。

搜索引擎存在的一个问题是需要提供对单个搜索引擎或者多于一个搜索引擎的性能的度量。本发明的一目标是通过监视查询重构以提供对一个或多个搜索引擎的质量评估，从而提供对此问题的解决方案。本发明的另一个目标是根据对查询重构的分析控制一个或多个搜索引擎的操作。

发明内容

根据本发明的第一方面，提供了一种用于评估一个或多个搜索引擎的质量的方法，该方法包括：监视搜索引擎的用户的重构会话，其中一重构会话是用户发出以满足单个信息需求的至少两个对搜索引擎的查询的系列；确定用于搜索引擎的重构会话参数；以及分析该重构会话参数。

该方法可任选地包括根据所述分析控制搜索引擎的操作。

重构会话参数可以是在重构会话中的查询重构的比率，该比率是通过将作为重构会话的一部分的查询的数量除以查询日志中的查询的总数计算出的。另一个重构会话参数可以是重构会话持续时间，其是用每个重构会

话的查询数量或一重构会话的持续时间计算的。可将统计方法应用于这些重构会话参数。

重构会话参数可与被重构的查询的内容的性质或趋势有关。例如，同义词、拼写错误、扩展项或收缩项的使用。

重构会话参数可与被重构的查询中语法的使用的性质或趋势有关。例如，减号、加号或引号的使用。

该方法可包括将与重构会话有关的数据记录在搜索引擎的外部或内部的日志内。

所述监视重构会话的步骤可包括识别在阈值时间或阈值类似性内的重构查询，并将这些查询归组为重构会话。

分析重构会话参数可包括确定对于单个搜索引擎参数是否随时间改变，或者根据单个搜索引擎内的不同设置确定该参数。所述监视可在被搜索的数据集合的更新之后执行。控制搜索引擎的操作可控制单个搜索引擎的操作参数。

分析重构会话参数可包括比较两个或更多个搜索引擎的参数。控制搜索引擎的操作可从两个或更多个搜索引擎选择供使用的搜索引擎。

控制搜索引擎的操作可包括一个或多个以下操作：如果重构会话参数改变到预定阈值之外则提供警报；为搜索引擎启动爬虫操作；向查询细化过程添加输入查询项；确定用户输入指令；或启动搜索引擎内的索引改变。

根据本发明的第二方面，提供了一种用于评估一个或多个搜索引擎的质量的系统，该系统包括：搜索引擎的用户提交的查询的查询日志；用于监视搜索引擎的用户的重构会话的装置，其中一重构会话是用户发出以满足单个信息需求的至少两个对搜索引擎的查询的系列；用于确定搜索引擎的重构会话参数的装置；以及用于分析重构会话参数的装置。

该系统任选地包括用于根据所述分析控制搜索引擎的操作的装置。

可在搜索引擎内部或在搜索引擎外部设置查询日志。该系统可包括用于从查询日志检索数据的装置。

所述用于分析重构会话参数的装置包括确定对于单个搜索引擎参数是

否随时间改变，或者根据单个搜索引擎内的不同设置确定该参数。所述用于监视的装置可在已更新的被搜索的数据集合上执行。

该系统可包括两个或更多个搜索引擎，并且所述用于分析重构会话参数的装置可包括比较两个或更多个搜索引擎的参数。

所述搜索引擎可以是因特网搜索引擎、内联网搜索引擎、网站搜索引擎、或专用于文件的任何集合的搜索引擎。

根据本发明的第三方面，提供了一种存储在计算机可读存储介质上的计算机程序产品，其包括用于执行以下步骤的计算机可读程序代码装置：监视搜索引擎的用户的重构会话，其中一重构会话是用户发出以满足单个信息需求的至少两个对搜索引擎的查询的系列；确定用于搜索引擎的重构会话参数；以及分析该重构会话参数。

该计算机程序产品还可包括根据所述分析控制搜索引擎的操作。

根据本发明的第四方面，提供了一种用于控制一个或多个搜索引擎的操作的系统，该系统包括：用于接收搜索引擎的用户对重构会话的分析的装置，其中一重构会话是用户发出以满足单个信息需求的至少两个对搜索引擎的查询的系列；以及用于根据所述分析控制搜索引擎的操作的装置。

所述用于控制搜索引擎的操作的装置可通过提供用于一个或多个以下操作的装置来控制所述操作：从两个或更多个搜索引擎中选择供使用的搜索引擎；如果重构会话参数改变到预定阈值之外则提供警报；为搜索引擎启动爬虫操作；向查询细化过程添加输入查询项；确定用户输入指令；或在搜索引擎内提供索引改变。

