

[19] 中华人民共和国国家知识产权局

[51] Int. Cl⁷

G06F 17/30



[12] 发明专利申请公开说明书

[21] 申请号 200510004066.9

[43] 公开日 2005 年 7 月 13 日

[11] 公开号 CN 1637744A

[22] 申请日 2005.1.7

[21] 申请号 200510004066.9

[30] 优先权

[32] 2004. 1. 9 [33] US [31] 10/754,159

[71] 申请人 微软公司

地址 美国华盛顿州

[72] 发明人 H·陈 R·钱德拉西卡

S·H·科斯顿

[74] 专利代理机构 上海专利商标事务所有限公司

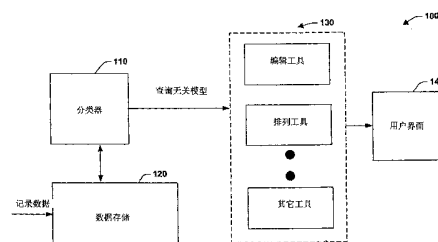
代理人 谢喜堂

权利要求书 3 页 说明书 10 页 附图 9 页

[54] 发明名称 为在大量电子文档中搜索而确定文档相关性的机器学习方法

[57] 摘要

本发明涉及应用自动化学习过程用于确定文档相关性并帮助信息检索的系统和方法。提供了一种促进确定文档相关性的机器学习方法的系统。所述系统包括一个接收人工选定条目集以使用作高相关性文档的正面测试案例的存储组件。一培训组件用人工选定条目作为正面测试案例、用一个或多个其它条目作为负面测试案例来培训至少一个分类器以便提供查询无关模型，其中其它条目可由例如统计搜索选定。另外，受训后的分类器可用于协助人们标识和选择新的正面案例、或用来过滤或重新排序根据基于统计的搜索所得到的结果。



ISSN 1008-4274

1. 一种便于机器学习方法确定文档相关性的系统，其特征在于，包括：
一存储组件，其接收人工或机器选定条目集并用作正面测试案例；以及
5 一培训组件，其用所述人工或机器选定条目作为正面测试案例、用一个或多个其它条目作为负面测试案来培训至少一个分类器，以便提供一查询无关模型。
2. 如权利要求 1 所述的系统，其特征在于，所述负面测试案例通过一统计搜索选定。
3. 如权利要求 1 所述的系统，其特征在于，所述受训分类器被用来帮助人们
10 选择新的正面案例。
4. 如权利要求 1 所述的系统，其特征在于，所述受训分类器被用来过滤基于统计搜索所获得的文档。
5. 如权利要求 3 所述的系统，其特征在于，所述过滤器的输出被排序，使得正面案例被排列在负面案例之前。
- 15 6. 如权利要求 4 所述的系统，其特征在于，所述输出根据它们是正面案例的几率进行排序。
7. 如权利要求 1 所述的系统，其特征在于，所述存储组件包括用户感兴趣的相关网站、文档或数据条目的记录。
8. 如权利要求 7 所述的系统，其特征在于，所述存储组件包括集中存储信息
20 或来自分散源的信息，诸如各种网站、文档收藏、百科全书、本地数据源和远程数据源等。
9. 如权利要求 1 所述的系统，其特征在于，所述分类器被用来自动分析所述存储组件中的数据，以便协助与用户界面交互的一个或多个工具。
10. 如权利要求 9 所述的系统，其特征在于，所述工具包括管理工具、编辑工
25 具、和排列工具中的至少一个。
11. 如权利要求 9 所述的系统，其特征在于，所述工具以在线方式和离线方式中的至少一种被采用。
12. 如权利要求 1 所述的系统，其特征在于，所述分类器根据正面和负面测试数据进行培训，以便根据诸如文档或建议其它有用信息网站的链接确定条目的相关性。
30 性。

13. 如权利要求 12 所述的系统, 其特征在于, 还包括一人工选定文档或条目集以培训机器学习分类器。

14. 如权利要求 12 所述的系统, 其特征在于, 所述分类器应用于新词语, 以标识最佳方案或相关文档。

5 15. 如权利要求 12 所述的系统, 其特征在于, 还包括在各种培训迭代中对新模型加以自举, 以促进建立用于更准确地进行信息检索活动的学习表达的成长模型。

16. 如权利要求 15 所述的系统, 其特征在于, 还包括由编辑人员人工选定的最佳方案。

10 17. 如权利要求 16 所述的系统, 其特征在于, 还包括一组件, 以最佳的可能性显示用户可能认为会感兴趣到查看或检索的文档或条目类型。

18. 如权利要求 1 所述的系统, 其特征在于, 所述分类器包括以下学习技术的至少一种: 支撑向量机器 (SVM) 技术、朴素贝叶斯技术、贝叶斯网络技术、决策树技术、基于相似性的技术、基于向量的技术、隐藏马尔可夫模型技术、和/或其它学习技术。

15

19. 如权利要求 1 所述的系统, 其特征在于, 还包括一组件, 以执行信息的后处理以确定一文档或一网站与用户或管理员的相关性。

20. 如权利要求 19 所述的系统, 其特征在于, 所述后处理包括根据预定几率阈值进行排序, 其中相关几率较高的条目显示在几率较低的条目之前。

20 21. 如权利要求 19 所述的系统, 其特征在于, 还包括加到所显示的条目上以指示文档或网站的相关性或重要性的显式标注。

22. 一种带有存储于其上的计算机可读指令的计算机可读介质, 其特征在于, 所述指令用于实现权利要求 1 所述培训组件和所述存储组件。

23. 一种基于计算机的信息检索系统, 其特征在于, 它包括:

25 用来确定数据项的培训集的装置;

用来自动对所述培训集分类的装置;

用来从所述已分类培训集确定新条目的装置; 以及

用来根据信息检索请求呈现所述新条目的装置。

24. 如权利要求 23 所述的系统, 其特征在于, 还包括用来测试所述已分类培
30 训集的装置。

25. 一种促进自动化信息检索的方法, 其特征在于, 包括:

处理来自数据记录的 n 个查询, n 为整数;

从所述 n 个查询标识相关的候选信息; 以及

培训所述分类器用来为随后的搜索标识其它相关的候选信息。

26. 如权利要求 25 所述的方法, 其特征在于, 还包括把一分析转送给确定一
5 则信息对于给定查询或主题是否值得呈现给编辑人员。

27. 如权利要求 25 所述的方法, 其特征在于, 还包括从可能文档或网站的列表中
提取相关候选信息、并自动把所述最佳方案置于其它统计排序信息之前。

28. 如权利要求 25 所述的方法, 其特征在于, 还包括按照一文档成为文档的
几率对结果重新排序, 其中相应文档被下载、且提取词语并在其中查寻出现在文档
10 中词语。

29. 如权利要求 25 所述的方法, 其特征在于, 还包括确定至少一个要分类的
类别。

30. 如权利要求 29 所述的方法, 其特征在于, 还包括采用培训数据集的一子
集来测试所述分类类别。

- 15 31. 一种具有存储于其上的一数据结构的计算机可读介质, 其特征在于, 所述
数据结构包括:

一第一数据域, 其涉及一相关性类别的培训数据集;

