



[12] 发明专利申请公开说明书

[21] 申请号 200410102291.1

[43] 公开日 2005 年 6 月 22 日

[11] 公开号 CN 1629845A

[22] 申请日 2004. 12. 16

[21] 申请号 200410102291.1

[30] 优先权

[32] 2003. 12. 16 [33] US [31] 10/736,915

[71] 申请人 微软公司

地址 美国华盛顿州

[72] 发明人 M·卡尔布西 S·萨拉斯

[74] 专利代理机构 上海专利商标事务所有限公司

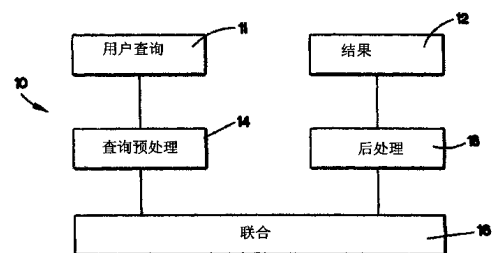
代理人 谢喜堂

权利要求书 3 页 说明书 9 页 附图 3 页

[54] 发明名称 查询识别器

[57] 摘要

一种用于在包括具有专用数据的远程定位数据源的多个数据源上执行查询的系统和方法。这些数据源可包括,例如,涉及零售、旅游、娱乐等的数据。这些数据源也可仅包含对给定的团体或团体组本地的数据。该系统包括接受查询并修改其形式以提供结果查询的预处理器。这一修改可包括例如拼写纠正。该系统基于分类符项的存在以及用户根据先前提交的查询或查询中识别的模式提供的分类数据对结果查询进行分类。该系统然后选择一数据源子集,在该子集上可基于该查询分类执行查询。该查询然后在所选择的数据源上执行。



1. 一种分析来自用户的查询并响应所述查询的系统，其特征在于，它包括：
 - a) 一查询处理器，用于评估一输入查询并向一数据源发送所述查询的经修改
 - 5 形式，以创建对所述查询的经修改形式的响应；以及
 - b) 多个识别器，用于评估所述查询或所述查询的经修改形式，并标识所述查询中的数据；
 - c) 所述查询处理器包括一识别器代理程序，用于将所述查询发送到所述多个识别器中指定的一个或多个。
- 10 2. 如权利要求 1 所述的系统，其特征在于，所述识别器代理程序将所述查询分类为具有一定程度的确定性的某一类型的查询。
3. 如权利要求 1 所述的系统，其特征在于，所述查询识别器基于在所述查询中标识的数据添加信息，以创建所述输入查询的经修改的形式。
4. 如权利要求 3 所述的系统，其特征在于，所述查询识别器向所述输入查询
- 15 添加类别。
5. 如权利要求 3 所述的系统，其特征在于，所述查询识别器添加涉及包含在所述输入查询中的信息的信息。
6. 如权利要求 1 所述的系统，其特征在于，它还包括一 web 爬行检测器，用于搜索文档并对所述文档进行分类，以及一数据库，用于基于来自所述查询处理
- 20 器的输入提供结果。
7. 如权利要求 1 所述的系统，其特征在于，所述查询处理器在向数据源发送所述经修改的查询之前修改所述查询以包括一类别。
8. 如权利要求 1 所述的系统，其特征在于，所述查询处理器改变所述查询中拼错的单词的拼写。
- 25 9. 如权利要求 1 所述的系统，其特征在于，所述识别器将查询输入单词与储存在一数据库中的单词进行匹配，并以一置信级别对所述单词进行分类。
10. 如权利要求 9 所述的系统，其特征在于，所述数据库包括用于建立所述置信级别的用户等级的历史。
11. 如权利要求 9 所述的系统，其特征在于，所述用户等级的历史基于用户
- 30 对基于先前所提交的查询所呈现的先前的结果的选择。

12. 如权利要求 10 所述的系统, 其特征在于, 所述数据库包含单词索引以及对所述单词的查询类别的指示, 其中, 对同一单词存在多个类别。

13. 如权利要求 1 所述的系统, 其特征在于, 它包括一模式识别器, 用于识别输入查询中的字符模式以对所述查询进行分类。

5 14. 如权利要求 1 所述的系统, 其特征在于, 一个或多个模式识别器包含指定语言的项的数据库, 用于在所述识别器中的标识。

15. 一种分析来自用户的查询并用数据响应所述查询的系统, 其特征在于, 它包括:

10 a) 一查询处理器, 用于评估所述查询并向一数据源发送所述查询的一种形式, 以创建对所述查询的经修改的形式的响应; 以及

 b) 一查询识别器, 用于评估所述查询或所述查询的经修改的形式, 并基于所述查询中标识的语言符号标识一查询类型。

 c) 一数据库, 它储存所述语言符号, 并关联包含所述语言符号的查询的类型的置信级别, 并在确认输入查询中的语言符号和所述数据库中的语言符号之间的匹
15 配时, 向所述查询识别器提供一置信级别。