附图说明

下面将参照附图仅作为示例说明本发明的实施例，在附图中：

图 1 是示出搜索引擎及其用户感知的知识范围的示意图；

图 2 是示例性的 Web 体系结构的框图；

图 3 是可根据本发明使用的搜索引擎体系结构的框图；

图 4 是根据本发明的系统的框图；以及

图 5 是根据本发明的方法的流程图。

具体实施方式

如上所述，图 1 示出搜索引擎的各个用户 102 的不同的知识库和搜索引擎 101 自身的知识。搜索引擎的用户从他们的知识库开始进行搜索查询。因此，在搜索引擎检索到用户查找的信息之前，经常需要该查询的重构。单个查询的重构被称为重构会话。所描述的方法和系统使用用户的重构会话提供的信息来评估搜索引擎的质量。

参照图 2，其示出 Web 体系结构 200 的示例性实施例。客户计算机系统 201 通常包括中央处理单元（CPU）210，并具有操作系统、存储器、输入/输出接口、总线、输入/输出设备。客户计算机系统 201 包括浏览器应用 202，该应用经由使用网络 205（例如因特网）的连接 209（例如 TCP（传输控制协议）连接）与主机服务器系统 204 交互。客户计算机系统 201 包括图形用户界面（GUI）203，其显示浏览器应用 202 提供的信息。

主机服务器系统 204 的功能是将浏览器应用 202 请求的信息发送给客户计算机系统 201。主机服务器系统 204 是通常包括中央处理单元（CPU）211 并具有操作系统和数据库 206 的计算机系统。主机服务器系统 201 包括服务器应用 207，其处理来自客户计算机系统 201 的浏览器应用 202 的请求并与主机操作系统通信。主机服务器系统 204 是 HTTP（超文本传输协议）服务器，其使用 HTTP 传输 208 将信息发送给客户机浏览器应用 202。在万维网的上下文中，主机服务器系统 204 是 Web 服务器。

通常，客户机浏览器应用 202 请求主机服务器系统 204 返回 HTML（超文本标记语言）文件。主机服务器系统 204 接收该请求并返回响应。主机服务器系统 204 从其数据库 206 检索请求的信息 212，并将该信息 212 发送给客户机浏览器应用 202，该客户机浏览器应用在客户机的 GUI 203 内显示该信息 212。

参照图 3，其示出搜索引擎系统 300 的示例性实施例。所提供的服务器系统 301 通常包括中央处理单元（CPU）302 并具有操作系统和数据库

303。服务器系统301提供了搜索引擎308，该搜索引擎包括：用于经由网络205从服务器310、311、312收集信息的爬虫应用304；用于在数据库303中创建收集的信息的索引或目录的应用305；以及搜索查询应用306。

数据库303中存储的索引通过从服务器310、311、312中的文件提取的信息引用这些文件的URL（统一资源定位器）。

搜索查询应用306经由网络205接收来自客户机201的查询请求320，将其与数据库303中存储的索引中的条目相比较，并在HTML页面中返回结果。当客户机201选择到文件的链接时，客户机浏览器应用202被直接路由到存放该文件的服务器310、311、312。

搜索查询应用306使用搜索引擎303保持从客户机机接受到的搜索查询的查询日志 307。作为另一种选择，可通过首先在日志中保存查询并然后将信息发送给搜索引擎300，保持与搜索引擎300分开的查询日志。

了解客户机的查询重构的最好方式是分析搜索引擎 303 的查询日志307。为了调查查询日志307中的重构，必须首先将日志307划分为重构会话。用于提取这些会话的方法除了依赖于每个查询的文本和时间戳之外还依赖于查询日志307为每个查询提供的信息。相关的附加信息是单个会话或单个用户的标识。

所描述的实施例集中于其中没有提供附加信息的情况，并且它不依赖于搜索引擎自身之外的任何事物。这种情况的一个示例是开盒即用的搜索引擎，其假设不了解运行它的应用。

最好的情况是搜索引擎在其日志内保持会话信息，实际上跟踪何时用户返回到搜索结果的页面以及改变查询。在此情况下，不需要进行额外的处理，并且将查询归组为重构会话是简单直接的。不过，一些用户会在单个被记录的会话内寻求满足若干信息需求，在此情况下它们可能需要被划分。

更常见的可能性是日志通过一些标识符例如IP（网际协议）地址包含标识其用户的信息。在此情况下，假设在用户发出一查询之后，他们在短时间范围内发出的所有其他查询将是该查询的重构。一旦已确定该时间界

限，就可用简单的算法将查询归组。在许多情况下，即使已知IP地址，也不能使用该IP地址来识别单个用户，例如通过代理服务器的请求。在这种情况下，必须如下文所述地近似得出会话。

查询日志常常不会包含任何用于识别用户的信息。对于这种日志，仅能通过在该日志中找到很可能是其他查询的重构的查询来近似得出会话。

观察到大多数重构使查询的大部分未改变，而使用近似字符串匹配算法。工作良好的一种形式的算法是*tf*idf*加权三字母组匹配。**Jaro-Winkler**算法也表现得很好并被调查。当用户完全重写查询时。这种方法不能发现重构。