一第二数据域, 其涉及关于所述相关性类别的一新数据项集; 以及

一第三数据域, 其涉及所述新数据项集的几率排序。

为在大量电子文档中搜索而确定文档相关性的机器学习方法

技术领域

- 5 本发明一般涉及计算机系统，尤其涉及采用自动化分类技术来帮助寻找信息的用户和/或管理这些信息的管理员进行有效的信息检索。

背景技术

- 10 搜索系统或信息检索系统是使用户能够找出与主题相关的所需信息的通用工具。为了找到所需信息，搜索引擎或其它搜索系统常被用来使用户能指引用户设计的查询。不幸地是，当因为用户可能对如何创作或设计特定查询并无把握、或者因为查询可能是含糊不清的而有许多不相关文件被检索到时，常常会给用户造成挫折感。这常常致使用户连续不断地修改查询，以便把检索到的搜索结果限定到适当数量的文件。

- 15 作为此左右为难窘境的一个示例，在搜索系统输入查询字段中键入字或短语而检索出数以千计的文件—在因特网情形中则可能为数百万个网站—作为可能的候选文件也不是什么不同寻常之事。为了搞清检索到的大量候选文件，用户常常会用其它字的组合进行试验以进一步减少清单长度，因为尽管许多检索结果可能共享相同元素、词语或短语但在主题上却几乎没有或根本没有上下文相似性。这种方法
20 对执行搜索的用户和系统而言都是不准确和浪费时间的。不准确性体现在检索出成千上万的（如果不是数以百万的）用户不感兴趣的无关文件/网站。在大规模数据库中搜索可能的却无关的文件时，被浪费的还有时间及系统处理资源。

- 在信息检索（IR）领域中有一个共识：确保显示给用户的文档是根据相关性排序的，最相关的排列在最前面。在涉及搜索大量文档的一些应用中，诸如在一家
25 公司的全公司域（corporate domain）内搜索时，编辑人员检查最常用的搜索词，并选择未来响应于这些查询词应当被显示的文档。例如，编辑人员可选择对用户常见问题的解答。可以理解，对数百或数千词语做人工处理是浪费时间、低效的。

发明内容

- 30 为了提供对本发明一些方面的基本理解，以下给出了本发明的简单内容。本

部分并非是本发明的广泛总览。它并非旨在确定本发明的关键/重要元素或描绘本发明的范围。其唯一目的是以简化形式揭示本发明的一些概念，作为后面的更详细说明的序幕。

本发明涉及通过一种或多种学习方法促进信息检索和管理的系统和方法，该学习方法根据正面和负面的测试数据进行培训，以便诸如根据文档或提示其它有用信息网站的链接来确定条目的相关性。在一个方面，本发明采用了一个人工选定的文档集或条目集（称为“最佳方案（best bets）”）来培训机器学习文本分类器。该分类器可处理人工选定或机器选定的最佳方案（即正面案例）和由常规统计搜索选定的其它文档（即负面案例），来建立标识最佳方案文档的模型。在训练之后，分类器（或多个分类器）和模型可被应用到新的查询词上，以标识最佳方案文档。这包括在各种培训迭代（iteration）中对新模型进行自举（bootstrapping），以帮助建立可被用来作更准确信息检索的学习表达（learned expression）的成长型模型。

在一个示例方面，分类器可在若干应用之一中使用。这包括例如在离线场景中帮助编辑人员选择最佳方案的工具。在在线示例中，可在常规统计搜索的输出之上运行一过滤器，其中“最佳方案”被放置、显示、或排列在被确定具较低重要性的非最佳方案或条目之前。可选地，所有文档可根据它们成为最佳方案的几率进行排列。这些场景和其它因素的任何一种的效果是增加排列在最前的文档的精确度（即相关性）。

本发明可根据有助于信息管理和检索的各种不同方法来实现。例如，搜索分析员或管理员可以用分类工具进行文本搜索，然后接收最佳方案或相关性排列的建议，并选择相关联最佳方案的网站或链接。在在线情形中，搜索网站的最终终端用户可被提供以最佳或经排列的结果列表。因而，用查询无关（query-independent）的分类来确定在给定上下文中的最佳/最相关网站。

关于自举过程，在起始时向系统提供人工选定的最佳方案。通过使用在每次培训迭代中从新模型导出的机器建议并经人工校验的最佳方案，培训数据随着每次迭代而增加。最佳方案可由编辑人员仔细检查所有相应内容而以人工方式选定，或者可由系统通过例如观察用户往往选择的是哪些文档而推断哪些条目可能是最佳的。在后一情形中，本发明提供了最佳的可能性来显示用户可能会认为对其有足够的兴趣以至于希望查阅或检索这些文档或条目类型。

为了达到前述及相关目标，本发明的某些说明性方面在此结合以下说明及附图来描述。但这些方面仅仅指示了本发明可在其中实践的各种方式，且本发明旨在

包括所有这些方式。结合附图参阅本发明的以下具体说明，本发明的其它优点和新颖特征会变得清楚。

附图说明

- 5 图 1 是根据本发明一个方面的查询无关模型和系统的示意块图。
图 2 是根据本发明一个方面的示例文档排序过程的示图。
图 3 是根据本发明一方面的示例建模过程的流程图。
图 4 是根据本发明一方面的示例培训和测试过程的示图。
图 5 是根据本发明一方面的用于网站排序的示例性用户界面的示图。
10 图 6 是根据本发明一方面的用于在线排序的示例性用户界面的示图。
图 7 是根据本发明一方面的示例应用的示图。
图 8 是根据本发明一个方面的适当操作环境的示意块图。
图 9 是本发明可与之交互的采样计算环境的示意块图。

15 具体实施方式

本发明涉及应用自动学习过程用来确定文档相关性并帮助信息检索活动的系统和方法。在一方面中，提供了一种促进机器学习方法以确定文档相关性的系统。该系统包括接收用作正面测试案例的人工选定条目集和/或机器选定条目集的存储组件。用人工选定条目（或机器选定条目）作正面测试案例并用一个或多个其它条
20 目作负面测试案例，培训组件培训至少一个分类器以便提供查询独立模型，其中其它条目可通过统计搜索来选定。还有，受训分类器可用于协助人们来选择新的正面案例，或用来过滤基于统计的搜索。如果指示相关性或重要性的期望或显式标记能加到返回的结果上，过滤器的输出就可排列成使得正面案例被排列在负面案例之前。如果需要，输出也可根据其作为正面案例的几率进行排序。

- 25 当在本申请中使用时，术语“组件”、“分类器”、“模型”、“系统”等等都意指计算机相关实体，即硬件、硬件和软件的组合、软件、或执行中的软件。例如，组件可以是但不限于是：运行于处理器上的进程、处理器、对象、可执行程序、执行线程、程序和/或计算机。作为说明，运行于服务器上的应用和服务器都是组件。一个或多个组件可驻留于执行的进程和/或线程中，且组件可位于一台计算机上和/
30 或分布在两台或多台计算机之间。还有，这些组件可从存储有各种数据结构的各种计算机可读介质执行。组件可经由本地和/或远程的进程进行通信，比如根据有一

个或多个数据分组的信号进行通信（例如，来自一个与本地系统中的、分布式系统中的另一组件交互的组件、和/或一个通过该信号跨越诸如因特网的网络与其它系统交互的组件的数据）。

先参阅图 1，根据本发明的一方面示出了与查询无关的模型和系统 100。系统 100 包括分析存储在数据存储器 120 中的数据以提供一查询无关模型的一个或多个分类器 110。这些数据可包括用户感兴趣的相关网站、文档、其它数据项等等的记录（随着时间的流逝而聚集的数据）。数据存储器 120 可包括用于集中存储（例如公司.com）的信息或来自分散源的信息，诸如各种网站、文档收藏、百科全书等等。分类器 110 被用来自动分析存储器中的数据以便帮助与用户界面 140 交互的一个或多个工具。

