



(12) 发明专利申请

(10) 申请公布号 CN 103562916 A

(43) 申请公布日 2014. 02. 05

(21) 申请号 201280025518. X

(51) Int. Cl.

(22) 申请日 2012. 04. 25

G06F 17/30(2006. 01)

(30) 优先权数据

13/117, 042 2011. 05. 26 US

(85) PCT国际申请进入国家阶段日

2013. 11. 26

(86) PCT国际申请的申请数据

PCT/IB2012/052068 2012. 04. 25

(87) PCT国际申请的公布数据

W02012/160456 EN 2012. 11. 29

(71) 申请人 国际商业机器公司

地址 美国纽约

(72) 发明人 J·J·陶

(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 李玲

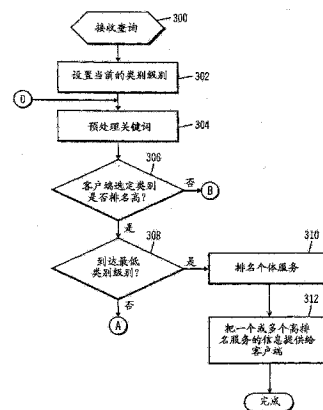
权利要求书4页 说明书14页 附图7页

(54) 发明名称

混合与迭代关键词以及类别搜索技术

(57) 摘要

本申请涉及混合与迭代关键词以及类别搜索技术。提供了用于提供改进查询的推荐的技术。接收具有查询关键词和选定类别的查询。响应于参考用于每个选定类别的查询相关指标值确定选定类别排名高,利用关键词相关指标计算查询与每个子类别的查询相关指标,基于查询相关指标给每个子类别排名,以及提供排名后的子类别,在选择要与查询一起提交的新类别时使用。



1. 一种用于提供改进查询的推荐的方法,包括:  
接收具有查询关键词和选定类别的查询;以及  
响应于参考用于每个选定类别的查询相关指标值确定选定类别排名高,  
利用关键词相关指标计算查询与每个子类别的查询相关指标;  
基于查询相关指标给每个子类别排名;以及  
提供排名后的子类别,以在选择要与查询一起提交的新类别时使用。
2. 如权利要求 1 所述的方法,进一步包括:  
确定是否到达选定类别中的最低类别级别;以及  
响应确定已经到达最低类别级别,  
给处于最低类别级别的个体服务排名;以及  
从排名后的个体服务提供一个或多个高排名的服务。
3. 如权利要求 1 或权利要求 2 所述的方法,进一步包括:  
响应于参考用于每个选定类别的查询相关指标值确定选定类别中至少有一个选定类别排名低,  
识别一对同义词,该对同义词中第一个同义词与查询关键词关联而该对同义词中第二个同义词与类别关键词关联;以及  
提供用来自该对同义词中的同义词代替查询关键词的推荐。
4. 如权利要求 3 所述的方法,其中该对同义词是为对非排名靠前的选定类别具有低关键词相关指标值的查询关键词识别的。
5. 如前面任何一项权利要求所述的方法,进一步包括:  
对于非排名靠前的选定类别中具有高关键词相关指标值的类别关键词,识别不在查询中而且没有为其识别出同义词的类别关键词;以及  
提供把所识别出的类别关键词添加到查询的推荐。
6. 如前面任何一项权利要求所述的方法,进一步包括:  
识别对非排名靠前的选定类别具有低关键词相关指标值并且对排名靠前但未选的类别具有高关键词相关指标值的查询关键词;  
从识别出的查询关键词,识别没有为其识别出同义词的查询关键词;以及  
提供把没有为其识别出同义词的查询关键词从查询中除去的推荐。
7. 如前面任何一项权利要求所述的方法,进一步包括:  
提供关于用于在修正查询时使用的查询关键词的推荐;以及  
接收基于所提供的推荐的修正后的查询。
8. 如前面任何一项权利要求所述的方法,进一步包括:  
接收不执行进一步迭代的指示;  
以当前形式执行查询;以及  
提供一个或多个服务的列表。
9. 如前面任何一项权利要求所述的方法,进一步包括:  
迭代通过各级别类别。
10. 一种用于提供改进查询的推荐的系统,包括:  
处理器;以及

匹配引擎,可以耦合到处理器并且可操作执行操作,所述操作包括:  
接收具有查询关键词和选定类别的查询;以及  
响应于参考用于每个选定类别的查询相关指标值确定选定类别排名高,  
利用关键词相关指标计算查询与每个子类别的查询相关指标;  
基于查询相关指标给每个子类别排名;以及  
提供排名后的子类别,以在选择要与查询一起提交的新类别时使用。

11. 如权利要求 10 所述的系统,其中所述操作进一步包括:

确定是否到达选定类别中的最低类别级别;以及  
响应于确定已经到达最低类别级别,  
给处于最低类别级别的个体服务排名;以及  
从排名后的个体服务提供一个或多个高排名的服务。

12. 如权利要求 10 或权利要求 11 所述的系统,其中所述操作进一步包括:

响应于参考用于每个选定类别的查询相关指标值确定选定类别中至少有一个选定类别排名低,

识别一对同义词,该对同义词中第一个同义词与查询关键词关联而该对同义词中第二个同义词与类别关键词关联;以及

提供用来自该对同义词中的同义词代替查询关键词的推荐。

13. 如权利要求 12 所述的系统,其中该对同义词是为对非排名靠前的选定类别具有低关键词相关指标值的查询关键词识别的。

14. 如权利要求 10 至 13 中任何一项所述的系统,其中所述操作进一步包括:

对于非排名靠前的选定类别中具有高关键词相关指标值的类别关键词,识别不在查询中而且没有为其识别出同义词的类别关键词;以及

提供把所识别出的类别关键词添加到查询的推荐。

15. 如权利要求 10 至 14 中任何一项所述的系统,其中所述操作进一步包括:

识别对非排名靠前的选定类别具有低关键词相关指标值并且对排名靠前但未选的类别具有高关键词相关指标值的查询关键词;

从识别出的查询关键词,识别没有为其识别出同义词的查询关键词;以及

提供把没有为其识别出同义词的查询关键词从查询中除去的推荐。

16. 如权利要求 10 至 15 中任何一项所述的系统,其中所述操作进一步包括:

提供关于用于在修正查询时使用的查询关键词的推荐;以及  
接收基于所提供的推荐的修正后的查询。

17. 如权利要求 10 至 16 中任何一项所述的系统,其中所述操作进一步包括:

接收不执行进一步迭代的指示;

以当前形式执行查询;以及

提供一个或多个服务的列表。

18. 一种用于提供改进查询的推荐的计算机程序产品,所述计算机程序产品包括:

计算机可读存储介质,其中包含计算机可读程序代码,所述计算机可读程序代码包括:

计算机可读程序代码,当被计算机的处理器执行时,可配置成执行:

接收具有查询关键词和选定类别的查询；以及  
响应于参考用于每个选定类别的查询相关指标值确定选定类别排名高，  
利用关键词相关指标计算查询与每个子类别的查询相关指标；  
基于查询相关指标给每个子类别排名；以及  
提供排名后的子类别，以在选择要与查询一起提交的新类别时使用。