简单的说，重构会话提取算法被赋予两个阈值—时间阈值和类似性阈值。如果一系列查询均在时间阈值内发生，并且每两个连续的查询都处于类似性阈值内，则将该一系列查询归组到单个会话。

```
Sessions <- ϕ
Log <- {按时间排序的所有查询}
while (Log != ϕ)
  Q1 <- 从Log移除第一个查询
  Q_start <- Q1
  New Session <- {Q1}
  for each Q2 in Log
      if (time(Q2)-time(Q_start) < time threshold)
        if (compare(Q1, Q2) < similarity threshold)
           New Session <- New Session U {Q2}
           Log <- Log \ {Q2}
           Q1 = Q2
  if (|New Session| > 1)
      Session <- Session U {New Session}
```

在下面给出的示例中，此分析中报告的发现是在10分钟的时间阈值内完成的。已实验了从5分钟直到30分钟的各种窗口尺寸，并且已发现在长

度、持续时间和持续时间分布方面在所有时间阈值上各值几乎相同。唯一随时间阈值改变的值是重构会话在整个查询日志之中的百分比，其随着时间的增加而轻微地增加。使用了 10 分钟时间阈值，因为它代表了查询重构特性，并且在提取错误方面更为可靠。例如，几个不同的用户在非常短的时间范围内提交相同查询的可能性不大。时间范围越短，则会话提取就越精确并且处理越快。

示例

此示例跟踪了具有两个非常不同的用户群体的两个不同搜索引擎—计算机公司的内联网搜索引擎和相同计算机公司的外部网站搜索引擎—的内联网和 Web 查询日志。内联网搜索引擎每个月唯一地从公司的雇员那里接收到大约 50 万条查询。外部因特网网站每个月从全世界的公司顾客那里接收到约几百万条查询。

这里分析的日志是从具有两个不同用户群体的两个不同搜索引擎获得的。内联网搜索引擎被采样并且在不同的几天内有大约 200000 条查询被记入日志。公共网站仅被记录了大约 1 周并且收集了超过 500000 条查询。内联网搜索日志是从主机器生成的；公共网站搜索日志是从作为一些机器的群集的一部分的两个不同的机器获得的。两个搜索引擎的用户在性质上是不同的。内联网用户非常有技术意识，而公共网站搜索引擎的用户则是来购买产品、寻求技术支持、和了解公司的财务状况。

下面给出可被分析的会话参数的示例，以及为了评估质量或获得关于用户行为的信息而在搜索引擎之间进行的比较。

分析了每个内联网搜索日志中的重构的比率。记入日志被限制为每个日志大约 25000 条查询。会话中的查询的百分比是通过将被发现为重构的一部分的查询的数量除以日志中的查询的总数计算出的。

仅仅计算来自不同引擎的日志的平均可得到惊人相似的结果，其中提交给内联网搜索引擎的查询的 31.7％是重构会话的一部分，而 31.3％的查询是公共网站搜索引擎上的重构会话的一部分。

也分析了工作日之间的差别并且在搜索引擎之间进行了比较。

以每个会话的查询数度量的重构会话长度是人们愿意花费的与搜索引擎交互的时间的指示。由于结果的“下一页”的所有发生以及查询的重构都包含在计算的会话内（但是要求每个会话具有至少一个重构），所以还可提供关于决定完全改变查询而不是浏览搜索引擎提供的结果的过程的指示。

在每个日志中，监视每个会话的查询的数量的样本方差和标准偏差。

还比较了内联网和公共网站内的每个会话的查询的平均数量。

可有助于解释两个不同引擎之间的轻微差别的一个因素是浏览搜索结果的比率。由于“下一结果页”被计算为会话中新发出的一个查询，所以还度量了浏览内联网搜索结果和公共网站搜索结果的比率之间的差。

用于包括发送给搜索引擎的所有查询的一般日志的该比率对于内联网和公共网站是大约14％到16％。此发现表明用户浏览搜索结果和发出查询重构之间正相关。

重构会话持续时间是用户选择与搜索引擎协商信息需求所花费时间长度的量度。为此，使用每个会话中的第一条和最后一条查询的时间戳来计算会话持续时间。

比较日志上的重构会话的中间持续时间和平均持续时间的一致性。

可获得每个会话的查询的平均数，并用会话持续时间去除该平均数以近似得出平均用户在每个查询中浏览搜索结果和确定是否满足了信息需求将花费的时间。此参数可在搜索引擎之间相比较。

查询的重构反应了用户对搜索引擎的感知。用户在无意中使用两种不同方法来解决发现信息的问题。一种方法是试图解读作者群体如何描述集合中的概念。另一种方法相当于试图对搜索引擎开发者群体选择的排列和分析集合的信息的方式进行逆向工程。第一种方法相当于使用内容重构与创作者对话，而第二种方法相当于使用语法重构与开发者交谈。此划分可帮助更好地理解每种方法提出的问题。也可检测和分析内容和语法重构。