16. 如权利要求 15 所述的系统, 其特征在于, 所述查询处理器在向所述数据源发送所述经修改的形式之前修改所述查询的形式, 所述经修改的形式包括一查询类别指示符。

17. 一种评估查询的方法, 其特征在于, 它包括:

20 a) 评估一输入查询并向一数据源发送所述查询的经修改的形式, 以通过使用多个查询识别器中指定的一个或多个评估所述查询来创建对所述查询的经修改的形式的响应, 所述查询识别器评估所述查询中的语言符号以生成所述查询的经修改的形式。

25 18. 如权利要求 17 所述的方法, 其特征在于, 所述输入查询以一定程度的确定性被标识为某一类型的查询。

19. 如权利要求 17 所述的方法, 其特征在于, 它包括基于所述输入查询中标识的数据添加信息以创建所述输入查询的经修改的形式。

20. 如权利要求 17 所述的方法, 其特征在于, 它包括向所述输入查询添加类别。

30 21. 如权利要求 17 所述的方法, 其特征在于, 它包括添加涉及包含在所述输

入查询中的信息的信息。

22. 如权利要求 17 所述的方法，其特征在于，它还包括搜索文档并对所述文档进行分类，并维护一用于基于输入提供概率的数据库，所述输入基于所述输入查询的内容从所述查询处理器导出。

5 23. 如权利要求 17 所述的方法，其特征在于，所述查询处理器改变所述输入查询中拼错的单词的拼写。

24. 如权利要求 17 所述的方法，其特征在于，所述查询识别器之一将查询输入单词与储存在一数据库中的单词进行匹配，并以一置信级别对所述单词进行分类。

10 25. 如权利要求 24 所述的方法，其特征在于，所述数据库包括用于建立所述置信级别的用户等级的历史。

26. 如权利要求 25 所述的方法，其特征在于，所述用户等级的历史基于用户对基于先前提交的输入查询所呈现的先前的结果的选择。

15 27. 如权利要求 26 所述的方法，其特征在于，所述数据库包含单词索引以及对所述单词的查询类别指示，其中，对同一单词存在多个类别。

28. 如权利要求 17 所述的方法，其特征在于，它包括识别输入查询中的字符模式以修改所述查询。

29. 如权利要求 28 所述的方法，其特征在于，一个或多个模式识别器包含指定语言的项的数据库，用于在所述识别器中的标识。

20 30. 一种具有用于在多个数据源上执行查询的指令的计算机可读媒质，其特征在于，它包括用于执行以下动作的指令：

预处理所述查询以形成一结果查询；

基于分类符项的存在以及用户根据先前提交的查询提供的分类数据对所述结果查询进行分类；以及

25 选择一数据源子集，在该子集上基于查询分类来执行所述查询。

查询识别器

5 技术领域

本发明涉及有效地向查询提供答复的自动化查询分析器。

背景技术

10 查询搜索引擎的一个目标是提供对查询的快速响应。面对较慢的响应搜索引擎的在线用户可通过试图重新提交搜索、停止搜索并去往另一搜索引擎、或可能试图重新表示搜索以寻找更快的结果来作出反应。如果可足够快地向用户返回结果，以防止用户尝试对获取结果的速度的可感知的问题的这些解决方案，则这是合乎需要的。

15 Wen 等人的名为“群聚搜索引擎的用户查询（Clustering User Queries of a Search engine）”的出版物描述了其目标为提高搜索引擎检索准确度的过程。Wen 等人的文献群聚查询，使得可向询问该查询的人呈现一预公式化的 FAQ（常见问题解答）文档。例如，如果群聚过程确定一查询正在询问“新汽车”，则返回“新汽车”FAQ 文档作为对“新汽车”查询的响应。该方法先假定用于每一查询聚类的 FAQ 文档的存在，也先假定用于提交给搜索引擎的每一查询的匹配聚类的存在。
20 网站 www.ask.com 提供了用户可用于寻找查询结果的方法，并且该网站可使用类似于 Wen 等人的文章中所描述的那些技术。

发明内容

25 如果形成查询搜索引擎的一部分的分析软件可准确地依照其类别来标识查询，则搜索引擎可更快速地响应查询。

一种示例性系统分析来自用户的查询，并用数据响应该查询。查询处理器评估查询，并向另一数据源发送一查询形式用于创建查询的修改形式的响应。该系统实现一评估查询或查询的修改形式并标识查询类型的识别器组件。在一个示例性实施例中，查询处理器包括一用于向多个识别器的指定的一个或多个发送查询的识别
30 器代理程序。

一个这样的识别器是单词或语言符号（token）匹配识别器。该系统将查询输入单词或语言符号与储存在数据库中的单词相匹配，并用一置信度级别对这些单词进行分类。置信度级别从定义用户等级的历史的数据库记录中导出，以使用先前提提交的查询。