在一示例中，工具 130 可包括帮助管理员选择用户在搜索信息时选择向他们显示的相关网站的编辑工具。这些网站可包括以最可能上下文或重要性（即最佳方案）向用户显示的排序列表，其中可能性可从分类器 110 确定。另一工具 130 可包括帮助在线用户或终端用户基于分类器 110 所确定的信息排序或分类来确定条目重要性的排序工具。

一般而言，分类器根据正面和负面测试数据进行培训，以便诸如从文档或建议其它有用信息网站的链接来确定条目的相关性。在一方面，这包括一组人工选定的文档或条目（称为“最佳方案”）用来培训机器信息文本分类器 110（或多个分类器）。分类器 110 可处理存储在数据存储器 120 中的人工或机器选定最佳方案（即正面案例）以及由常规的统计搜索选定的其它文档（即负面案例）。在培训之后，分类器 110 能应用于新查询词，以标识最佳方案或相关文档。这包括在各种培训迭代中自举新模型，以推动建立可用来进行更准确的信息检索的学习表达的成长模型，详述如下。

在一示例中，分类器 110 可在若干应用之一中利用。例如，这包括帮助编辑人员或机器在离线场景中选择最佳方案的工具 130。在在线示例或用户查询示例中，过滤器（未示出）可运行于常规统计搜索的输出之上，其中“最佳方案”置于、显示、或排序在确定为较不重要的非最佳方案或条目之前。可选地，所有文档或者一个文档子集可根据它们成为最佳方案的几率进行排序。这些和其它场景的任何一种的效果是提高排列在最前面的文档的精确度（即相关性）。

如果需要，用户界面 140 也可用来更新、改变或更改数据存储器 120，并分析、处理和接收分类的结果。界面 140 可包括具有一个或多个显示对象（未示出）的显

示(未示出)以有助于对系统 100 的操作,其中显示对象可包括有多种可配置尺寸、形状、色彩、文本、数据和声音的诸如可配置图标、按钮、滑块、输入框、选项、菜单、制表键(tab)等等。此外,可提供包括多个其它输入或控制的一个或多个用户输入(未示出)用来调节和配置本发明的一个或多个方面。这可包括从鼠标、
5 键盘、语音输入、网站、浏览器、远程 web 服务和/或其它诸如话筒、照相机或视频输入的装置接收用户命令来影响或改变系统 100 的操作。

分类器 110 可根据多种技术实现。根据本发明的一方面,所熟知的支撑向量机器(Support Vector Machines, SVM)被用作分类器。可以理解其它分类器方法也可采用,诸如朴素贝叶斯(Naïve Bayes)方法贝叶斯网络(Bayes Net)方法、决策树方
10 法、基于相似性的方法、基于向量的方法、隐藏马尔可夫(Markov)模型方法、和/或其它学习方法。SVM 方法通过学习或训练阶段进行配置。分类器是把输入属性向量 $x = (x_1, x_2, x_3, x_4, x_n)$ 映射到“输入属于一个类”的置信的函数—即 $f(x) = \text{confidence}(\text{class})$ (置信(类))。在主题分类情形中,属性是查询中的字或其它源自查询中字的域专用属性(例如词性、关键词语的存在),而类是感兴趣的类别或领域。SVM 和其它学习方法的一个重要方面是采用一个经标注的实例的培训
15 集以自动地学习分类函数。

培训集可包括指示用来询问特定主题的可能和/或真实的元素或元素组合(例如如字或短语)的词 1 到 N 的子集。每个词可与一个或多个主题相关联(例如(Q1,T2,T3,T9)、(Q7,T2,T6)、(Q2,T5))。在学习期间,学习了把输入特征映射到
20 类的置信的函数。因而,在学习了模型之后,主题分别被表示为输入特征的加权向量。注意,词、查询和/或主题的其它实现是可能的。例如,可采用另一种归纳法来不但培训主题和副主题的查询,而且培训关联于目标内容和/或文档的原始文本。换言之,系统可用几个查询做种子,但却向其提供多个原始文本,而且在后来还可添加查询和原始文本来增强该系统。

25 对于主题分类,常采用二元特征值(例如,字在主题中出现或未曾出现)、或实值特征(例如带有重要性权重 r 的字出现)。由于主题集合会包含大量唯一的词语,当把机器学习技术应用到主题分类时一般采用特征选择。为了减少特征数量,特征可基于频率总计数被移除,并基于对类别的拟合根据少量特征进行选择。对类别的拟合可通过交互信息、信息增益(information gain)、卡方(Chi-square)检验
30 法和/或任何其它统计选择技术来确定。这些较小的描述则可作为 SVM 的输入。注意,线性 SVM 提供了适当的归纳准确率并提供快速学习。其它类非线性 SVM 包

括多项式分类器和径向基函数，并也可利用在本发明中。

对于包括 SVM 的多种学习方法而言，每个类别的模型都可表示为属性权重的向量 w (例如 $w_1, w_2, \dots, w_v$)。因而，每一类别都有带权重的学习向量。当得悉权重后，通过计算 x 和 w 的点积对新查询进行分类，其中 w 是相应类学习权重的向量，而 x 是表示新查询的向量。S 型函数 (sigmoid function) 也可被提供用来把 SVM 的输出变换为几率。几率提供跨类别或类的可比得分。

SVM 是参数化函数，其函数形式在培训前定义。培训 SVM 一般需要一个经标记的培训集，因为 SVM 将从示例集拟合 (fit) 函数。培训集包括 E 个示例， E 为整数。每个示例包括输入向量 x 、以及类别标记 y ，其描述输入向量是否在类别中。对于每个类别，在用 E 个示例培训的 SVM 中有 E 个自由参数。为了得出这些参数，要求解二次规划问题 (QP)，这是众所周知的。有多种众所周知解决 QP 问题的技术。这些技术可包括顺序最小优化技术 (Sequential Minimal Optimization)，以及诸如多级组块 (chunking) 的其它技术。

现在参阅图 2，示例文档重新排列过程 200 根据本发明的一方面被示出。这可包括信息的后处理，用来确定文档或网站对用户或管理员的相关性。在该方面中，诸如条目、文档、记录、词语等的数据在 210 接收并根据如上所述的分类技术在 220 自动排序。在 230，成为相关条目的几率高的条目在具较低排序的条目之前被呈现给用户 (240)。这种排序可根据预定几率阈值获取，其中成为相关条目的几率较高的条目呈现在较低几率的条目之前。然而，可以理解，与根据几率的隐式排序不同，可在所显示的条目上加显式标记用来指示文档或网站的相关性或重要性 (例如采用指示文档相关性的色彩、数字、符号)。

图 3 和 4 是根据本发明的一方面示出建模和培训过程的流程图。为便于解释，方法论被显示和描述成一系列动作。可以理解和明白的是，本发明并不限于所述动作和/或动作之顺序，例如动作能以各种顺序和/或同时发生，并与未在此显示和描述的其它动作一起发生。例如，本领域技术人员将理解和明白，这些方法论可另外通过状态图表示为一系列相互关连的状态或事件。此外，要实现根据本发明的方法论，并不是所有的图示动作都是必需的。

参阅图 3，示例建模过程 300 根据本发明的一方面示出。在 310，至少一个被用来提取或确定最佳方案文档或网站本质的文本分类器被构建。在 320，要分析的测试数据存储被选中。例如，这可包括有各种主题或词语的适当域，或包括分布在各个本地或远程数据库的多个域。在 330，一个或多个类别从区域内选中 (例如，