与内容相关的重构可以有以下几种类型：查找同义项，简单地拼错项，

扩展查询以使搜索范围变窄，以及简化查询以拓宽搜索范围。

语法重构包括在查询中插入搜索运算符例如减号、加号和引号。

现参照图 4，系统 406 示出为本发明的示例性实施例。系统 406 包括用于分析和控制一个或多个搜索引擎 402、403 的应用 401。应用 401（或一系列应用）可相对于分析中的一个或多个搜索引擎 402、403 经由网络 405 远程地或本地地设置在客户机系统或服务器系统上。如在上文给出的示例中，分析中的搜索引擎 402、403 可以是因特网搜索引擎、公共网站搜索引擎、内联网搜索引擎、专用于任何文件的集合的搜索引擎或上述各引擎的组合。

应用 401 包括用于检索分析中的一个或多个搜索引擎 402、403 的查询日志 407、408 的装置 410。在此示例性实施例中，查询日志 407、408 被示为在搜索引擎 402、403 的内部；但是查询日志 407、408 可设置在搜索引擎外部，例如设置在用户系统或外部服务器上。可从设置在包括搜索引擎的群集内机器子集中获得分析的查询日志。应用 401 包括用于分析来自查询日志 407、408 的数据的分析装置 411。分析装置 411 包括用于监视重构会话 412 的装置、用于确定会话比率或其他会话参数 413 的装置以及比较装置 414。应用 401 可包含依赖于需要的分析的其他形式的数据操作。

在一个示例性实施例中，应用 401 还包括用于控制分析中的搜索引擎 402、403 的控制装置 420。控制装置 420 可作为另一种选择与分析装置 411 分开设置，例如设置在搜索引擎 402、403 本地或远程的另一系统上。控制装置 420 可根据一个或多个下面的基于分析结果的操作控制搜索引擎 402、403。

- 控制装置 420 可根据分析从多个搜索引擎中选择搜索引擎。
- 控制装置 420 可根据一个搜索引擎的分析选择用于单个搜索引擎的操作参数。
- 如果所监视的重构会话的参数根据预先设置的阈值改变，控制装置 420 可发出警报。
- 如果分析指示出需要重构的重复地未被识别的输入查询，控制装置

420 可启动爬虫应用。

• 如果分析识别出查询重构中的重复地改正的项，控制装置 420 可向搜索引擎的查询细化过程自动添加输入查询项。

• 控制装置 420 可根据对查询重构中的语法参数的分析，选择将包含在用户界面内的指令（例如查询语法示例）。

• 控制装置 420 可根据查询重构的高的重构比率启动索引改变。

图 5 是一个或多个计算机过程执行的分析重构会话的方法的流程图 500。在 501，从查询日志接收到查询重构会话数据。在 502 监视数据，并在 503 确定预定义的重构会话参数。所述监视和确定 502 和 503 可执行有限的一段时间或持续进行。在 504 分析被确定的参数，并且在 505 根据分析的结果控制一个或多个搜索引擎的操作。

简单的用于搜索引擎的质量测试将是监视查询日志以度量查询的重构比率。如果该比率随着时间增加，则这需要更全面地分析重构的性质。另一种使用重构比率量度的方法是比较两个不同的搜索引擎的性能，或在相同集合上具有相同用户群体而具有不同设置的相同搜索引擎的性能。假设较好的搜索引擎或搜索设置将需要用户付出较少的重构努力。还可能在定期更新索引之后运行重构比率分析，以了解用户是否错过先前存在那里的且没有被索引的或者被不同地命名的某些内容。

重构会话的分析还揭示了用于内容增强的丰富的源。例如，可能用户主要通过产品的旧的或常见名称要求产品，而索引仅包含用新产品名称标示的信息。这是一个可通过分析重构列表相当容易发现的非常常见的问题。可将此重要信息转发给网站编辑者，并建议向它们的现有内容添加项。

通过分析会话，可发现可搜索的集合中没有包含的项和主题。此信息使得能够通过添加新的文件和新的内容来增强集合。

知道哪些查询或孤立的项被搜索但是不易于被发现或者根本不被发现可提供可能需要专注式爬虫（focused crawler）的证据。爬虫可被配置为优选包含从重构会话中提取出的所需的项的文件。另外，爬虫可被设置为访问被识别为包含来自重构会话的项的新网站。

还可能存在这样的情况，即用户查找的信息只是缺失，并且更严格的分析可指示在可搜索信息中存在“漏洞”。在此情况下，应创建新内容以满足该信息需求。

可搜索的集合的管理员可通过分析重复地重新出现的重构序列来识别集合中不包含的主题。然后，管理员可指令将写入新内容以包含这些主题。也可购买或获得这样的内容例如帮助文件、驱动器的支持页等。也可在在线零售店中想象到这种情况，其中从会话识别新的趋势并扩展当前的存货以满足需要。