5 下文结合附图详细描述本发明的这些和其它目的、优点和特征。

附图说明

图 1 是适合使用本发明的示例性实施例的计算机系统的示意描述；

图 2 是依照本发明的示例性实施例构造的查询识别器的框图；

10 图 3 是查询预处理器的子组件的框图；以及

图 4 是图 2 的查询识别器的流程图。

具体实施方式

实践本发明的示例性环境

15 图 2 示出了依照本发明的一个示例性实施例构造的查询分析器的示意图。查询分析器 10 在从用户接收查询 11 时开始其分析。更典型地，分析器以在诸如服务器计算机等计算机系统 20（图 1）上执行的软件来构造，该计算机系统进而通过通信连接耦合至其它服务器或通过网络耦合至其它计算机。在最典型的示例中，用户登录到他或她的计算机并通过因特网广域网与担当服务器的远程计算机系统进
20 通信。

在服务器 20 上执行的搜索引擎软件可能结合其它联合搜索引擎提供对查询的快速响应。该响应以搜索结果 12 的形式提供给用户，通常通过诸如因特网等网络发送回用户。响应可被公式化为一系列文章或具有到那些文章的链接的网站汇总，或嵌入在搜索结果中的网站。可担当合适的查询响应计算机的计算机系统 20 在图
25 1 中描述，并在下文详细描述。

示例性计算机系统 20 包括定义用于评估查询的查询处理器 10 的软件。对查询 11 的接收的一个可能的响应是向基于查询的修改形式执行搜索的另一服务器重新发送该查询的修改形式。作为示例，搜索结果的其它来源可以是作为提供关于机票价格、酒店等的数据的旅行网站的主机的服务器。它可以是维护一个国家中的教堂列表的宗教网站。它可以是专用于汽车信息的网站，它进而具有到汽车经销商
30

的链接。其它，当然是非穷尽的类别为：新闻、本地、体育、百科全书、历史、书籍、电影、娱乐等等。

图 1 描述的服务器计算机系统 20 也可基于由服务器 20 维护的数据的内容直接评估查询，并提供响应或结果 12。该数据库包含以单词索引形式的信息，单词索引由在上千个网站上搜索因特网分类页内容的 web 爬行检测器 (web crawler) 获取。这一扫描以周期性的基础作出，以确保它是该网站的内容的最新表示。无论计算机系统 20 是搜索查询结果还是向另一计算机发送搜索结果，结果 12 由服务器 20 格式化并被发送回用户。由于这一结果包含到包含文档或信息的其它网站的链接列表，因此用户能够在文档上点击，并且用户的 web 浏览器被重定向到由该链接指向的因特网地址。

为在其它位置上有效地使用搜索引擎，计算机系统 20 使用多个识别器 220 (图 3)，用于为标识输入查询的类型的目的评估查询或查询的修改形式。一旦标识了查询的类型，分析器联合程序 16 决定可向哪一备选站点搜索引擎发送查询，或可选地决定查询应当由最初接收该查询的计算机系统 20 作进一步的评估。

查询处理

服务器 20 包括在输入查询上执行若干功能的查询处理器组件 14。图 4 是描述查询处理器 10 的操作的流程图。查询处理器接收 110 输入查询，并访问 120 关于特定用户的查询上下文信息，如查询的地理和互联网 (网页) 的起点、该用户最近访问的网站以及最近由用户输入的查询和由用户选择的对这些查询的结果。查询上下文信息的一部分包含在嵌入到查询所起源的源计算机的地址中的信息内。该地址是被分解成字段的 32 比特的串。由 IEFT 公布的 RFC #791 第 3.2 节描述了 IP 寻址系统的细节。

每一因特网服务提供商 (或者国家或者公司) 得到 IP 范围类 A、B 或 C，并按其自己的需要划分对其可用的 32 字节。在大多数实例中，由于公司因特网连接位置的存在，可将 IP 关联到城市。这一反向查找并非总是准确的，例如，所有的 AOL 用户具有在维吉尼亚起源的 IP 地址。

接下来，查询处理器在查询上执行若干功能来修改或扩充查询，以优化查询分析。这一扩充的目的是快速返回最可能与该特定用户相关的结果。

在阶段 130，查询处理器在查询上执行拼写检查，并改变查询中拼错的项的拼

写，或向查询扩充正确拼写的项。查询处理器扫描经纠正拼写的查询以找出应当被组合成短语的项 135。查询处理器可使用关于通常被执行的查询的信息来确定哪些项应当被组合成短语。

- 5 在步骤 140，查询处理器标识或识别短语中担当表示查询为某一类型的指示符的单词，类型如位置敏感的本地查询，或搜索要购买的项目的查询。这些单词或项的标识可促使查询处理器向查询扩充上下文专用信息，如基于查询所起源的地理起点的邮政编码和区域代码信息。