19. 如权利要求 18 所述的计算机程序产品，其中，所述计算机可读程序代码当被计算机的处理器执行时配置成执行：

确定是否到达选定类别中的最低类别级别；以及  
响应确定已经到达最低类别级别，  
给处于最低类别级别的个体服务排名；以及  
从排名后的个体服务提供一个或多个高排名的服务。

20. 如权利要求 18 或权利要求 19 所述的计算机程序产品，其中，所述计算机可读程序代码当被计算机的处理器执行时配置成执行：

响应于参考用于每个选定类别的查询相关指标值确定选定类别中至少有一个选定类别排名低，

识别一对同义词，该对同义词中第一个同义词与查询关键词关联而该对同义词中第二个同义词与类别关键词关联；以及

提供用来自该对同义词中的同义词代替查询关键词的推荐。

21. 如权利要求 20 所述的计算机程序产品，其中该对同义词是为对非排名靠前的选定类别具有低关键词相关指标值的查询关键词识别的。

22. 如权利要求 18 至 21 中任何一项所述的计算机程序产品，其中，所述计算机可读程序代码当被计算机的处理器执行时配置成执行：

对于非排名靠前的选定类别中具有高关键词相关指标值的类别关键词，识别不在查询中而且没有为其识别出同义词的类别关键词；以及

提供把所识别出的类别关键词添加到查询的推荐。

23. 如权利要求 18 至 22 中任何一项所述的计算机程序产品，其中，所述计算机可读程序代码当被计算机的处理器执行时配置成执行：

识别对非排名靠前的选定类别具有低关键词相关指标值并且对排名靠前但未选的类别具有高关键词相关指标值的查询关键词；

从识别出的查询关键词，识别没有为其识别出同义词的查询关键词；以及

提供把没有为其识别出同义词的查询关键词从查询中除去的推荐。

24. 如权利要求 18 至 23 中任何一项所述的计算机程序产品，其中，所述计算机可读程序代码当被计算机的处理器执行时配置成执行：

提供关于用于在修正查询时使用的查询关键词的推荐；以及

接收基于所提供的推荐的修正后的查询。

25. 如权利要求 18 至 24 中任何一项所述的计算机程序产品，其中，所述计算机可读程序代码当被计算机的处理器执行时配置成执行：

接收不执行进一步迭代的指示；

以当前形式执行查询；以及

提供一个或多个服务的列表。

## 混合与迭代关键词以及类别搜索技术

### 技术领域

[0001] 本发明的实施例涉及用于分布式计算系统与网络环境,而且更具体地说,用于任何基于网络的服务的混合与迭代关键词以及类别搜索技术。

### 背景技术

[0002] 在过去的二十年当中,互联网与万维网(WWW 或 Web)使信息技术有了巨大的变化。作为人-机和客户端-服务器技术,Web 使具有连接到互联网的计算机的任何人都能够从他或她的指尖访问互联网上任何公布的信息。

[0003] Web 页面是用例如超文本标记语言(HTML)编写的文档。由于有巨大量的 web 内容,因此大多数 web 用户依赖 web 搜索引擎通过关键词搜索来搜索有用的 web 页面。

[0004] 为了方便计算系统的集成并且向用户提供更加交互式的用户体验和丰富的内容,近年来已经推出了新的 web 技术,它们当中许多都基于可扩展的标记语言(XML)。这些技术中的两种是 Web 服务与异步 JAVASCRIPT 和 XML (AJAX)。(JAVASCRIPT 是 Oracle 和 / 或其附属公司的商标或注册商标。) Web 服务可以描述为机器到机器的分布式计算技术,这种技术克服了使计算机程序能够在异类计算环境中彼此通信的困难。这些困难是由不同的计算机平台、不兼容的通信协议和计算机程序使用的各种计算机语言引起的。作为战胜这些挑战的一种基于标准的全面解决方案,Web 服务被各行业广泛支持。Web 服务基于一系列标准,例如 XML、简单对象访问协议(SOAP)和 Web 服务描述语言(WSDL)。这些标准为运行在计算机和电子设备上的应用提供了经网络在它们之间交换信息的共用格式、语法和协议。与 Web 服务不同,AJAX 没有定义一组标准。AJAX 使 web 应用能够异步地把数据从客户端发送到服务器。AJAX 可以用于实现 REST (代表性状态传输)风格的 Web 服务。

[0005] 例如,一个公司可以为不同的厂家创建销售他们产品的市场网站。web 服务的例子包括 web 站点的灵活实现 web 服务和支付 web 服务,这些服务用于集成该市场网站与那些厂家的信息系统。

[0006] 为了方便公布和搜索 web 服务,已经开发出了通用描述、发现与集成(UDDI)标准。UDDI 标准定义如何创建 web 服务 UDDI 注册表,以便使 web 服务提供商能够公布他们的 web 服务并且使 web 服务消费者能够搜索并使用这些公布的 web 服务。

[0007] 非 UDDI 的 web 服务注册表可以提供 web 服务支配特征与语义 web 技术。这种 web 服务注册表或储存库存储附加的与 web 服务相关的元数据,以便支配 web 服务的生命周期。Web 本体语言(OWL)是一种基于本体论的标记语言,它最初是在学术研究中开发的,以便以机器可理解的形式在 web 上呈现数据。OWL 可以用于在 web 服务注册表中组织与 web 服务相关的元数据。

[0008] 在常规系统中,关键词搜索被 web 服务消费者用来在 web 服务注册表中找出恰当的 web 服务。Web 服务的关键词可以由 web 服务提供商手动指定。自动关键词生成过程可以用于从 web 服务元数据生成关键词。也可以使用手动方法与自动方法的组合,例如让提供商验证或修改所生成的关键词。

[0009] UDDI 与其它的注册表提供了查询应用编程接口(API)和/或图形用户接口(GUI),以使 web 服务消费者能够搜索在 web 服务注册表中公布的 web 服务。利用这些查询 API 或 GUI,用户提供进行搜索的具体 web 服务元数据域的关键词、串或其它数据。

[0010] 例如,UDDI 客户端可以查询 UDDI 注册表,以基于名字、它们所属的商业实体和它们所落入的类别找出 web 服务。在这个例子中,用户提供 web 服务、商业实体和/或类别的部分或完整名字,来构造这种搜索查询。

[0011] 现有的 web 服务注册表技术允许用户利用合成的查询搜索 web 服务。这种查询的搜索结果是构成合成查询的简单查询的搜索结果集合的交集或并集。

[0012] 本领域中需要发现服务例如 web 服务的改进技术。

[0013] US20100185619 公开了一种采样分析,它包括根据与多个查询关键词关联的页面视图值把多个查询关键词分类为多个查询关键词子集 - 例如,见摘要;附图 1-6;和行 [0023]-[0035]。

[0014] US20100174691 公开了至少部分地基于与缺陷有关的关键词把缺陷自动归类到一个类别中的方法与系统 - 例如,见摘要;附图 1-8;和行 [0015]-[0017]、[0030]-[0076]。