可使用在查询日志中发现的重构会话作为用于查询细化的候选。如果发出类似查询的一些用户在对结果感到满意之前结束了重构它们，则很可能更多的用户将遇到类似的困难。搜索引擎可利用以前的用户已经完成的检测工作自动将这些重构建议为细化。这种方法比当前的查询细化的方法更加以用户为中心，当前方法通常根据搜索引擎已经索引的内容确定将建议什么细化。

可分析在查询日志中发现的重构会话信息而不对日志中存储的用户信息进行假设。可以许多方式利用从它们得到的信息以提高用户体验并改进 Web 内容。该信息还可用作搜索引擎或搜索引擎已索引的内容的质量的量度。

本发明通常实现为包括用于控制计算机或类似设备的一组程序指令的计算机程序产品。这些指令可通过被预装载在系统上或记录在存储介质例如 CD-ROM 上而被提供，或被提供到网络例如因特网或移动电话网络上供下载。

可对前文内容进行改进和修改而不会背离本发明的范围。

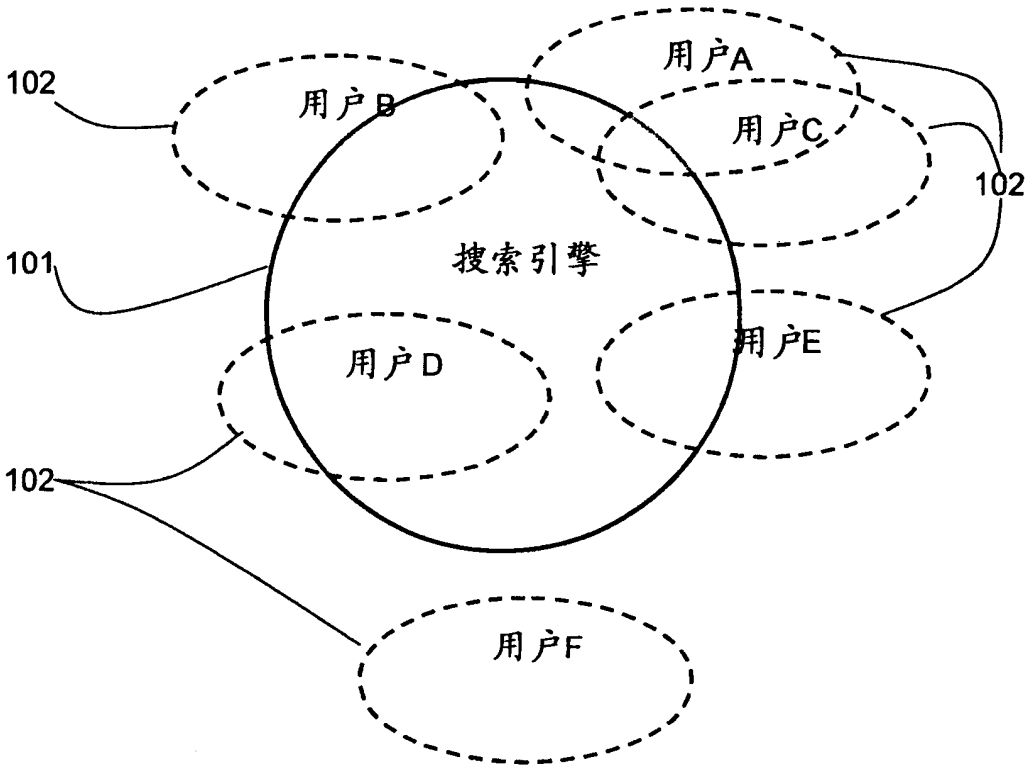


图 1

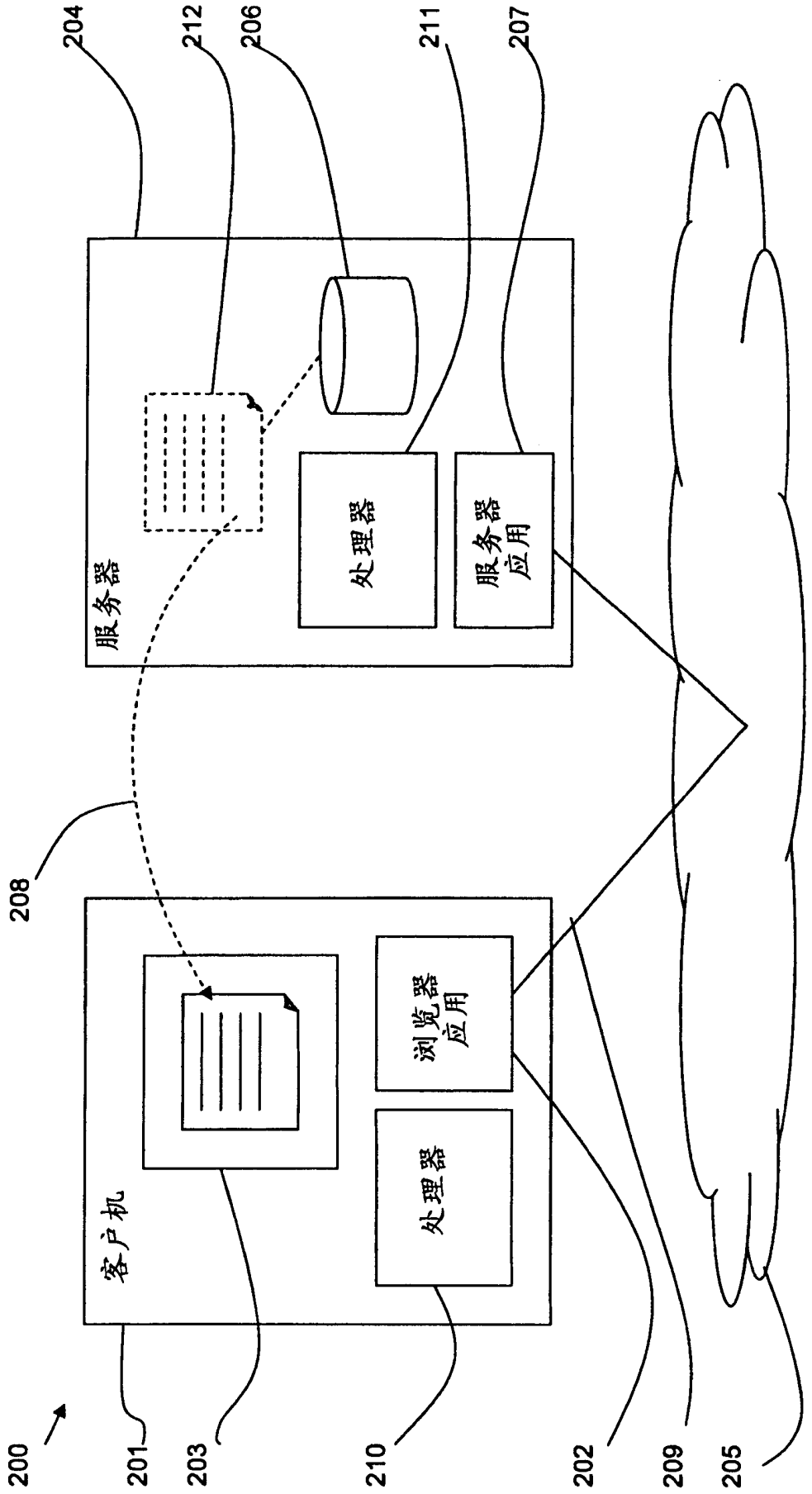

200
201
203
210
202
209
205
208
客户机
处理器
浏览器应用
服务器
处理器
服务器应用
204
212
206
211
207

图 2

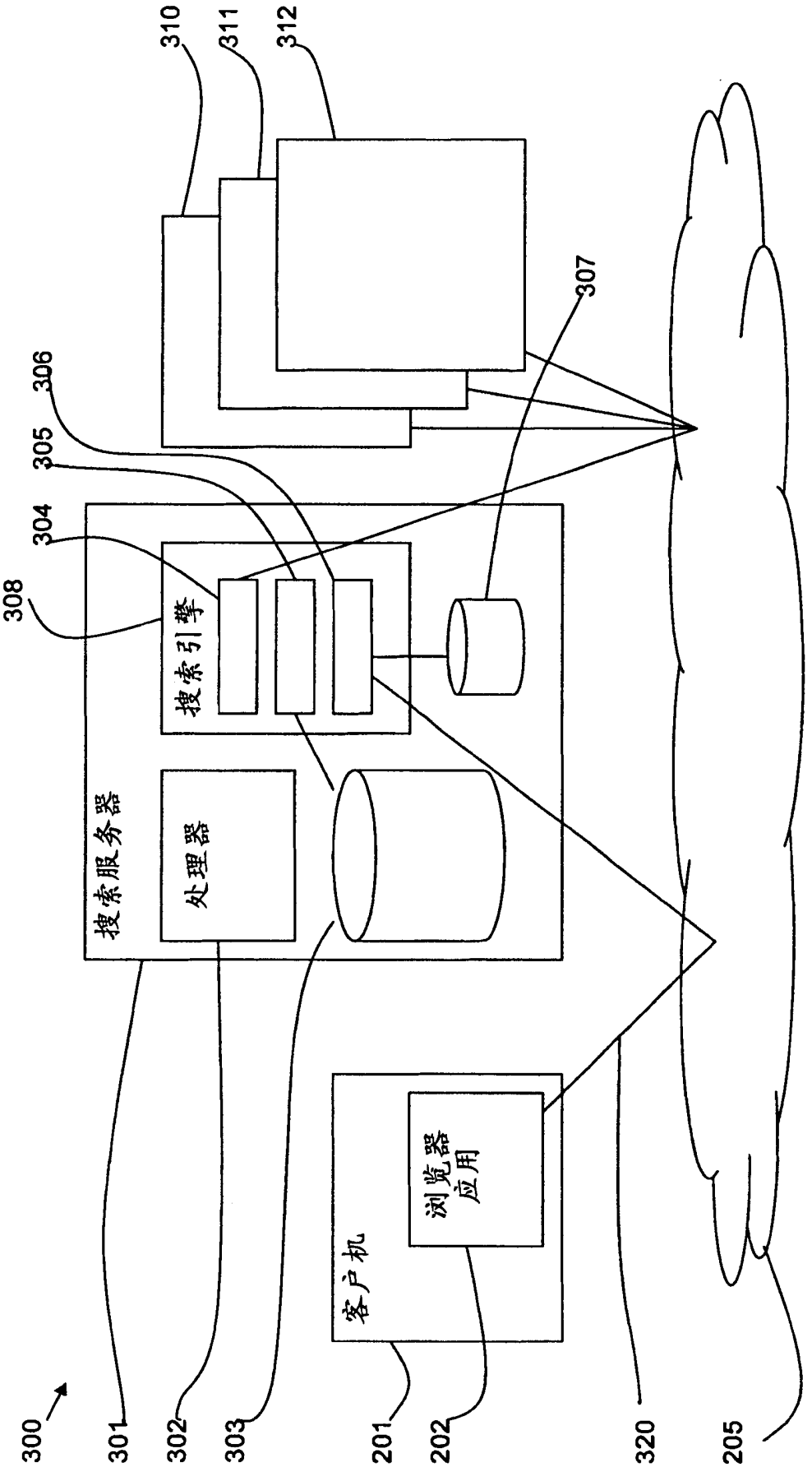

300
301
302
303
201
202
320
205
搜索服务器
处理器
搜索引擎
客户机
浏览器应用
304
305
306
307
308
310
311
312