- 10 在这一阶段，查询的每一短语由查询分析器 200 和识别器代理程序 210 断开、提取词干并分析，用于概念或类别匹配。这些严格基于内容的概念结合对特定用户收集的过去数据一起标识可能的联合结果，即，以代理方式将该查询安排在何处用于最有效的分析。联合被定义为将查询“切断”成单独的服务（内部或外部）以提供与查询有关的数据，用于产生查询结果。在识别阶段，多个查询识别器 221、222、223、224 等评估该查询，并为识别器代理程序 210 确定查询属于预定类别组之一的概率。

- 15 在查询预处理的分析级采用了三个单独的模块或组件。单词断开器将查询的每一短语分割成单独的单词，并将这些单词储存在输出阵列或列表中。词干提取（stemmer）器组件试图从单词断开器的输出阵列中找出每一单词的根，并创建对应的根单词阵列。最后，识别器组件试图对照储存在数据库 230 中的意向列表匹配根单词（或对没有根的单词匹配实际单词），以发现单词的意向。识别器组件也使用算法查询意向识别器来搜索模式。这一分析的结果提供了类别以及表示为百分比的置信度。考虑由用户输入的“compare price Buick and Satturn（比较 Buick 和 Satturn 的价格）”形式的查询。
- 20

下表 1 是识别器 221 在该查询上的分析的结果列表。

表 1

根或单词	类别	置信度
Satturn 或 Saturn	汽车	68%
Satturn 或 Saturn	科学	46%
Satturn 或 Saturn	神话	14%
Buick	汽车	91%
Price	购物	99%

Compare	购物	50%
---------	----	-----

在阶段 150 (图 4)，基于修改的查询并鉴于过去的查询确定用户的可能意向。例如，如果用户最近输入了搜索汽车的众多查询，或者该查询在汽车点 (carpoint) 网页的 web 搜索框中输入，则项 Saturn 的“汽车”意义是最适当的，并用于提供结果。

- 5 基于完成的查询和确定的查询类型，查询处理器在步骤 160 选择可在其上执行查询的一组数据源。查询是“compare price Saturn and Buick cat: cars: 80”形式的经修改的查询。这一查询形式指示预处理器 14 纠正了单词“Saturn”的拼写，并向查询扩充了该查询涉及类别“汽车 (cars)”的 80% 的置信度。

- 10 在阶段 170，可在数据源上并发地执行查询（如由识别器所增强的），或者可首先访问优选数据源，并在优选数据源不提供足够的结果或由于超负荷或技术困难而引起的“超时”的情况下，使用其它数据源。

- 15 数据源或提供商可以是运行在 web 服务器 20 上的内部提供商，或诸如 Encarta、Expedia、Overture、Inktomi、Yellow Pages 等外部提供商。诸如基于“en-us”意味着起源于美国的英语语言查询，向数据源提供增强查询和查询配置。从所有可能数据源的列表，可在增强查询和查询配置上构建两个列表。第一个列表是不依赖于其它数据源的源列表，第二个列表是的确依赖于其它源的那些源列表。第一个列表上的源首先被并行地调用，然后调用对第一列表中的来源具有依赖性的那些源。

- 20 为快速对流行的查询提供结果，查询处理器 10 高速缓存对流行查询的结果。搜寻类似于其结果被高速缓存的查询的结果的查询被首先定向到适当的高速缓存。这些高速缓存可根据高速缓存的信息改变的速率以不同的间隔更新，即每天或每小时。被标识为本地查询的查询被定向到黄页类型的目录数据源。被标识为汽车查询的查询被定向到汽车销售数据源。

- 25 返回的结果被去重复 (de-duplicate)，并由后处理组件 18 进行分级。基于上下文信息和查询类型向用户呈现结果。分级结果的呈现可基于记录的用户偏好来个性化。分级的结果也可被记录到一记录原始查询、结果查询、结果，并由用户选择结果的探测数据库 (instrumentation database)。探测数据库用于监控搜索引擎的成效。

识别器代理程序 210

返回到识别器代理程序 210，需要强调几点。首先，有多个识别器 221、222 等。在一个实施例中，代理程序 210 仅促使每一识别器评估查询的修改形式，并返回预测的查询类别。在一个替换实施例中，代理程序 210 基于从查询来源导出的其它信息选择识别器。例如，如果用户地址指示国家来源为“Spain（西班牙）”，

5 则向英语语言单词的列表匹配识别器发送该查询是无效的，因此代理程序使用对其可用的信息作出关于要使用的识别器的智能选择。某些代理程序不是基于单词而是算法的，并使用启发式规则来搜索诸如识别模式等意向。例如，如果在查询中出现五位数的串，则用于标识邮政编码的识别器将用高置信度来响应，这是涉及搜索关于国家的特定区域的本地搜索查询。以类似的方式，识别器搜索电话模式。

10 在一个示例性实施例中，识别器有两种类型—算法或列表匹配。算法查询意向识别器使用启发式规则来确定用户用他或她输入的单词要表达什么。一个示例是电话号码。检测是否输入了电话号码的规则可以是：三位数，其后跟随一分隔符，其后跟随七位数，或者三位数，其后跟随一分隔符，其后再跟随四位数。因此，如果用户输入“(425)882-8080”，则识别器代理程序用高置信度将该查询标记为电话