[0015] US20100131563 公开了包括多个搜索列表的一个搜索结果页面,所述多个搜索列表是通过施加到基于用户查询选择的初始匹配文档集合的多个群集操作生成的 - 例如,见摘要;附图 1-9;和行 [0014]-[0019]、[0046]-[0047]。

[0016] US20100036822 公开了用于从先前的结果数据推断类别同义词的系统与方法 - 例如,见摘要;附图 1-7;和行 [0008]-[0011]、[0036]-[0060]。

[0017] US20090222444 公开了在提交到一个或多个搜索引擎之前分解搜索查询。查询被分解成使得该查询明确地对应于包括在把搜索查询关联到搜索类别的查询本体中的类别 - 例如,见摘要;附图 1-17;和行 [0005]-[0012]、[0088]-[0124]。

[0018] US20070118509 公开了用于建议给数字图像加标签的关键词的合作定位关键词注册表服务 - 例如,见摘要;附图 1-8;和行 [0005]-[0007]、[0019]-[0063]。

[0019] US7,620,627 公开了一种生成在搜索内容时所使用的关键词的方法,该方法包括接收关键词、从类别文件获得对应于该关键词的类别、接收对应于这些类别中至少一个类别的选择并且生成与该选择关联的关键词 - 例如,见摘要;附图 1-3;和第 1 栏第 10-67 行、第 2 栏第 1-25 行。

[0020] US6,571,239 公开了解决被用户用来进行查询的关键词以及存储在系统关键词索引中、被手动或自动分类系统指定的那些关键词之间的失配问题的方法 - 例如,见摘要;附图 1-5;和第 1 栏第 5-67 行、第 2 栏第 1-15 行、第 3 栏第 40-55 行。

## 发明内容

[0021] 提供了用于提供改进查询的推荐的方法、计算机程序产品与系统。接收具有查询关键词和选定类别的查询。响应于参考用于每个选定类别的查询相关指标值确定选定类别排名高,利用关键词相关指标计算查询与每个子类别的查询相关指标,基于查询相关指标给每个子类别排名,以及提供排名后的子类别,以在选择要与查询一起提交的新类别时使用。

## 附图说明

[0022] 现在参考附图,其中相同的标号贯穿所有附图都代表对应的部分:

[0023] 图 1 说明了根据某些实施例、包括查询增强系统的计算体系结构。

[0024] 图 2 说明了根据某些实施例、在分类层次或群集层次中具有两个子类别的一个类别的图。

[0025] 图 3 以流程图说明了根据某些实施例、在查询增强系统中由关键词与分类类别匹配过程执行的逻辑。图 3 由图 3A、图 3B、图 3C 和图 3D 构成。

[0026] 图 4 以框图说明了根据某些实施例可以使用的计算机体系结构。

## 具体实施方式

[0027] 在以下描述中,参考构成本说明书一部分并且说明本发明几个实施例的附图。应当理解,在不背离本发明范围的情况下,可以使用其它实施例而且可以进行结构与操作变化。

[0028] 在实施例中,本文提到的服务指在信息系统上实现并且可以从电信网络访问的任何服务。服务包括但不限于 web 服务。

[0029] 图 1 说明了根据某些实施例、包括查询增强系统 120 的计算体系结构。服务客户端 100 通过通信网络与服务注册服务器 110 交互。服务注册服务器 110 包括查询增强系统 120 和服务注册表 170。在某些实施例中,服务客户端 100 可以通过由查询增强系统 120 提供的一个或多个用户接口与服务注册服务器 110 交互。

[0030] 查询增强系统 120 提供一种集成和迭代的过程,以为服务注册表 170 中所公布的一种或多种服务生成并识别恰当的关键词与分类和 / 或群集类别,并且所识别出的关键词用于增强(即,改进)查询。查询增强系统 120 提供了以集成的方式组合关键词搜索与类别选择 / 搜索的混合技术。另一方面,常规系统可以通过执行两个单独的查询、然后利用逻辑操作(例如“AND”或“OR”)组合结果来把关键词与类别搜索组合成合成查询。

[0031] 查询增强系统 120 具有四个组件:分类 - 群集与服务关键词数据存储 130 (“关键词数据存储”130)、关键词预处理器 140、关键词分类 - 群集匹配引擎 150 (“匹配引擎”150)和关键词词库 160。

[0032] 查询增强系统 120 利用基于迭代关键词与类别的过程来发现例如面向服务的信息系统与网络中可用的 web 服务。在可供选择的实施例中,查询增强系统 120 可以发现除 web 服务之外的项。

[0033] 服务注册表 170 使服务提供商能够公布他们的服务并且使服务消费者能够搜索并使用这些公布的服务。服务注册表 170 存储所公布服务的信息。

[0034] 匹配引擎 150 提供一种机制,以通过允许用户提供反馈来把关键词搜索与类别浏览集成到相互校正与自校正搜索过程中。

[0035] 关键词数据存储 130 保留由服务提供商提供的或者从服务元数据生成的服务关键词的信息,以及这些关键词与分类 / 群集类别之间的关系信息。关键词数据存储 130 中的信息是从存储在服务注册表 170 中的服务信息检索或导出的。

[0036] 关键词与分类 / 群集类别之间的关系信息定义为相关指标(indicator),这是与一个类别关联的关键词的权重(并且在本文中称为关键词相关指标)。相关指标也用作与一



个类别关联的查询的权重(并且在本文中称为查询相关指标)。

[0037] 关键词预处理器 140 被用来验证查询关键词是否有效。关键词词库 160 在识别关键词同义词的过程中被采用。参考同义词,词的意义依赖于它们在其中使用的上下文。例如,服务提供商使用的术语可能与服务客户端 100 使用的术语不同。

[0038] 服务可以分成分类或群集。分类可以由标准机构创建并且可以具有明确定义并且很好理解的子类别名字。群集或群集层次可以由服务注册服务器 110 创建,其中服务注册服务器 110 实现本文所述的迭代关键词与类别搜索技术。在某些实施例中,如果服务具有相似的关键词,就把它们分到相同的群集中。在某些实施例中,群集不具有良好定义的名字。当识别出一个群集并且需要将其呈现给用户时,该群集中一种典型样本服务的信息被发送到用户。用户通过指定该样本服务是否与用户试图找出的服务相似来决定是否选择该群集。在各种实施例中,用户可以是人、计算机程序、设备等。

[0039] 在某些实施例中,本文所使用的分类指相同类型的事物分到其中的标准类别,例如北美行业分类系统(NAICS)。通常,这些分类是由标准机构开发的而且,如果它们不是内置特征的话,可以插入服务注册表中。分类是一种归类机制,它可以在关键词搜索技术中使用。不管怎样,任何分类都具有有限和固定的级别。在一个分类的最精细的级别,有可能有数千或者更多的服务分到相同的子类别中。在这种情况下,在这个最精细的子类别中期望进一步的归类/分组。

[0040] 服务客户端 100 与查询增强系统 120 之间的通信是迭代的过程。服务客户端 100 与查询增强系统 120 来回传递关键词和分类-群集类别信息多次,以便为处于服务客户端 100 的用户试图找出的服务识别恰当的关键词和分类-群集类别。