图 3

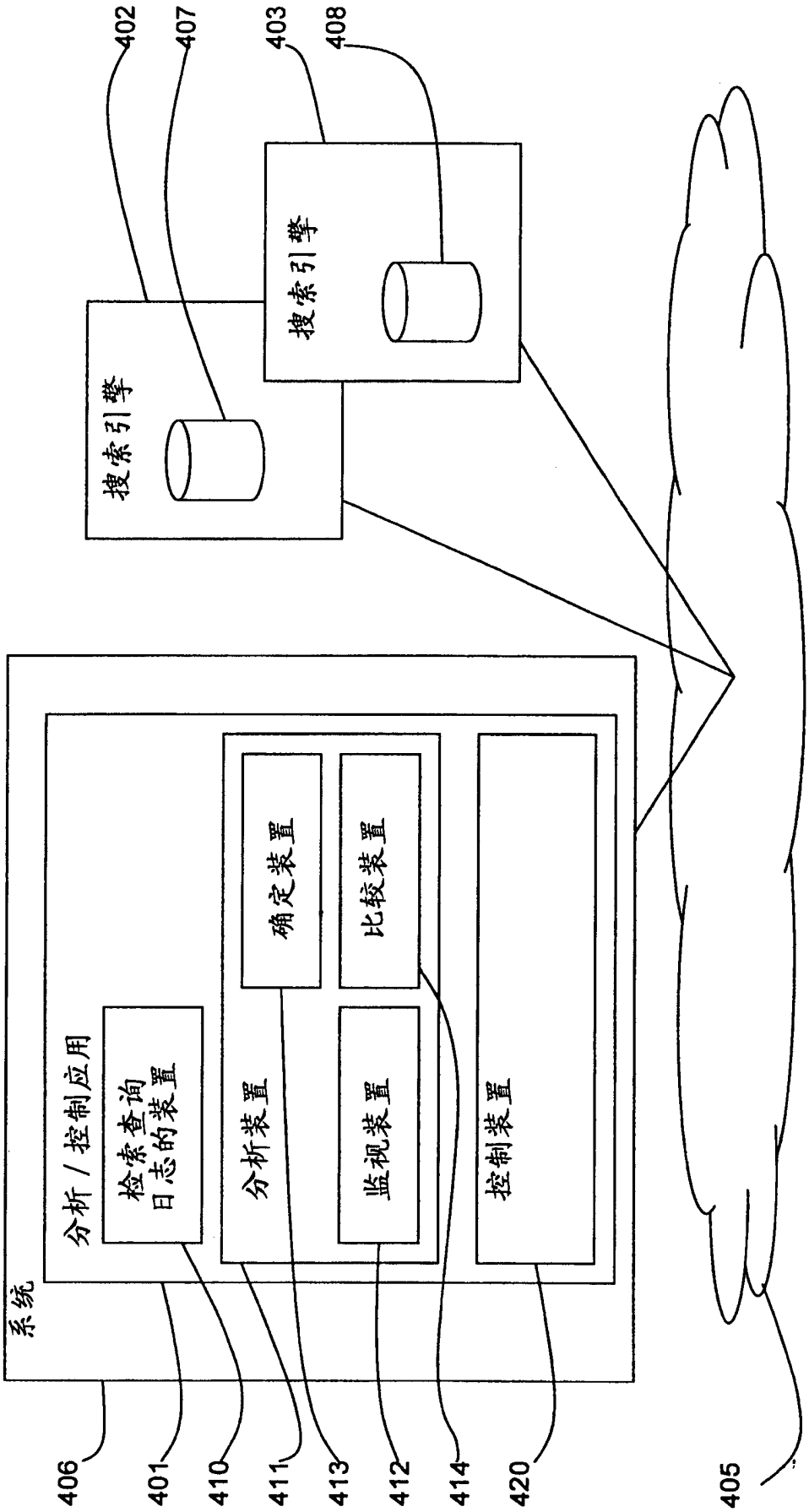

402
407
403
408
搜索引擎
搜索引擎
系统
分析 / 控制应用
检索查询日志的装置
分析装置
确定装置
比较装置
监视装置
控制装置
406
401
410
411
413