15 号码。这可帮助联合以代理方式来安排要联系哪一来源或提供商。算法查询意向识别器的其它示例是：

- 电话号码一如，“找出(425)880-8080 的名字（find name of (425)880-8080）”
- 邮政编码一如，“中国餐馆 98052（Chinese Restaurant 98052）”
- 电子邮件地址一如，“开发者 mcablu@microsoft.com（Developer
- 20 Mcablu@microsoft.com）”
- URL一如，“如何到达 yahoo.com（how to go to yahoo.com）”
- UPS 号码—“跟踪 Z9857103753300（Track Z9857103753300）”

如上所述，列表匹配查询意向识别器是基于字典查找模式的。对于字典中的每一条目，数据库具有其自己的单词或短语、候选类别和匹配概率。数据库

25 中的条目的一个子集可包括以下条目：

- Paris—city(80%); hotel(40%)（巴黎—城市（80%）；酒店（40%））
- Las Vegas—city(90%)（拉斯维加斯—城市（90%））
- Hotel—travel(80%)（酒店—旅游（80%））
- Jaguar—car(50%)（美洲虎—汽车（50%））
- 30 • Window—car(30%)（窗—汽车（30%））

- Jaguar—animal(50%) (美洲虎—动物 (50%))
- Restaurant—local(60%) (餐馆—本地 (60%))
- Hair Cut—local(50%) (剪发—本地 (50%))

5 如果用户输入诸如“Paris Hotel in Las Vegas (拉斯维加斯的巴黎酒店)”，则适当的查询识别器将指示该查询的特定部分包含城市 (Paris、Las Vegas)、包含酒店 (Paris) 并包含旅游 (Hotel)。识别器不仅报告每一单词或短语属于哪一类别，也报告短语中的位置。在上述示例中，该“Paris Hotel in Las Vegas”查询的结果为：

- 字符 1-5: 酒店 (40%)
- 10 • 字符 1-5: 城市 (80%)
- 字符 7-11: 旅游 (80%)
- 字符 16-24: 城市 (90%)

识别器代理程序将这一结果传递回联合处理器以采取关于查询分配的特定行动。

15 由识别器 221 (例如英语语言列表匹配识别器) 归于查询内的单词的置信度基于先前搜索的历史。数据库 230 基于数据库中维护的搜索历史维护单词列表和单词类别。从上述示例，数据库从过去的经验知道当向用户呈现包含单词“Saturn”的查询的结果时，他或她可能在当时有 68%的可能性对“汽车”类别感兴趣，因为他或她在呈现包含单词“Satrun”的查询的结果时，以那一频率点击到这一类别的链接。

20 表 1 的结果被概括为具有基于查询单词的组合置信度的结果。两个单词对汽车具有相对较高的置信度，而两个单词对购物具有较高的置信度。联合组件 16 可向两个专用搜索引擎发送该查询，一个涉及购物，而另一个涉及汽车。也可以知道存在一个适合“汽车购物”的特殊搜索站点。

25 查询意向识别阶段的其它使用是如果出现某一类别，执行该查询意向识别器的 web 服务器可选择性地显示 (或不显示) 广告。例如，如果查询的类别是“汽车”，则服务器可在“Results (结果)”网页上显示“Toyota ad (丰田汽车广告)”。对选择的另一响应是不显示内容。例如，如果识别器分析查询并确定它包含诸如“live sex (真人性行为)”等“成人项”，则该软件可使用这一信息来删除特定的联合或删除搜索结果页面的结果的元素。当前，呈现广告促销的服务器软件可逐
30 字提取查询的部分，并向为这一服务付款的广告商增加所提取的部分。识别器的使

用可通过自动增加未包含在查询中的内容以及在查询包含令人讨厌的语言的情况下对某些顾客删除广告来增强这一服务。

替换的示例性实施例不限于查询分类，并也关注查询扩充。考虑这两个示例：

示例 1：用户通过他或她的浏览器中的搜索文本框输入短语“Restaurants in
5 Redmond, WA（华盛顿州雷蒙德市的餐馆）”。识别器扩充该查询以形成短语
“Restaurants in Redmond, WA zip:98052:90 cat: local: 60”，其中，“zip:98052:90”
指这一查询有 90%的机会涉及邮政编码 98052，这是对搜索引擎的一条有用信息。
此外，“local: 60”的分类指这一查询以 60%的置信度为对本地搜索内容的请求。

示例 2：用户输入“News about Iraq（关于伊拉克的新闻）”，识别器以这一
10 方式扩充该查询“News about Iraq cat: news: 80 ranking: date: 30”，其中，“cat: news:
80”指它有 80%的机会是新闻类别，“ranking: date: 30”指分级器应当对日期字段
使用 30%的权值。