[0041] 当服务客户端 100 与查询增强系统 120 通信时,关键词预处理器 140 首先从服务客户端 100 接收查询并且检查关键词,以确信关键词有效,例如没有拼写错误。如果关键词预处理器 140 识别出关键词中的错误,则关键词预处理器 140 通知匹配引擎 150,而且匹配引擎 150 在发送回服务客户端 100 的消息中把该信息转发到服务客户端 100。

[0042] 匹配引擎 150 是实现匹配技术的组件。匹配引擎 150 从关键词预处理器 140 接收预处理之后的查询、从关键词数据存储 130 检索分类-群集类别关键词、从关键词词库提取关键词的同义词、比较查询中的关键词与类别中的关键词,并且生成更新后的关键词版本与排名后的类别列表。匹配引擎 150 把修改后的关键词与类别信息提供回服务客户端 100,用于进一步的反馈和调整。

[0043] 关键词数据存储 130 是其中存储用于每个服务与分类/群集类别的关键词的数据存储。查询增强系统 120 计算用于每个类别的每个关键词的关键词相关指标并且把这些相关指标的值存储在关键词数据存储 130 中。

[0044] 关键词词库 160 是由匹配引擎 150 用于找出两组关键词之间的同义词的词库。在某些实施例中,如果两个关键词具有相同或者非常相似的意义,则它们是同义词。

[0045] 在迭代过程中通过指定或选择多个关键词并且通过选择/指定服务所属的恰当分类或群集类别,查询增强系统 120 使服务消费者能够识别在服务注册表中公布的恰当服务。特别地,匹配引擎 150 识别最适合的服务所属的恰当关键词与类别。在相互校正过程中利用在每个分类与群集级别的关键词选择,对合适服务的搜索范围可以通过向下导航分类与群集层次而有效地缩小。同义词问题利用关键词词库 160 来解决。在该过程结束时,

查询增强系统 120 允许服务用户检索候选服务的列表。

[0046] 在某些实施例中,查询增强系统 120 为服务消费者和提供商都提供关于关键词选择和类别识别的帮助。通过给类别排名并且引导服务用户做出正确的关键词选择,查询增强系统 120 方便了服务搜索过程。

[0047] 对于查询增强系统 120,内置了利用被查询的关键词与类别之间的关系、给用户推荐(即,建议)和分析用户的反馈以便作出更有效的搜索的智能。

[0048] 查询增强系统 120 提供了收集并利用服务关键词与服务所属的层次性分类与群集类别之间关系的过程。相同的概念也可以应用到服务注册表中其它上下文相关的服务元数据。

[0049] 更具体地说,通过比较用户提供的关键词和与每个分类和 / 或群集类别关联的关键词,利用来自关键词词库 160 的额外帮助和来自用户的反馈,匹配引擎 150 可以快速识别出恰当的关键词和相关的服务。用户的反馈包括迭代地提供关于服务的关键词与类别信息。

[0050] 当服务注册表存储了数千或者甚至数百万种服务的信息时,要求用户给出关于他们想找出的服务的详细和具体信息常常是低效的。相反,允许用户给出多个关键词、然后基于匹配引擎 150 提供的推荐关键词进行直截了当的选择可能更实际。即,对于查询增强系统 120,匹配引擎 150 的任务是帮助识别对用户最适合的服务。

[0051] 查询增强系统 120 提供了匹配引擎 150,以利用有限的或者没有关于注册表结构以及注册表中所公布服务的确切细节的事先了解,使用户能够搜索在注册表中公布的服务(以类似于 web 搜索引擎如何用于搜索 web 页面的方式)。

[0052] 类似地,查询增强系统 120 所采用的关键词与归类信息改进技术不仅能帮助服务消费者搜索服务,而且能帮助服务提供商归档并分类它们的服务。

[0053] 图 2 说明了根据某些实施例、在分类层次或群集层次中具有两个子类别的一个类别的图。特别地,图 2 是说明两级类别层次的图。类别 A200 (在第一级别)具有两个子类别,类别 \_A1210 (在第二级别)和类别 \_A2220 (在第二级别)。

[0054] 图 2 用于描述两个实体例如关键词与类别之间关系的测量、相关指标。

[0055] 在某些实施例中,当一个服务在服务注册表中公布时,服务提供商提供该服务的关键词与分类信息。如果一个服务属于一个类别,则其所有关键词都属于这个类别并且属于其所有祖先类别。

[0056] 例如,如果汽车保险报价服务在注册表中公布,则服务提供商可以指定这个服务属于北美行业分类系统(NAICS)中的类别“保险代理 & 经纪业务”。这个类别是类别“保险运营商及相关活动”的子类别,而“保险运营商及相关活动”又是类别“金融与保险”的子类别。在这个例子中,为这个服务指定的所有关键词也都是用于类别“保险代理 & 经纪业务”及其父类别“保险运营商及相关活动”和祖父类别“金融与保险”的关键词。每个类别中的关键词可以具有与该类别关联的不同权重,称为关键词相关指标。

[0057] 相关指标可以描述为测量两个文档有多么相关的权重。每个文档都包含一组关键词。文档可以是单个关键词、服务、查询或类别。

[0058] 特别地,一个具有关键词列表的查询可以被看作是一个文档。一个包含关键词集合的类别也可以被看作是一个文档。基于其关键词的权重来测量文档相似性的不同方案已

经在信息检索研究中开发出来了,例如余弦相似性、欧氏距离、骰子系数和杰卡德指数。这些方案中有些可以用来测量查询与类别之间的相似性。在常规的信息检索系统中,在这些方案中一些当中所使用的关键词的权重是利用基于词频-逆文档频率(TF-IDF)的技术计算的。

[0059] 一个服务的关键词集合也可以被看作是一个文档。为了进一步给服务归类,当分类达到其最精细的级别时,多个相似性方案,例如余弦相似性或者欧氏距离,可以用于基于它们的关键词测量服务之间的相似性。如果单个群集中有太多服务,则这个群集中的服务可以进一步分成更小的群集。这个过程可以递归进行,以产生群集层次。多个群集或群集层次可以在分类层次中建立,以方便迭代关键词搜索。可以实现在层次性文档群集研究中开发出的技术,例如凝聚层次结构群集或基于层次性频率项的群集,以便创建群集层次。

[0060] 关键词相关指标(RI)是一个与 TF-IDF 相似的测量,这是一种用于评估一个项对文档集合中的一个文档有多重要的加权方案。

[0061] 关于相同的记号,与一个类别关联的关键词的关键词频率(KF)类似于 TF。KF 测量一个关键词在一个类别中出现得多频繁。在某些实施例中,它被定义为类别中包含该关键词的服务总数的与该类别中每个服务具有的所有关键词的个数之和的商。等式(1)规定了  $KF_i$  :

$$[0062] \quad KF_i = \frac{\sum_{j=1}^N n_{ij}}{\sum_{j=1}^N \sum_{k=1}^M n_{kj}}$$