412
414
420
405

图 4

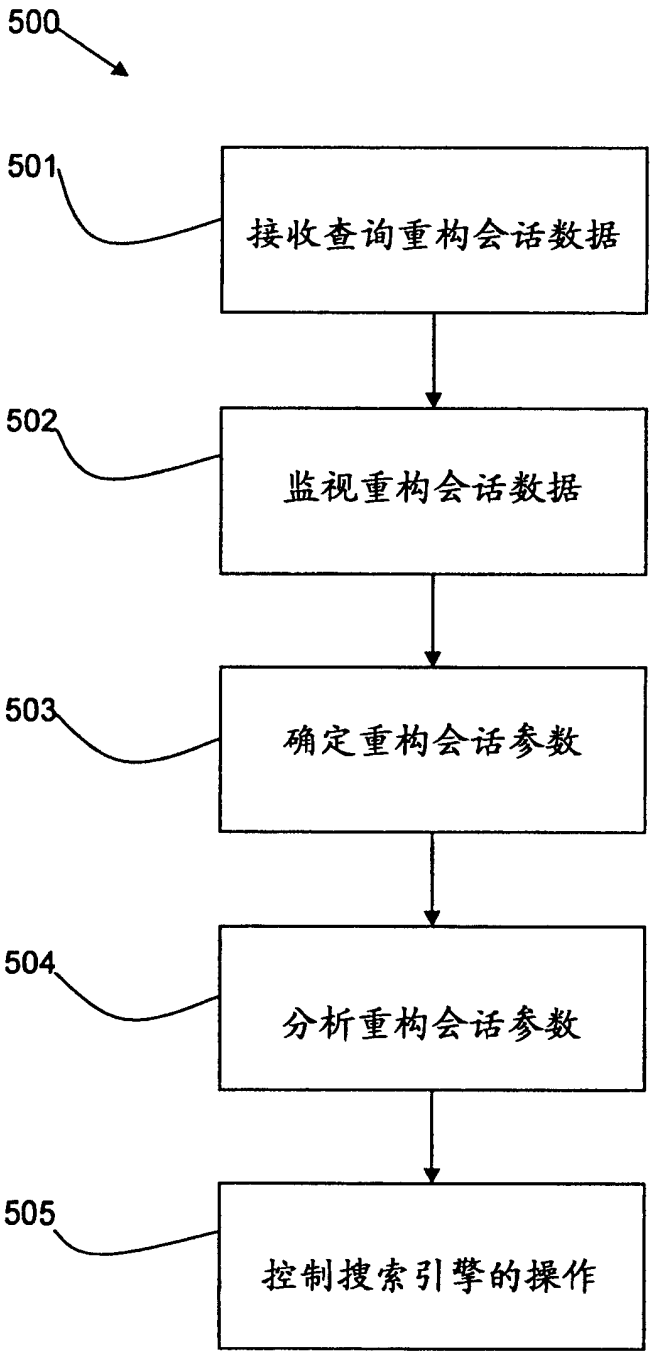

500
501
接收查询重构会话数据
502
监视重构会话数据
503
确定重构会话参数
504
分析重构会话参数
505
控制搜索引擎的操作

图 5