团体、下载、合伙人、产品信息、支持、技术、培训、公司信息等等)。在 340, 至少一个分类器被分别分配给类别以便学习类别中最佳方案的特征。在 350, 已知培训数据集被用来培训和测试类别分类器, 在下面参照图 4 更详细地描述。

图 4 根据本发明的一方面示出了示例培训和测试过程 400。进行到 410, 数据库中的词语 (term) 被拆分为培训数据集和测试数据集 (例如, 80%的词语用来培训, 20%的词语用来测试分类器)。在 420, 每类的和每个词语的最佳结果文档被提取出来 (例如, 提取最佳的 10 个文档)。在 430, 相应类别分类器在 420 的提取结果之上进行培训 (例如, 通过采用字特征)。在 440, 以上所述的培训数据由受训分类器分析, 以确定分类器的准确率。例如, 这包括由分类器确定的类别准确性与基线估计之间的比较, 如下表所示:

类别 (#特征)	类别准确性	类别基线
团体(1K)	97.81	88.86
下载(1K)	92.06	76.02
合伙人(1K)	96.86	91.11
产品信息(1K)	89.25	75.22
支持(1K)	96.09	79.28
技术(1K)	93.36	86.33
培训(1K)	89.8	88.78
公司信息(10K)	96.93	92.40

图 5 根据本发明的一方面示出了用于网站排序的示例用户界面 500。在该方面中, 示例搜索词语(例如, blaster worm (胚胎虫))在 510 根据本发明被提交给适合的数据网站。该网站被拆分成各种类别 520 和 530。在每个类别中的结果 540、541 等可包括到提供与所需词语相关的更多信息的网站的最佳方案链接 (best bet link)。

图 6 根据本发明的一个方面描绘了在线排序的示例性用户界面 600。在该方面中, 界面 600 根据用户提交的查询词语 620 排序类别 610 中的文档。如上所述, 这可包括界面 600 上的隐式排序, 其中相关性强的文档呈现在其它文档之前。

图 7 根据本方面的一方面描述了示例应用 700。应用 700 包括一个或多个可能使用场景, 尽管可以理解的是还可采用其它应用。在 710, 一个案例应用到查询和与这些查询相关的文档, 其中的文档是先前根据最佳方案分析时没有进行过处理的。在 720, 离线处理可包括处理来自记录的最前面的 n 个查询 (n 为整数)、获取这些查询的搜索结果、从结果中标识最佳候选结果、以及将该分析转送给确定一则

信息对于给定查询或主题是否值得显示的编辑人员。在 730, 在线过程可包括从可能文档或网站列表中提取最佳方案, 并自动把最佳方案置于其它诸如统计处理的技术所处理的可能条目之前。在 740, 在线技术还可包括按条目成为最佳方案的几率来重新排序结果。

5 参照图 8, 实现本发明各方面的示例性环境 810 具有计算机 812。计算机 812 具有处理单元 814、系统存储器 816、及系统总线 818。系统总线 818 耦合系统组件包括, 但不限于将系统存储器 816 耦合到处理单元 814。处理单元 814 可以是各种可用处理器的任一种。双微处理器和其它多处理器架构也可被用作处理单元 814。

10 系统总线 818 可以是若干类总线结构的任一种, 包括存储器总线或存储器控制器、外围总线或外部总线、和/或使用各种可用总线架构任一种的本地总线, 这些总线架构包括, 但不限于, 16 位总线、工业标准架构 (ISA)、微信道架构 (MSA)、扩展 ISA (EISA)、智能磁盘设备 (IDE)、VESA 局部总线 (VLB)、外围部件互连 (PCI)、通用串行总线 (USB)、加速图形接口 (AGP)、个人计算机存储卡国际
15 协会总线 (PCMCIA)、和小型计算机系统接口 (SCSI)。

系统存储器 816 具有易失存储器 820 和非易失存储器 822。包含在计算机 812 硬件间传送如起动时信息的基本例程的基本输入/输出系统 (BIOS), 存储在非易失存储器 822 上。作为说明且无限制, 非易失存储器 822 可包括只读存储器 (ROM)、可编程 ROM (PROM)、电可编程 ROM (EPROM)、电可擦除 ROM (EEPROM)
20 或闪存。易失存储器 820 具有作为外部高速缓存的随机存取存储器 (RAM)。作为说明且无限制, RAM 有多种形式可用, 诸如同步 RAM (SRAM)、动态 RAM (DRAM)、同步 DRAM (SRAM)、双数率 SDRAM (DDR SDRAM)、增强 DSRAM (ESDRAM)、同步链接 DRAM (SLDRAM)、和直接 Rambus RAM (DRRAM)。

计算机 812 还具有可移动/不可移动、易失/非易失计算机存储介质。图 8 示出了例如磁盘存储器 824。磁盘存储器包括, 但不限于, 象磁盘、软盘、磁带、Jaz
25 盘、Zip 盘、LS-100 盘、闪存卡、或存储棒。另外, 磁盘存储器 824 可具有单独存储介质或结合其它存储介质包括, 但不限于, 诸如光盘 ROM 设备 (CD-ROM)、可记录光盘驱动器 (CD-R 盘)、可重写光盘驱动器 (CD-RW 盘)、或数字化视频 ROM 盘 (DVD-ROM) 的光盘。为便于磁盘存储器 824 与系统总线 818 的连接,
30 通常可使用诸如接口 826 的可移动或不可移动接口。

应当理解, 图 8 描述了作为用户和在适当操作环境 800 中描述的基本计算机

资源间中间体的软件。这种软件具有操作系统 828。可存储在磁盘存储器 824 中的操作系统 828，其作用是控制和分配计算机系统 812 的资源。系统应用 830 得益于操作系统 828 通过存储在或系统存储器 816 或磁盘存储器 824 中的程序模块 832 和程序数据 834 对资源的管理。可以理解本发明可用各种操作系统或操作系统组合来实现。

用户通过输入设备 836 将命令或信息输入到计算机 812 中。输入设备 836 包括，但不限于，诸如鼠标、跟踪球、铁笔、触板、键盘、麦克风、操纵杆、游戏垫、卫星接收器、扫描仪、电视调谐器卡、数字相机、数字录像照相机、网络相机等定点设备。这些和其它输入设备经系统总线 818 通过接口端口 838 与处理单元 814 相连。接口端口 838 包括例如串行端口、并行端口、游戏端口、和通用串行总线 (USB)。输出设备 840 使用象输入设备 836 的某些相同类型端口。因此，例如 USB 端口可用来为计算机 812 提供输入，并将信息从计算机 812 输出到输出设备 840。输出适配器 842 被提供用来说明某些输出设备象监视器、扬声器、和打印机，与其它输出设备 840 之间，需要特殊适配器。输出适配器 842 包括，作为说明且无限制，视频卡和声卡在输出设备 840 和相同总线 818 间提供了连接方法。应该注意的是诸如远程计算机 844 的其它设备和/或设备系统提供输入和输出功能。