[0063] 在等式(1)中,  $KF_i$  是一个类别中关键词  $i$  的关键词频率;  $M$  是该类别中关键词的总数;  $N$  是该类别中服务的总数;  $i$  和  $k$  都是 1 和  $M$  之间的整数; 对于任何  $i \in [1, M]$  和  $j \in [1, N]$ , 如果该类别中的服务  $j$  有关键词  $i$ , 则  $n_{ij}$  是 1, 而如果该类别中的服务  $j$  没有关键词  $i$ , 则  $n_{ij}$  是 0; 而且对于任何  $k \in [1, M]$  和  $j \in [1, N]$ , 如果该类别中的服务  $j$  具有关键词  $k$ , 则  $n_{kj}$  是 1, 而如果该类别中的服务  $j$  没有关键词  $k$ , 则  $n_{kj}$  是 0。

[0064] 等式(1)假设每种服务都具有相同数量的关键词。如果情况不是这样,则在某些实施例中服务被同等地对待,代替如果服务  $j$  具有关键词  $k$  则为 1,  $n_{kj}$  (或  $n_{ij}$ ) 的值可以调整成使得,对于每种服务  $j$ ,  $n_{kj}$  之和,  $\sum_{k=1}^M n_{kj}$ , 对于所有服务都是相同的。

[0065] 但是,服务集合中关键词的唯一性测量,像逆文档频率(IDF),在某些实施例中可能没有在类别级别中定义,因为包括大量服务的类别可能包含相同的关键词。因此,包含该关键词的类别的个数与整个分类系统中类别的总数之商总是等于或者接近等于 1,在某些实施例中,使得这种测量是低效的。如果包含个体关键词的文档在服务级别定义,就可以避免这个问题。在某些实施例中,文档集合可以在整个分类/群集域的范围定义。不管怎样,比较相同父类别中的类别都是典型的。通过把文档集合定义为测量其关键词唯一性的一个/多个具体类别的一个/多个父类别在更精细的级别测量关键词唯一性更准确。即,在某些实施例中,在一个或多个具体类别的逆服务频率(ISF)被定义为一个/多个父类别中服务总数与包含该关键词的服务个数之商的的对数。等式(2)规定  $ISF_i$  :

$$[0066] \quad ISF_i = \log \frac{N}{\sum_{j=1}^N n_{ij}}$$

[0067] 在等式(2)中,  $ISF_i$  是一个分类系统中一个或多个具体类别中关键词  $i$  的逆服务频率;如果服务  $j$  有关键词  $i$ , 则  $n_{ij}$  为 1, 而如果服务  $j$  没有关键词  $i$ , 则  $n_{ij}$  为 0; 而且  $N$  是在一个或多个父类别中注册的服务的总数。

[0068]  $RI_i$  是关键词  $i$  与一个类别的关键词相关指标。等式(3)规定  $RI_i$  :

$$[0069] \quad RI_i = KF_i \times ISF_i$$

[0070]  $RI_Q$  是一个查询与一个类别的查询相关指标, 用来测量该查询与类别之间的相似性。包含多个关键词的查询的  $RI_Q$  是这些关键词与这个类别的关键词相关指标的聚集。在某些实施例中, 类似于与在信息检索研究中开发的矢量空间模型一起使用的余弦相似性,  $RI_Q$  是代表查询与类别的两个多维矢量之间角度的余弦。等式(4)规定  $RI_Q$  :

$$[0071] \quad RI_Q = \frac{\sum_{i=1}^{M_Q} RI_i}{(M_Q \times (\sum_{i=1}^{M_Q} RI_i^2 + \sum_{j=M_Q+1}^{M_C} RI_j^2))^{\frac{1}{2}}}$$

[0072] 在等式(4)中,  $RI_i$  是关键词  $i$  与这个类别的关键词相关指标; 关键词  $i$  是在该查询中指定的关键词;  $M_Q$  是该查询中的关键词总数;  $RI_j$  是关键词  $j$  与这个类别的关键词相关指标; 关键词  $j$  是该类别中但不在查询中的一个关键词; 而  $M_C$  是该类别和查询中不同关键词的总数(即,  $M_C$  的大小是查询关键词集合与类别关键词集合的并集)。

[0073] 由于基于矢量空间模型的等式(4)可能对具有更多服务和关键词的大类别生成过于小的  $RI_Q$  值, 因此, 在某些实施例中, 如果进行比较的类别的尺寸显著不同, 则  $RI_Q$  的值需要调整。

[0074] 说明以上概念的一个简单例子在图 2 中示出。如果类别  $\_A1210$  有 5 个服务, 而类别  $\_A2220$  有 10 个服务, 则类别  $\_A200$  总共有 15 个服务。这些服务都是唯一的。假设在类别  $\_A1210$  中只有一个服务包含关键词“保险(insurance)”, 而类别  $\_A2220$  中有五个服务包含相同的关键词“保险”。这意味着类别  $\_A$  有六个包含关键词“保险”的服务。此外, 每个服务都包含 10 个关键词。在类别  $\_A200$  的父类别中, 假设每五个服务中有一个服务包含关键词“保险”。关键词“保险”与每个类别的关键词相关指标是:

$$[0075] \quad \text{类别 } \_A: RI_{\text{insurance}} = \frac{6}{10 \times 15} \times \log 5 = 0.028$$

$$[0076] \quad \text{类别 } \_A1: RI_{\text{insurance}} = \frac{1}{10 \times 5} \times \log \frac{15}{6} = 0.008$$

$$[0077] \quad \text{类别 } \_A2: RI_{\text{insurance}} = \frac{5}{10 \times 10} \times \log \frac{15}{6} = 0.020$$

[0078] 图 3 以流程图说明了根据某些实施例、在查询增强系统 120 中由关键词与分类类别匹配过程执行的逻辑。图 3 由图 3A、图 3B、图 3C 和图 3D 构成。

[0079] 在图 3 中, 查询增强系统 120 处理从服务客户端 100 接收到的查询。在某些实施

例中,匹配迭代在每个分类系统中的一个具体分类级别执行(即,服务注册表 170 可以具有多个分类系统)。查询增强系统 120 在每次迭代结束时向服务客户端 100 返回更新后的关键词与分类类别信息。然后,服务客户端 100 可以通过选择所显示的查询关键词、类别关键词与类别来提供反馈,而且该选择在下一迭代中使用。在到达其最低类别级别之前,如利用方框 308-312 所说明的,查询增强系统 120 所采用的查询改进过程可以迭代多次,或者服务客户端 100 可以退出过程,如利用方框 334-342 所说明的。

[0080] 处理在方框 300 开始(图 3A),查询增强系统 120 从服务客户端 100 接收查询 Q。该查询中的关键词将被称为查询关键词,而分类类别中的关键词将被称为类别关键词。在某些实施例中,在第一次迭代中,由服务客户端 100 生成的查询可以包括查询关键词,而且可以包括或者可以不包括由服务客户端 100 指定的分类类别。但是,在第一次迭代和每次后续的迭代结束时,查询增强系统 120 都可以推荐类别和类别关键词,而且服务客户端 100 可以选择要作为查询关键词包括在查询中的这些类别关键词中的一个或多个。在某些实施例中,后续迭代中的每个查询都包含由服务客户端 100 指定或确认的关键词和分类类别。查询中这些指定或确认的类别被称为选定类别。对于每个分类系统,查询都有可能具有多于一个的选定类别。