计算机系统 20

15 如参考图 1 所见的，用于实践本发明的代表性计算机系统 20 包括一个或多个
处理单元 21、系统存储器 22 以及将包括系统存储器的各类系统组件耦合至处理单
元 21 的系统总线 23。系统总线 23 可以是若干种总线结构类型的任一种，包括存
储器总线或存储器控制器、外围总线以及使用各类总线体系结构的局部总线。

系统存储器包括只读存储器（ROM）24 和随机存取存储器（RAM）25。基本
20 输入/输出系统（BIOS）26，包含如在启动时协助在计算机 20 内的元件之间传输
信息的基本例程，可储存在 ROM 24 中。

计算机 20 还包括用于对硬盘（未示出）进行读写的硬盘驱动器 27、用于对可
移动磁盘 29 进行读写的磁盘驱动器 28 以及用于对可移动光盘 31 如 CD ROM 或其
它光媒质进行读写的光盘驱动器 30。硬盘驱动器 27、磁盘驱动器 28 以及光盘驱动
25 器 30 分别通过硬盘驱动器接口 32、磁盘驱动器接口 33 和光盘驱动器接口 34 连接
至系统总线 23。驱动器及其相关的计算机可读媒质为计算机 20 提供了计算机可读
指令、数据结构、程序模块和其它数据的非易失存储。尽管这里描述的示例环境采
用了硬盘、可移动磁盘 29 以及可移动光盘 31，本领域的技术人员应当理解，也可
以在示例性操作环境中使可储存可由计算机访问的数据的其它类型的计算机可读
30 媒质，包括盒式磁带、闪存卡、数字视频盘、Bernoulli 盒式磁盘、随机存取存储器