计算机 812 使用与一台或多台远程计算机如远程计算机 844 的逻辑连接在网络化环境中操作。远程计算机 844 可以是个人计算机、服务器、路由器、网络 PC、工作站、基于微处理器设备、同等设备或其它相同网络节点等等，而且通常包括所述与计算机 812 相关的许多或全部元件。为简化起见，仅存储器存储装置 846 与远程计算机 844 一起图示。远程计算机 844 经网络接口 848 与计算机 812 逻辑连接，然后通过网络连接 850 进行物理连接。网络接口 848 包括诸如局域网 (LAN) 和广域网 (WAN) 的有线和/或无线通信网络。LAN 技术包括光纤分布式数据接口 (FDDI)、铜线分布式数据接口 (CDDI)、以太网/IEEE1102.3、令牌环/IEEE1102.5 等等。WAN 技术包括，但不限于，点对点链接、象综合服务数字网 (ISDN) 及其后变种的电路切换网络、报文分组交换网络、和数字用户线路 (DSL)。

通信连接 850 指向被用来将网络接口 848 连接到总线 818 的硬件/软件。尽管为了清晰地说明，通信连接 850 被示为位于计算机 812 内，但它也可在计算机 812 之外。与网络接口 848 连接所必需的硬件/软件具有，仅作为示例，诸如包括常用电话级调制解调器、电缆调制解调器和 DSL 调制解调器的调制解调器、ISDN 适配器、和以太网卡等内部和外部技术。

图9是本发明可与之交互的采样计算环境900的示意块图。系统900包括一台或多台客户机910。客户机910可以是硬件和/或软件(例如线程、过程、计算装置)。系统900还可包括一台或多台服务器930。服务器930也可以是硬件和/或软件(例如线程、过程、计算装置)。服务器930可采用本发明覆盖线程来执行变换。

5 在客户机910和服务器930间的一可能通信可能是以适于在两个或多个计算机过程间传送的数据包形式进行。系统900包括可用来便于客户机910和服务器930间通信的通信框架950。客户机910可与用来存储客户机910本地信息的一个或多个客户数据存储960作可操作连接。类似地,服务器930可与用来存储服务器930本地信息的一个或多个客户数据存储940作可操作连接。

10 以上所述包括本发明的诸多示例。当然,为描述本发明而对每一能想到的组件或方法论组合进行描述是不可能的,但本领域普通技术人员明白本发明的更多排列和组合是可能的。因此,本发明意欲包含所有这样的在所附权利要求书精神和范围内的变更、修改、和变化。此外,就用于具体实施方式或权利要求书的术语“具有”而言,这种术语意在以类似于术语“包括”在权利要求书中作连接词的方式作