[0081] 在方框 302,匹配引擎 150 设置当前的类别级别。在某些实施例中,如果服务客户端 100 在开始的时候没有指定任何分类类别,则新的查询在每个分类系统的根 / 顶类别级别开始。在每次迭代,依赖于服务客户端 100 反馈是否用于识别一个或多个恰当的子类别,当前的类别级别在每个层次分类系统中向下移动零个、一个或多个级别。在某些实施例中,如果服务客户端 100 自己改变查询中的一个或多个关键词(即,不是由查询增强系统 120 建议的推荐改变),则查询可以被查询增强系统 120 作为新查询来对待。

[0082] 在方框 304,关键词预处理器 140 接收查询并且预处理关键词。如果关键词预处理器 140 识别出拼写错误或者停用词,则关键词预处理器 140 通知匹配引擎 150 忽略拼写错误的词或停用词,并且把该信息转发到服务客户端 100,以便校正查询。

[0083] 在方框 306,匹配引擎 150 接收预处理之后的查询并且确定该查询的客户端选定类别关于与每个类别的查询相关指标值是否排名高。在某些实施例中,确定是对于所有选定类别。如果一个或多个选定类别具有低关联查询相关指标值,则处理继续到方框 322(图 3C),以便找出推荐的关键词变化,以便消除查询的关键词与选定类别之间的失配。

[0084] 说明以上失配场景的一个例子是以下的例子:由客户端 100 提交的查询 Q 包含三个关键词:“保险(insurance)”、“小汽车(car)”和“驾驶员(driver)”。选定类别是图 2 中所绘出的类别 \_A2220。但是,关于与查询的查询相关指标,也在图 2 中绘出的类别 \_A1210 排名高,如以下所示出的。假设关键词与两个类别之间的关键词相关指标是:

[0085] 类别 \_A1:  $RI_{insurance} = 0.008$ ,  $RI_{car} = 0.025$ ,  $RI_{driver} = 0.030$

[0086] 类别 \_A2:  $RI_{insurance} = 0.020$ ,  $RI_{car} = 0.005$ ,  $RI_{driver} = 0.004$

[0087] 为了简化,假设等式(4)中项  $(\sum_{i=1}^{M_Q} RI_i^2 + \sum_{j=M_Q+1}^{M_C} RI_j^2)$  的值对于类别 \_A1 和类别 \_A2 都

是 0.01。计算出的查询 Q 与这两个类别的查询相关指标是:

$$[0088] \quad \text{类别\_A1: } RI_Q = \frac{0.008 + 0.025 + 0.030}{(3 \times 0.01)^{\frac{1}{2}}} = \mathbf{0.364}$$

$$[0089] \quad \text{类别\_A2: } RI_Q = \frac{0.020 + 0.005 + 0.004}{(3 \times 0.01)^{\frac{1}{2}}} = \mathbf{0.167}$$

[0090] 如果所有选定类别都具有高查询相关指标值,则处理继续到决定方框 308。

[0091] 在方框 308,匹配引擎 150 确定是否已经到达每个分类系统的最低类别级别。如果是,则处理继续到方框 310,否则,处理继续到方框 314,以在每个分类层次中向下穿过一级或多级(图 3B)。

[0092] 在某些实施例中,每个分类系统的每个类别的最低类别级别包含一个或多个个体服务。在方框 310,匹配引擎 150 给处于最低类别级别的个体服务排名。在方框 312,响应该查询,匹配引擎 150 把一个或多个高排名服务的信息提供给服务客户端 100。然后,处理完成。

[0093] 在方框 314,匹配引擎 150 从关键词数据存储 130 为每个选定类别的每个子类别提取(即,检索)查询中每个关键词的关键词相关指标( $RI_i$ )。在方框 314,提取可以包括提取数据并计算关键词相关指标。在方框 316,匹配引擎 150 利用关键词相关指标计算查询与那些子类别中每一个的查询相关指标( $RI_Q$ )。对于一个类别来说,查询相关指标描述该类别与一个查询相关多紧密(即,它们关于它们共同具有的关键词多相似)。

[0094] 在某些实施例中,一个分类系统中处于选定类别相同级别的其它类别不推荐给服务客户端 100 或者被其选择。但是,如果用于这些类别的查询相关指标达到某个阈值,显示这些类别可能也与该查询相关,则查询与这些类别的每个子类别的查询相关指标也要计算并排名。

[0095] 在方框 318,匹配引擎 150 基于每个分类系统中子类别的查询相关指标给它们排名。

[0096] 在某些实施例中,如果一个分类系统中某个子类别的查询相关指标值显著高于其它子类别的查询相关指标值,则在来自服务客户端 100 的确认之前匹配引擎 150 可以假设这个子类别是选定类别。在这种实施例中,匹配引擎 150 在层次中再向下移动当前类别级别一级,而且匹配引擎 150 计算并排名这个子类别的子类别的相关指标。这个过程可以重复,直到匹配引擎 150 到达一个点,在这个点,没有子类别具有显著高的查询相关指标值,而且需要来自服务客户端 100 的选择。

[0097] 在方框 320,匹配引擎 150 确定由服务客户端 100 指定或确认的每个选定类别是否包括至少一个排名高的子类别(基于方框 318 中的排名)。如果是,则处理继续到方框 330,否则,处理继续到方框 322。

[0098] 在方框 330,匹配引擎 150 把当前类别级别的子类别级别设置成当前类别级别。在方框 332,匹配引擎 150 把未改变的关键词与新排名的类别列表(在当前类别级别改变之前被称为子类别列表)发送到服务客户端 100,以使服务客户端 100 能够选择或验证新子类别。

[0099] 在方框 334,匹配引擎 150 确定是否从服务客户端 100 接收到具有指定的或确认的关键词与类别的查询。如果是,则过程从方框 334 (图 3D)循环回到方框 304 (图 3A),以执

行改进查询的另一次迭代,否则,处理继续到方框 336。在方框 336,匹配引擎 150 确定服务客户端 100 是否指示查询应当以当前形式还是以之前的形式执行,而不需要执行进一步的迭代(例如,经用户接口)。如果是,则处理继续到方框 338,否则,处理完成。在某些实施例中,在没有接收到反馈的一段时间后,匹配引擎 150 确定服务客户端 100 最好退出该过程,而不以当前形式或者以之前的形式执行查询。在这种情况下,匹配引擎 150 退出,而不执行查询。

[0100] 在方框 338,匹配引擎 150 以当前形式或者以之前的形式执行查询。在某些实施例中,服务客户端 100 能够识别要执行的特定之前形式。在方框 340,匹配引擎 150 基于它们的查询相关指标给服务排名。在方框 342,匹配引擎 150 把一个或多个高排名服务的列表提供给服务客户端 100(例如,经用户接口)。即,匹配引擎 150 提供关于所述一个或多个服务的信息。然后,处理完成(即,由服务客户端 100 退出)。在某些实施例中,来自该列表的一个或多个服务可以由服务客户端 100 选择消费(即,提供选定的服务)。