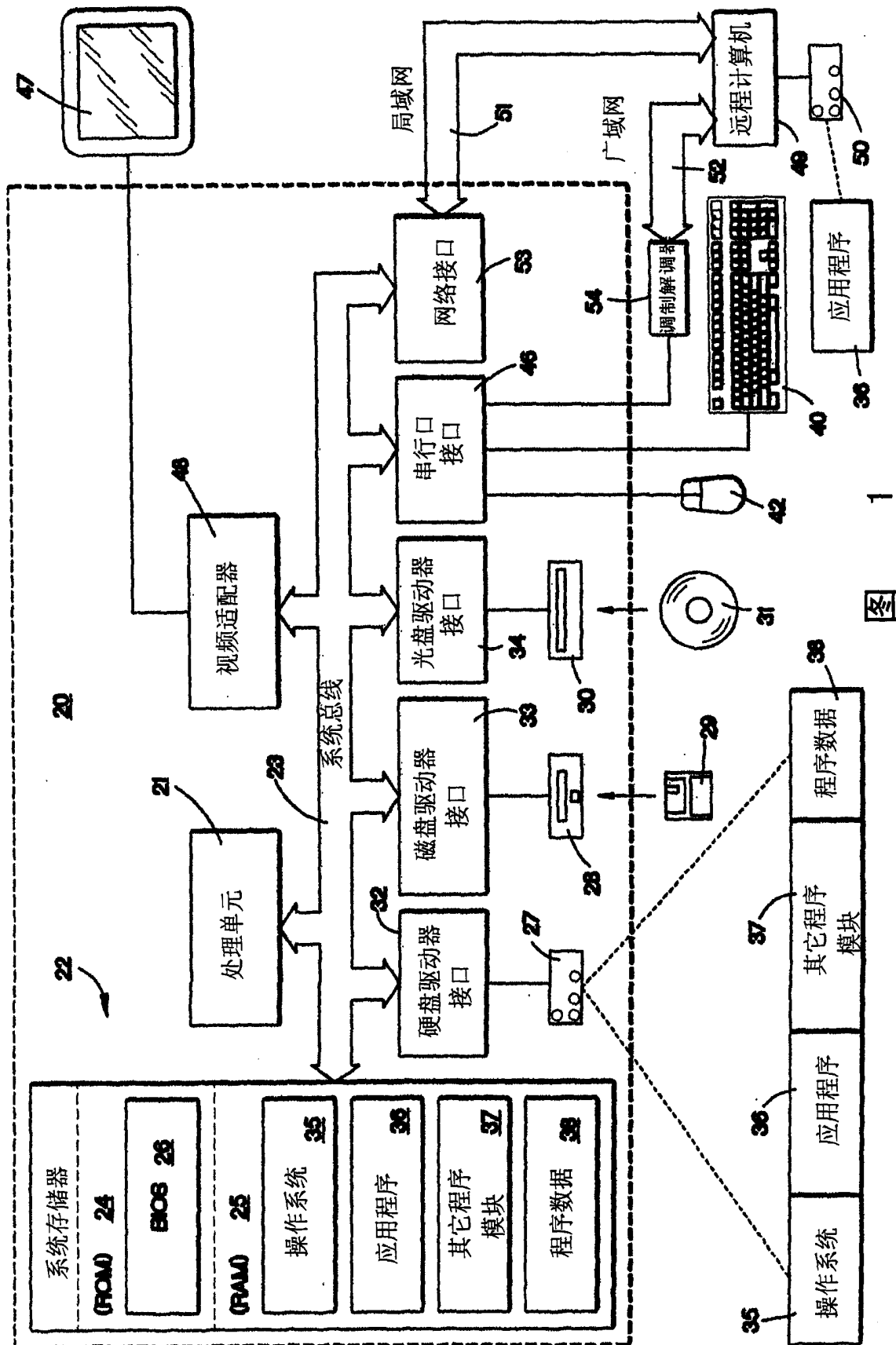
(RAM)、只读存储器(ROM)等等。

多个程序模块,包括数据挖掘软件组件 12 可储存在硬盘、磁盘 29、光盘 31、ROM 24 或 RAM 25 中,包括操作系统 35、一个或多个应用程序 36、其它程序模块 37 以及程序数据 38。用户可以通过诸如键盘 40 等输入设备以及定位设备 42 向计算机 20 输入命令和信息。其它输入设备(未示出)可包括麦克风、操纵杆、游戏垫、圆盘式卫星天线、扫描仪等等。这些和其它输入设备通常通过耦合至系统总线的串行端口接口 46 连接到处理单元 21,但也可通过其它接口连接,如并行端口、游戏端口或通用串行总线(USB)。监视器 47 或其它类型的显示设备也通过接口,如视频适配器 48 连接到系统总线 23。除监视器之外,个人计算机通常包括其它外围输出设备(未示出),如扬声器和打印机。

计算机 20 可以在使用到一个或多个远程计算机,如远程计算机 49 的逻辑连接的网络化环境中操作。远程计算机 49 可以是另一个人计算机、服务器、路由器、网络 PC、对等设备或其它公用网络节点,并通常包括许多或所有上述与计算机 20 相关的元件,尽管在图 1 中仅示出了存储器存储设备 50。图 1 描述的逻辑连接包括局域网(LAN) 51 和广域网(WAN) 52,这里示出作为示例而非局限。这类网络环境常见于办公室、企业范围计算机网络、内联网以及因特网。

当在 LAN 网络环境中使用时,计算机 20 通过网络接口或适配器 53 连接至局域网 51。当在 WAN 网络环境中使用时,计算机 20 通常包括调制解调器 54 或其它装置,用于通过广域网 52,如因特网建立通信。调制解调器 54 可以是内置或外置的,通过串行端口接口 46 连接至系统总线 23。在网络化环境中,描述的与计算机 20 相关的程序模块或其部分可储存在远程存储器存储设备中。可以理解,示出的网络连接是示例性的,也可以使用在计算机之间建立通信链路的其它装置。

从上述描述中可以看到,在中间查询结果上构建并维护统计信息可获得更有效的查询计划。尽管参考某一程度的细节描述了本发明,然而本发明意图包括落入所附权利要求书的精神或范围之内的所揭示的设计的所有修改和替换。