15 包含意义解。

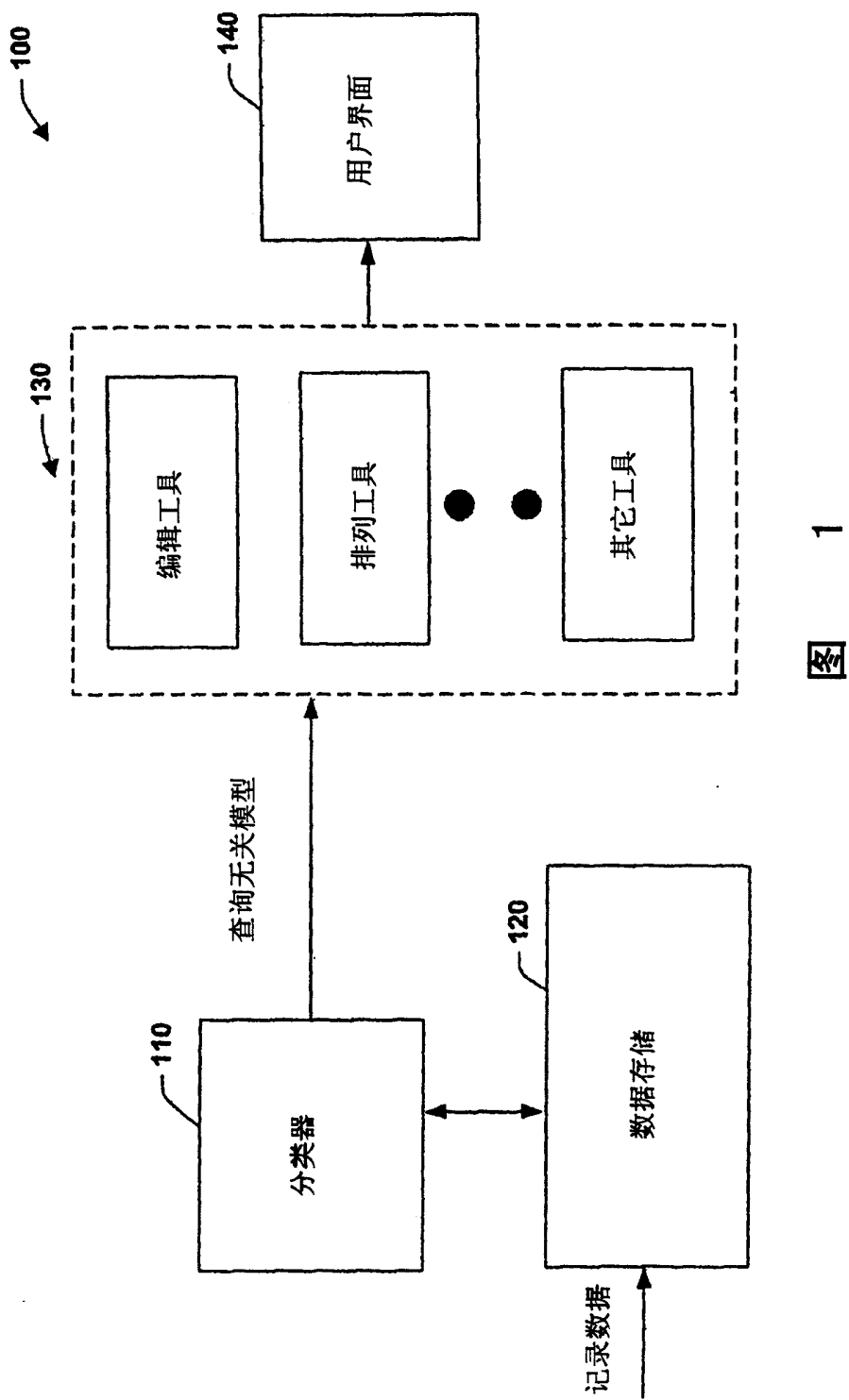


图 1

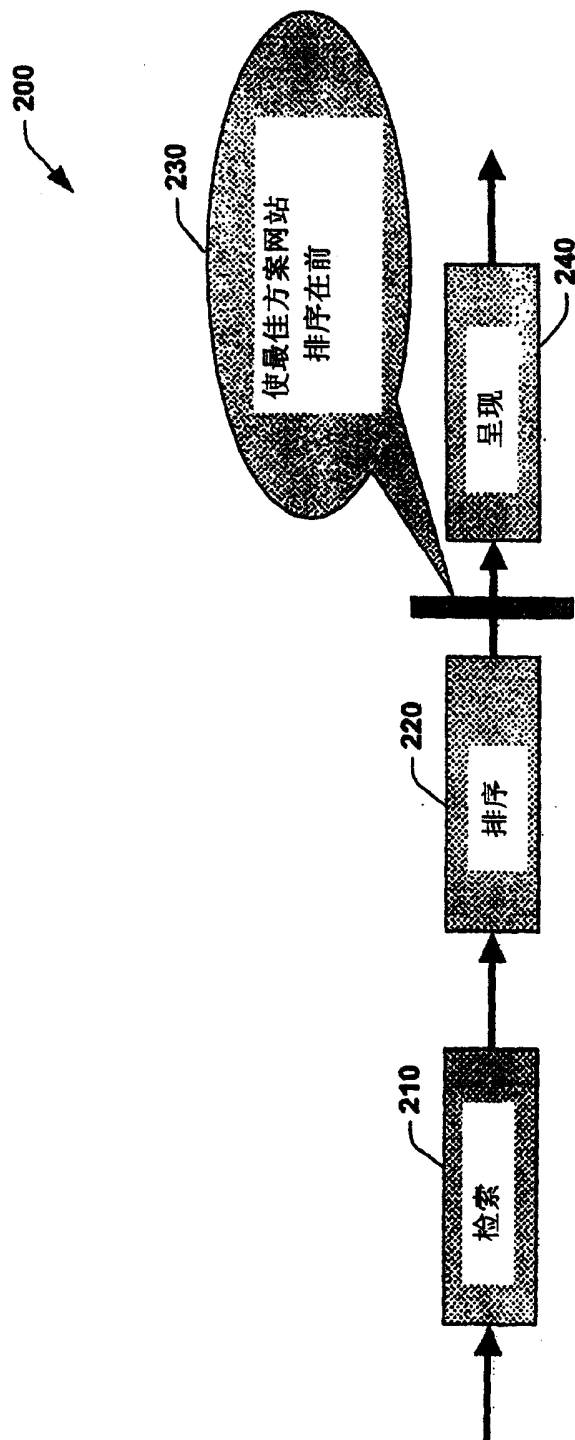


图 2

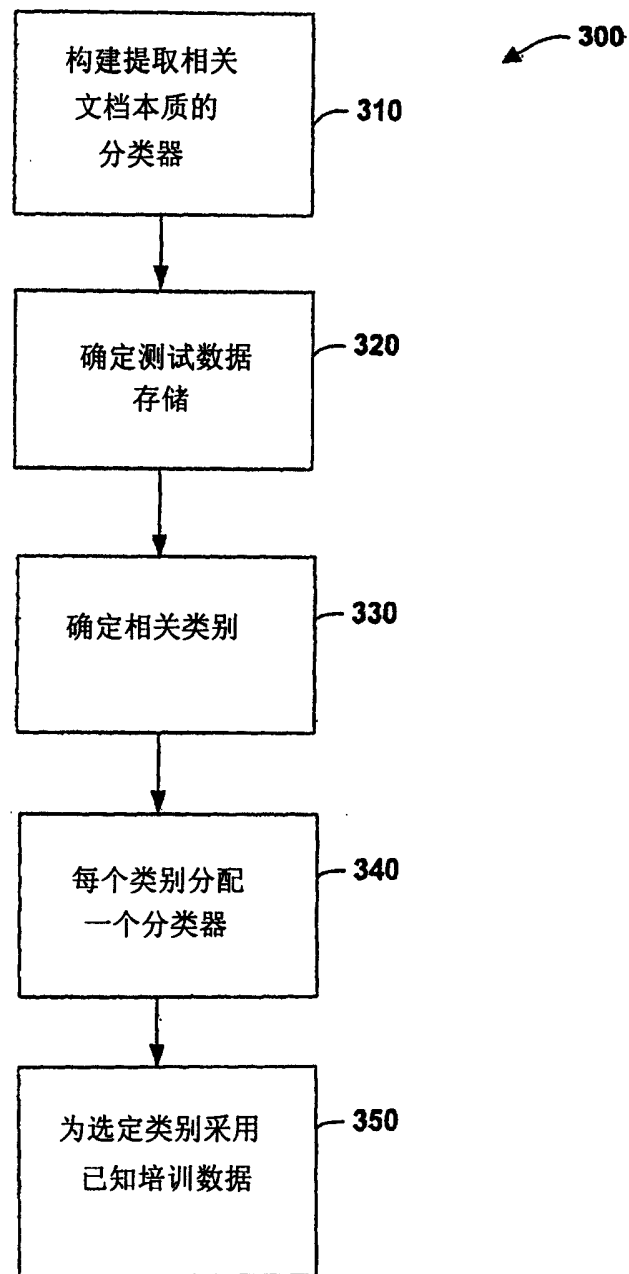