[0101] 如果存在不具有高排名子类别的一个或多个选定类别,则处理从方框 320 继续到方框 322。这是由服务客户端 100 指定的关键词与属于相关类别或子类别的关键词之间的失配造成的。在方框 322,匹配引擎 150 识别类别关键词中可以用于替换查询中匹配查询关键词的查询关键词的同义词。特别地,匹配引擎 150 识别对非排名靠前的选定类别具有低关键词相关指标值的查询关键词,以便匹配非排名靠前的类别中可以用于代替查询中关键词的查询关键词。非排名靠前的选定类别可以被描述为其查询相关指标值不在最高的当中,或者其子类别的查询相关指标值中没有一个是最高当中的选定类别。

[0102] 为了识别同义词,匹配引擎 150 从关键词数据存储 130 提取属于查询中非排名靠前的选定类别的关键词的关键词相关指标值。然后,匹配引擎 150 定位在非排名靠前的类别中但不在查询中具有高关键词相关指标值的类别关键词,而且反之亦然(即,定位对非排名靠前的类别具有低关键词相关指标值的查询关键词)。匹配引擎 150 还从关键词词库 160 提取同义词并且识别这两组关键词(即,查询关键词与类别关键词)之间的任何同义词。如果识别出一对同义词,每个组有一个(即,一个来自查询关键词,一个来自类别关键词),则产生建议服务客户端 100 用属于非排名靠前的类别的同义词代替(即,替换)查询中的关键词的推荐。例如,假设非排名靠前的选定类别是类别 \_A2220。关键词“汽车”在类别 \_A2220 中具有高关键词相关指标值,但是在查询 Q 中没有。相反,关键词“小汽车”在查询 Q 中,但是在类别 \_A2220 中具有低关键词相关指标值。关键词词库 160 指示汽车与小汽车是同义词。因此,产生建议服务客户端 100 用关键词“汽车”代替查询中的关键词“小汽车”的推荐。

[0103] 在方框 324,匹配引擎 150 识别非排名靠前的类别中可以添加到查询的新关键词。特别地,对于非排名靠前的类别中具有高关键词相关指标值的关键词,如果它们不在查询中而且在方框 322 没有找到它们的同义词,则产生建议服务客户端 100 在查询中添加这些新关键词的推荐。在某些实施例中,当服务客户端 100 的选择与匹配引擎 150 的排名之间发生类别失配时,推荐从方框 324 提供。作为一个例子,关键词“报价”可以被识别为这种关键词而且为服务客户端 100 产生把关键词“报价”添加到查询 Q 中的推荐。

[0104] 在方框 326,匹配引擎 150 识别作为要从查询中除去的候选的查询关键词。特别地,匹配引擎 150 检查查询中的关键词并且定位对非排名靠前的类别具有低关键词相关指

标值并且对排名靠前但未被服务客户端 100 选择的类别具有高关键词相关指标值的关键词。如果在方框 322 没有为这些关键词在非排名靠前的类别中找到同义词,则产生建议服务客户端 100 除去查询中这些关键词的推荐。作为一个例子,关键词“驾驶员”可以被识别为这种关键词并且为服务客户端 100 产生从查询 Q 除去关键词“驾驶员”的推荐。

[0105] 在方框 328,匹配引擎 150 向服务客户端 100 提供关键词改变推荐和排名后的类别列表(例如,经用户接口)。然后,服务客户端 100 可以通过选择要在查询中使用的关键词和要与该查询关联的类别来提供反馈。例如,假设服务客户端 100 采纳用“汽车”代替“小汽车”、添加关键词“报价”和除去关键词“驾驶员”的变化推荐。还假设关键词“汽车”与类别 \_A1210 和类别 \_A2220 的关键词相关指标分别是 0.006 和 0.026,而关键词“报价”的关键词相关指标分别是 0.004 和 0.032。于是,在关键词改变之后,查询 Q 与这两个类别的查询相关指标是:

$$[0106] \quad \text{类别 } \_A1: RI_Q = \frac{0.008 + 0.006 + 0.004}{(3 \times 0.01)^{\frac{1}{2}}} = \mathbf{0.104}$$

$$[0107] \quad \text{类别 } \_A2: RI_Q = \frac{0.020 + 0.026 + 0.032}{(3 \times 0.01)^{\frac{1}{2}}} = \mathbf{0.450}$$

[0108] 在改变之后,非排名靠前的类别类别 \_A2220 具有比类别 \_A1210 更高的相关指标。类别 \_A2220 变成被匹配引擎 150 认为是排名靠前的类别。服务客户端 100 与匹配引擎 150 之间的类别选择失配在采纳查询关键词变化推荐之后校正。

[0109] 从方框 328,处理继续到方框 334 (图 3D)。

[0110] 如果当前的类别级别不在层次性分类系统中,而是在群集或群集层次中,则该过程是类似的。区别在于,在某些实施例中,群集类别可能没有有意义的名字。于是,匹配引擎 150 需要选择类别中的一种典型服务来代表这个类别并且把该服务信息提供给服务客户端 100。通过指定所提供的服务是否类似于服务客户端 100 正在寻找的服务,服务客户端 100 决定是否选择该类别。

[0111] 当过程到达分类或群集层次最精细的类别时,当前的类别不具有子类别,但是它们包含个体服务。每个服务都可以作为一个子类别来对待而且也将计算并排名它们的查询相关指标。在这种实施例中,服务客户端 100 将接收从其进行选择的排名后的服务列表,而不是类别列表。

[0112] 给定包括一个或多个关键词的用户查询,实施例提供了在服务储存库中对相关条目的有效搜索与检索。除了开发关键词 - 查询 - 类别匹配方案,实施例还考虑到服务储存库中服务的关键词之间已有的层次关系,这导致用于精细化查询以便提高查询准确性与质量的有效技术。

[0113] 实施例提供了一种信息检索技术,这种技术帮助用户基于迭代用户反馈改进依靠具有层次性分类系统的数据存储发布的查询。

[0114] 对于这些实施例,用户不需要为了指定查询关键词和执行有效的关键词搜索而首先具有关于什么服务元数据以及它们如何存储在服务注册表中和 / 或这些服务所使用的确切关键词的具体了解。

[0115] 所属技术领域的技术人员知道,本发明的各个方面可以实现为系统、方法或计算



机程序产品。因此,本发明的各个方面可以采取完全硬件实施例、完全软件实施例(包括固件、驻留软件、微代码等)或硬件和软件方面结合的实施例的方式,这里可以统称为“电路”、“模块”或“系统”。此外,在一些实施例中,本发明的各个方面还可以采取在一个或多个计算机可读介质中体现的计算机程序产品的形式,该计算机可读介质中包含计算机可读的程序代码。

[0116] 可以采用一个或多个计算机可读介质的任意组合。计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质。计算机可读存储介质可以是例如,但不限于,电、磁、光、电磁、红外线、或半导体的系统、装置或设备,或者以上所述的任意组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括:具有一根或多根导线的电连接、便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦可编程只读存储器(EPROM或闪存存储器)、光纤、便携式光盘只读存储器(CD-ROM)、光存储设备、磁存储设备、固态存储器、磁带或者以上所述的任意合适组合。在本文档的背景下,计算机可读存储介质可以是包含或存储程序的任何有形介质,该程序可以被指令执行系统、装置或设备使用或者与其结合使用。