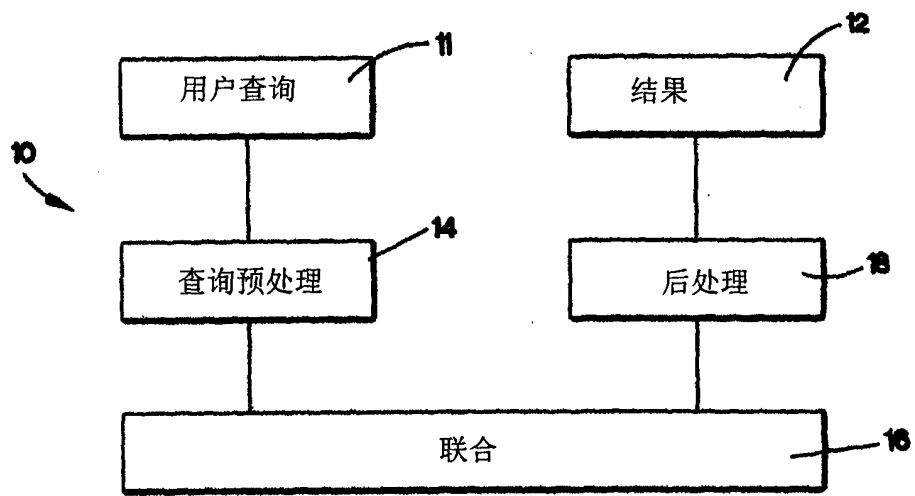


图 2

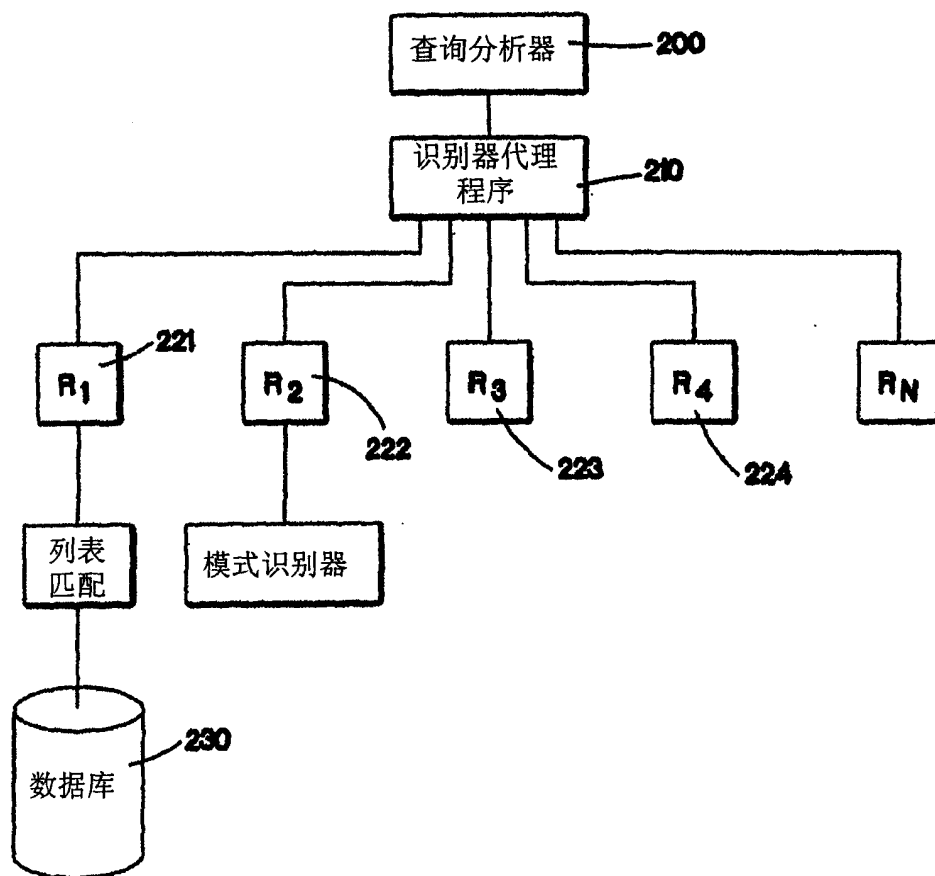


图 3

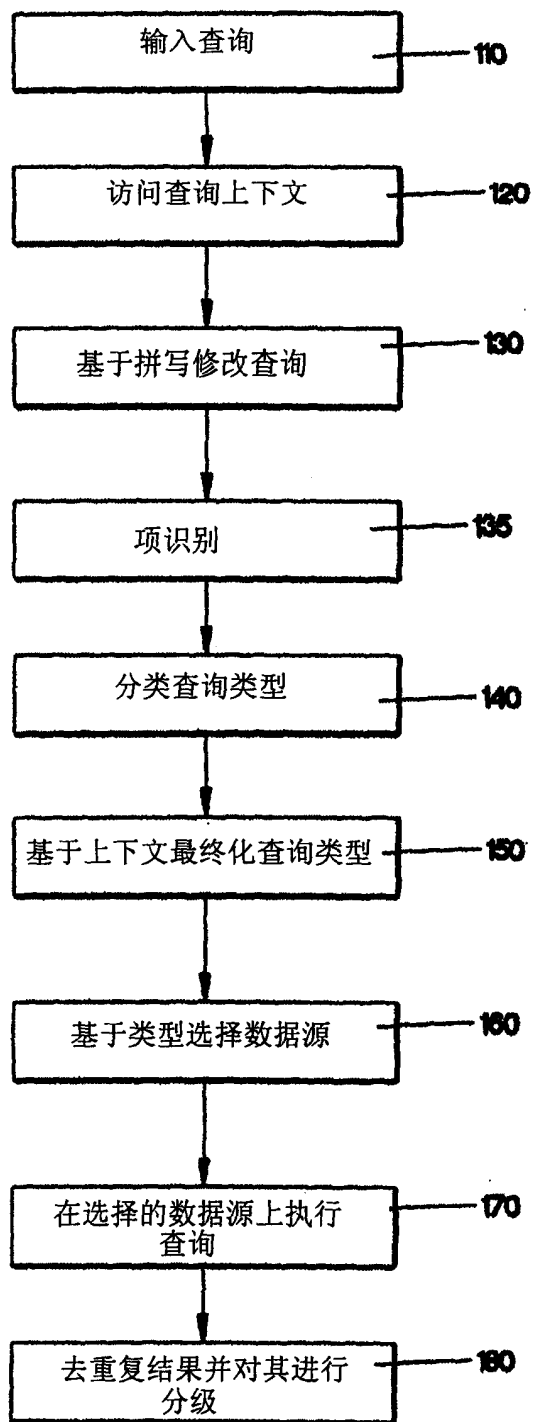


图 4