图 3

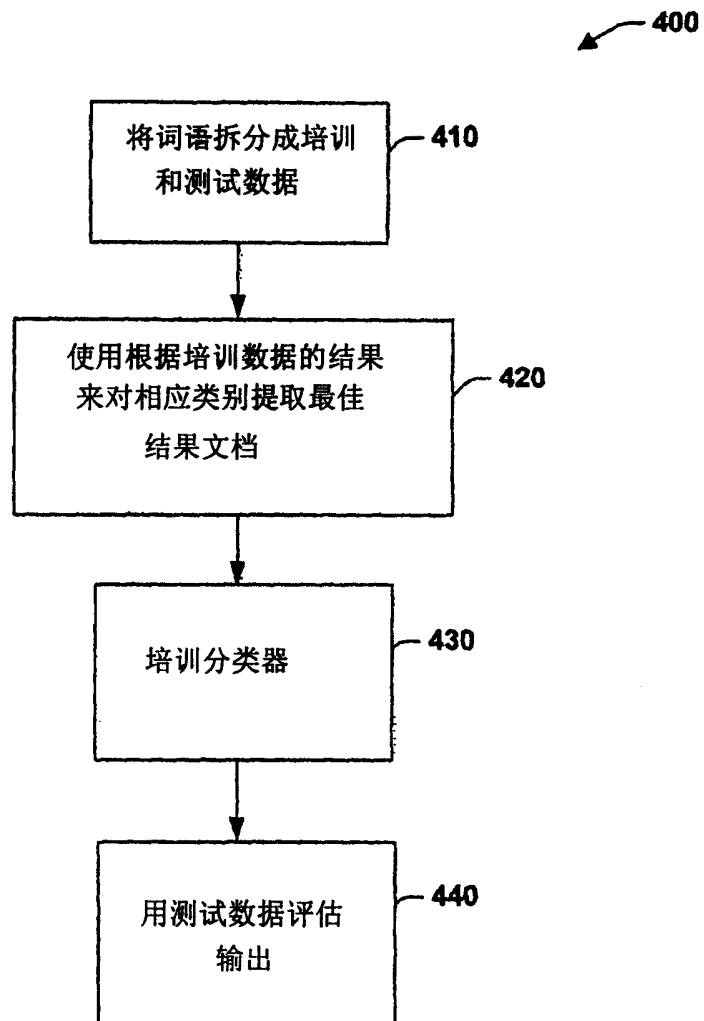


图 4

500

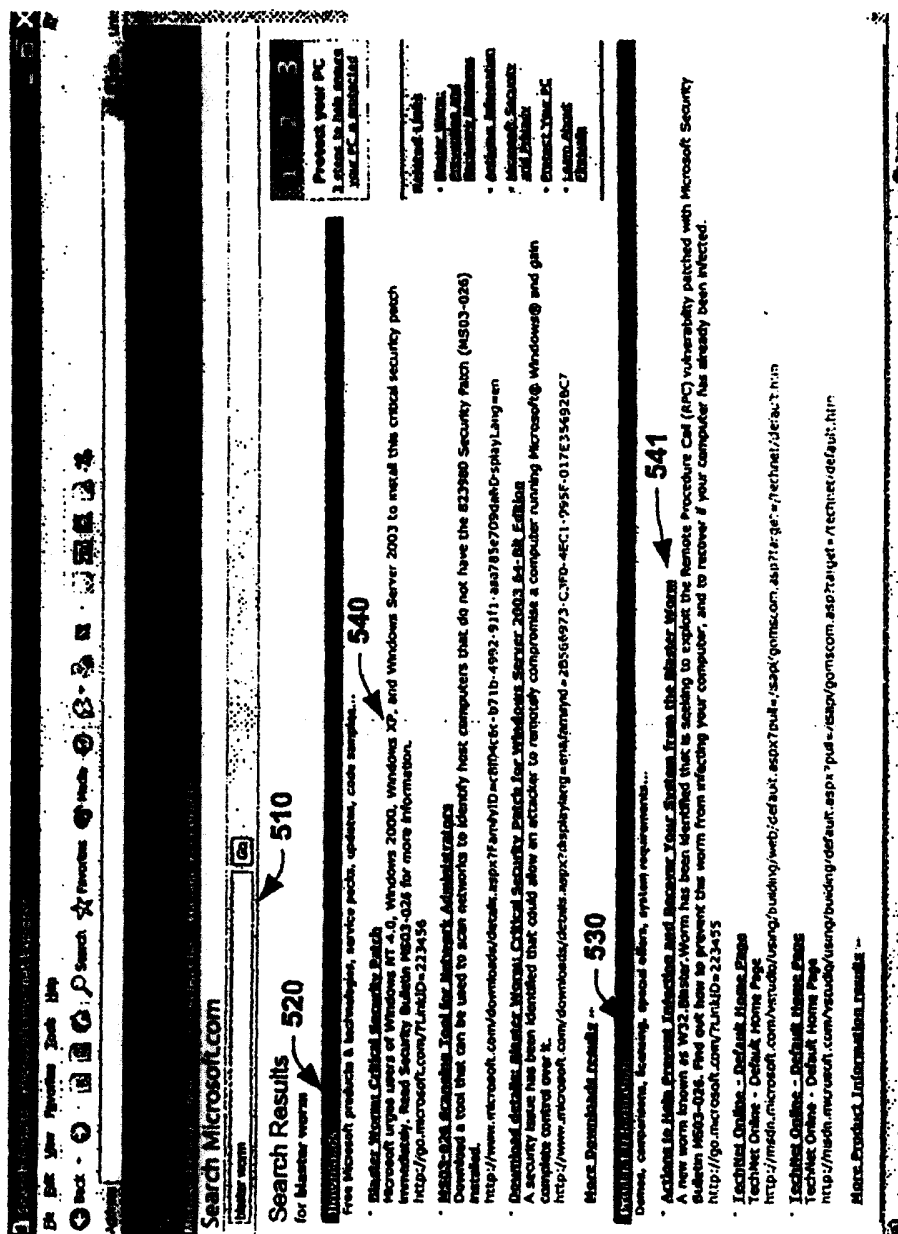
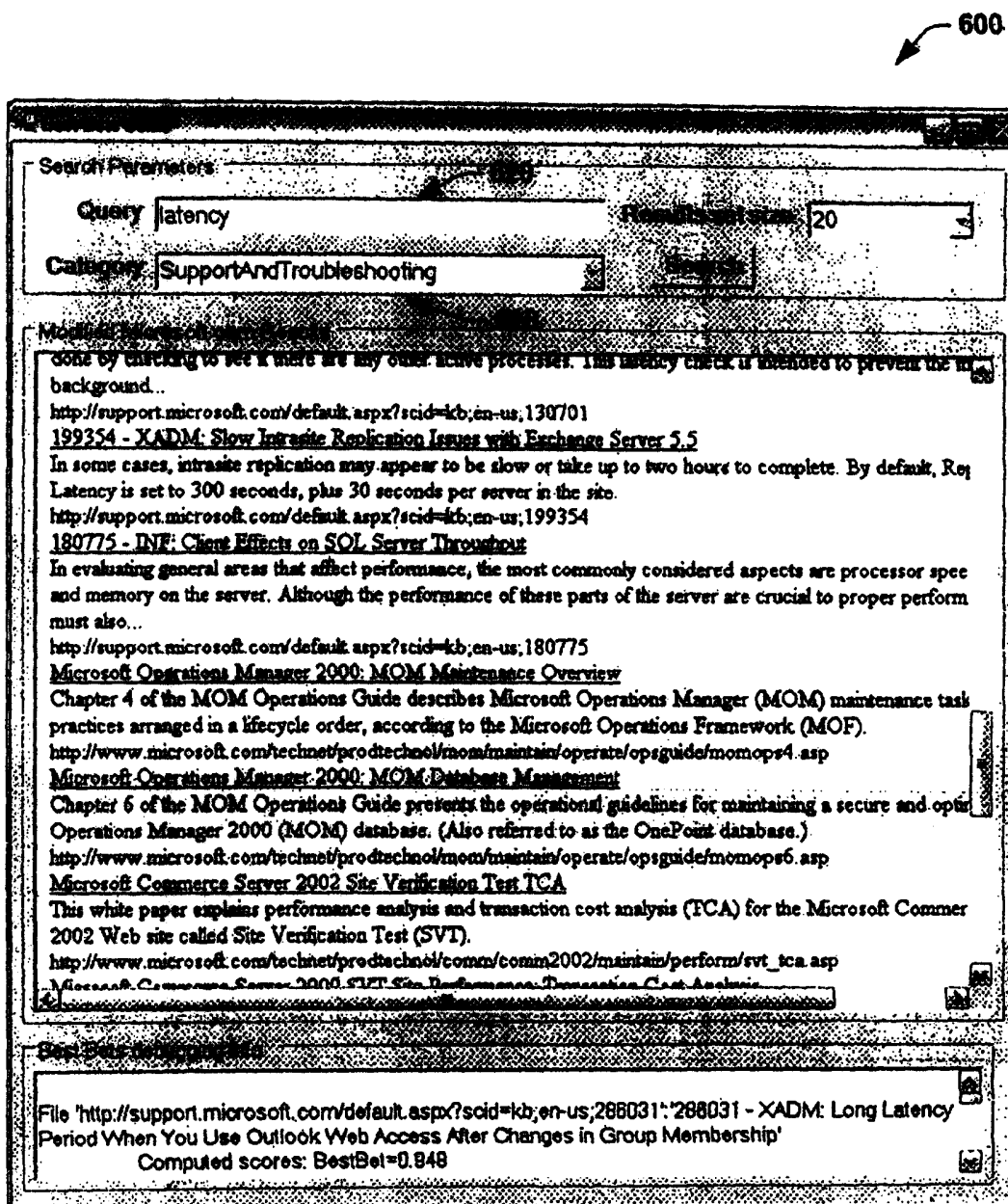


图 5



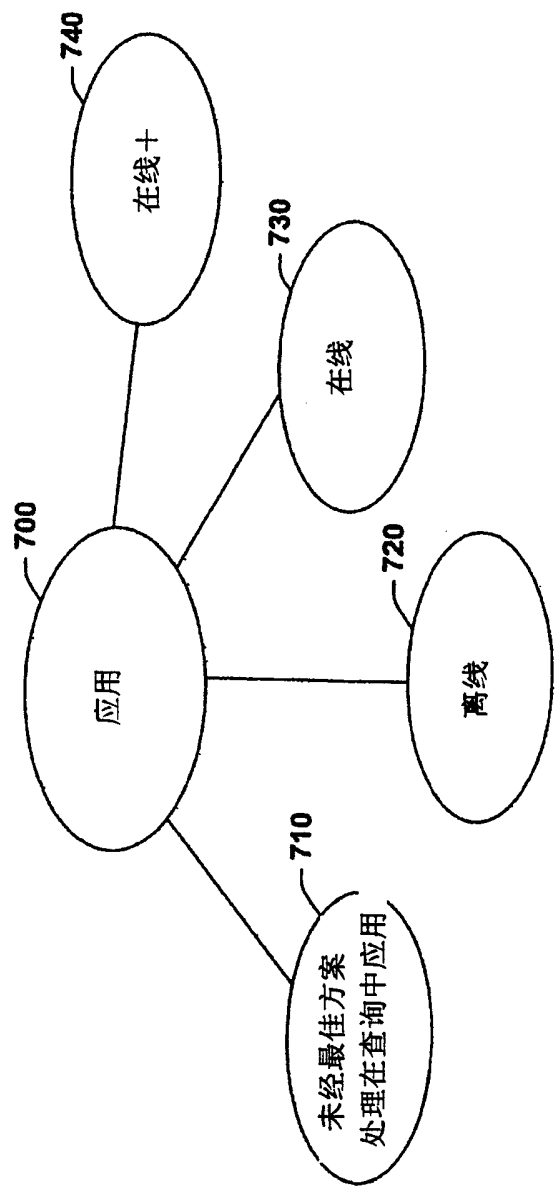


图 7

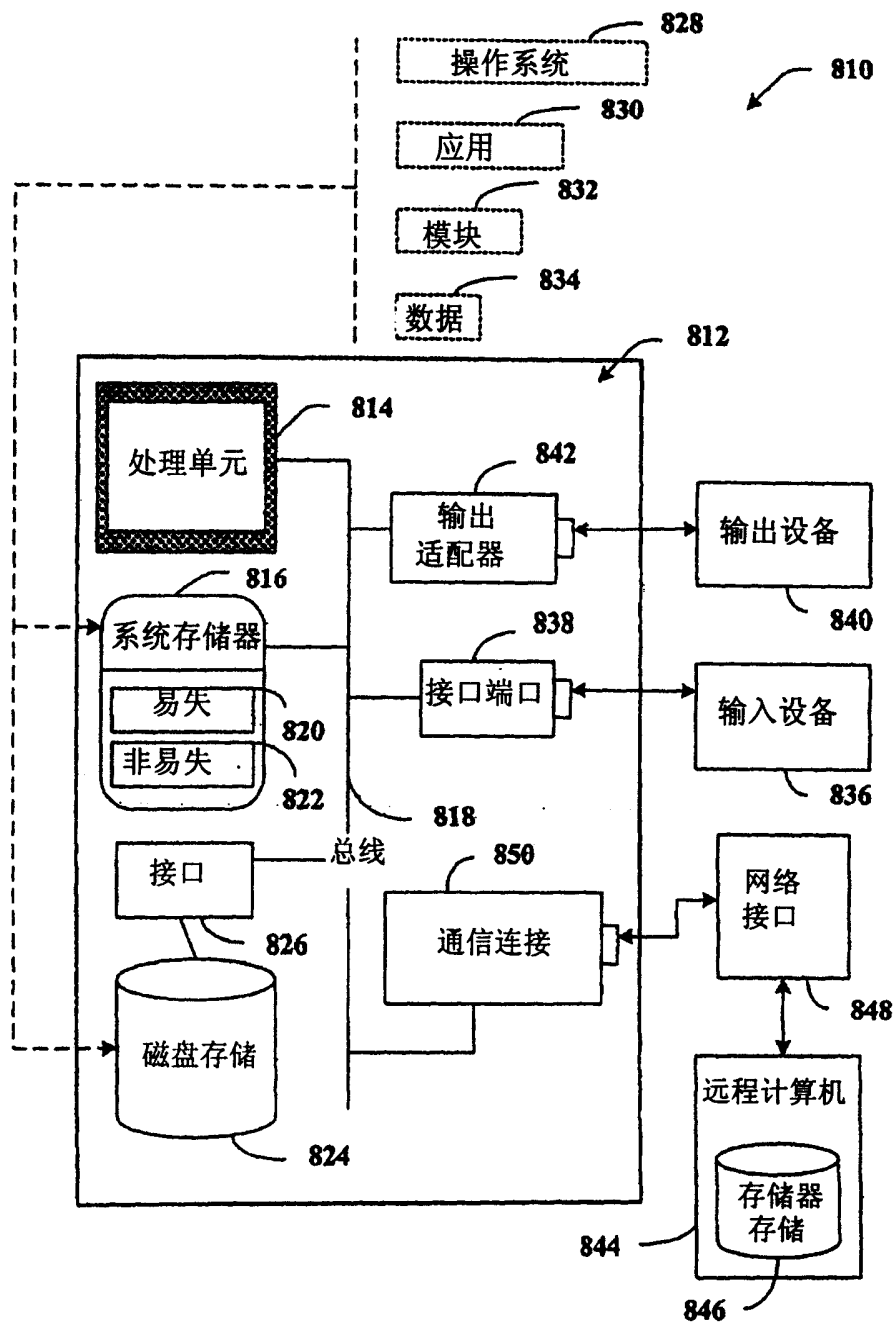


图 8

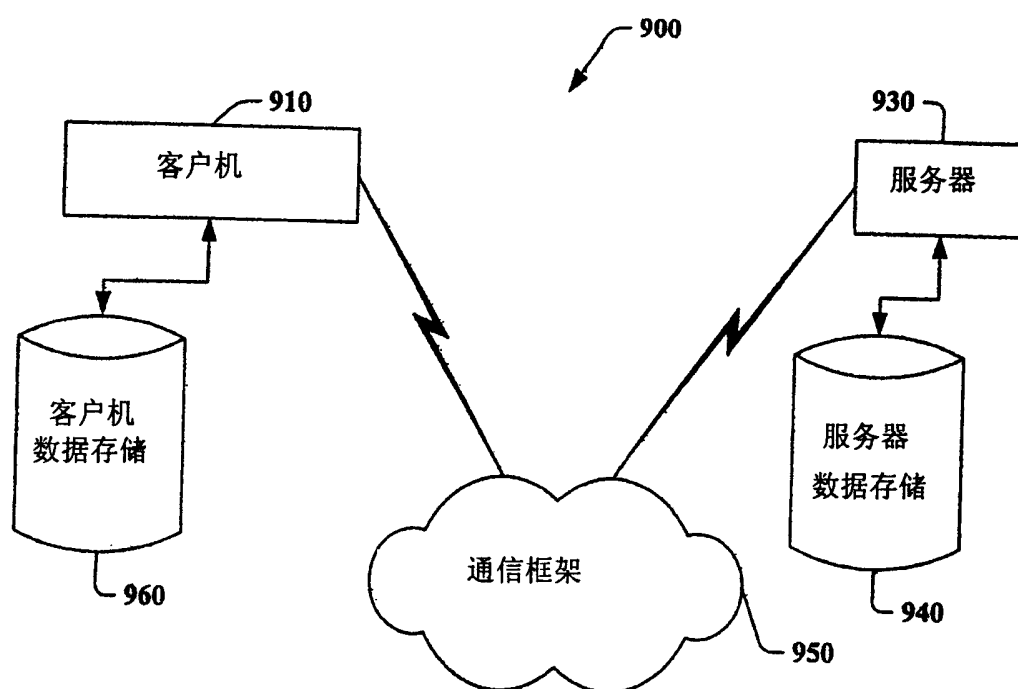


图 9