[0117] 计算机可读信号介质可以包括其中,例如,在基带中或者作为载波的一部分,体现计算机可读程序代码的传播数据信号。这种传播信号可以采取多种形式中的任何一种,包括但不限于电-磁、光或者其任意合适组合。计算机可读信号介质可以是非计算机可读存储介质而且可以传送、传播或者运输由指令执行系统、装置或设备使用或者与其联合使用的程序的任何计算机可读介质。

[0118] 计算机可读介质上包含的程序代码可以用任何适当的介质传输,包括,但不限于,无线、有线、光缆、RF等等,或者以上所述的任意合适组合。

[0119] 可以以一种或多种程序设计语言的任意组合来编写用于执行本发明操作的计算机程序代码,所述程序设计语言包括面向对象的程序设计语言—诸如Java、Smalltalk、C++等,还包括常规的过程式程序设计语言—诸如“C”语言或类似的设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中,远程计算机可以通过任意种类的网络——包括局域网(LAN)或广域网(WAN)—连接到用户计算机,或者,可以连接到外部计算机(例如,通过利用互联网服务提供商的互联网)。

[0120] 下面将参照根据本发明实施例的方法、装置(系统)和计算机程序产品的流程图说明和/或框图来描述本发明的实施例的各个方面。应当理解,流程图说明和/或框图的每个方框以及流程图说明和/或框图中各方框的组合,都可以由计算机程序指令实现。这些计算机程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器,从而生产出一种机器,使得这些计算机程序指令在通过计算机或其它可编程数据处理装置的处理器执行时产生实现流程图和/或框图中的一个或多个方框中规定的功能/动作的器件。

[0121] 也可以把这些计算机程序指令存储在计算机可读介质中,这些指令会指示计算机、其它可编程数据处理装置、或其他设备以特定方式工作,使得存储在计算机可读介质中的指令产生出包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的指令

的制造品(article of manufacture)。

[0122] 计算机程序指令也可以加载到计算机、其它可编程数据处理装置或者其它设备上,使得一系列操作处理(例如,操作或步骤)在计算机、其它可编程装置或者其它设备上执行,产生一种计算机实现的过程,使得在所述计算机或者其它可编程装置上执行的指令提供用于实现在所述流程图和 / 或框图的一个或多个方框中规定的功能 / 动作的过程。

[0123] 实现所述操作的代码可以进一步在硬件逻辑或电路系统(例如,集成电路芯片、可编程门阵列(PGA)、专用集成电路(ASIC)等)中实现。硬件逻辑可以耦合到处理器,来执行操作。

[0124] 关键词预处理器 140 和 / 或关键词分类 - 群集匹配引擎 150 (“匹配引擎 150”)可以实现为硬件(例如,硬件逻辑或电路系统)、软件或者硬件与软件的组合。

[0125] 在某些实施例中,匹配引擎 150 具有自己的处理器和存储器。

[0126] 图 4 说明了根据某些实施例可以使用的计算机体系结构 400。服务客户端 100 和 / 或服务注册服务器 110 可以实现计算机体系结构 400。计算机体系结构 400 适于存储和 / 或执行程序代码并且包括至少一个直接地或者通过系统总线 420 间接地耦合到存储器元件 404 的处理器 402。存储器元件 404 可以包括在程序代码实际执行过程中采用的本地存储器、大容量储存器和提供至少一些程序代码的临时存储以便减少执行过程中必须从大容量储存器检索代码的次数的快速缓存存储器。存储器元件 404 包括操作系统 405 和一个或多个计算机程序 406。

[0127] 输入 / 输出(I/O)设备 412、414 (包括但不限于键盘、显示器、定点设备等)可以直接地或者通过中间 I/O 控制器 410 间接地耦合到系统。

[0128] 网络适配器 408 也可以耦合到系统,以使得数据处理系统能够变得通过中间的私有或公共网络耦合到其它数据处理系统或者远端打印机或存储设备。调制解调器、电缆调制解调器和以太网卡仅仅是一些目前可用类型的网络适配器 408。

[0129] 计算机体系结构 400 可以耦合到储存装置 416 (例如,非易失性存储区域,例如磁盘驱动器、光盘驱动器、磁带驱动器等)。储存装置 416 可以包括内部存储设备或者附连或网络可访问的储存装置。储存装置 416 中的计算机程序 406 可以按本领域中已知的方式加载到存储器元件 404 中并且被处理器 402 执行。

[0130] 计算机体系结构 400 可以包括比所说明更少的组件、未在此说明的附加组件、或者所说明的组件与附加组件的某种组合。计算机体系结构 400 可以包括本领域中已知的任何计算设备,例如大型机、服务器、个人计算机、工作站、膝上型计算机、手持式计算机、电话设备、网络用具、虚拟化设备、存储控制器、个人数字助理(PDA)、平板计算机、口袋式个人计算机(PC)、网络 PC、手持式设备、机顶盒、消费者电子产品、微型计算机、超级计算机等。

[0131] 附图中的流程图和框图说明了根据本发明各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段或代码的一部分,所述模块、程序段或代码部分包含一条或多条用于实现规定的逻辑功能的可执行指令。还应当指出,在有些可供选择的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,依赖于所涉及的功能性,两个连续示出的方框实际上可以基本并行地执行,或者它们有时也可以按相反的顺序执行。还应当指出,框图和 / 或流程图说明中的每个方框、以及框图和 / 或流程图说明中方框的组合,可以由执

行规定功能或动作的基于硬件的专用系统或者专用硬件与计算机指令的组合来实现。

[0132] 本文所使用的术语仅仅是为了描述特定的实施例而不是要作为本发明的限制。如本文所使用的,除非上下文明确地另外指出,否则单数形式“一个”和“这个”要也包括复数形式。还应当理解,当在本说明书使用时,术语“包括”和/或“包含”规定了所述特征、整数、步骤、操作、元素和/或部件的存在,但是并不排除一个或多个其它特征、整数、步骤、操作、元素、部件和/或其组的存在或添加。

[0133] 以下权利要求中所有方法或步骤加功能元素的对应结构、材料、动作及等价物都是要包括用于结合具体所述的其它所述元素执行所述功能的任何结构、材料或行为。已经为了说明和描述的目的给出了本发明实施例的描述,但这不是详尽的或者要把本发明限定到所公开的形式。在不背离本发明范围与主旨的情况下,许多修改和变化对本领域普通技术人员都将是显而易见的。实施例的选择和描述是为了最好地解释本发明的原理和实践应用,并使本领域其他普通技术人员能够理解本发明具有适于预期特定使用的各种修改的各种实施例。

[0134] 本发明实施例的以上描述是为了说明和描述的目的而给出的。它不是详尽的或者要把实施例限定到所公开的精确形式。根据以上教导,许多修改与变化都是可能的。实施例的范围不是要由具体描述而是要由所附权利要求来限定。以上说明书、例子和数据提供了对实施例成分的制造与使用的完整描述。由于在不背离本发明主旨与范围的情况下可以产生许多实施例,因此实施例在于以下所附的权利要求或者任何后续提交的权利要求及其等价物中。

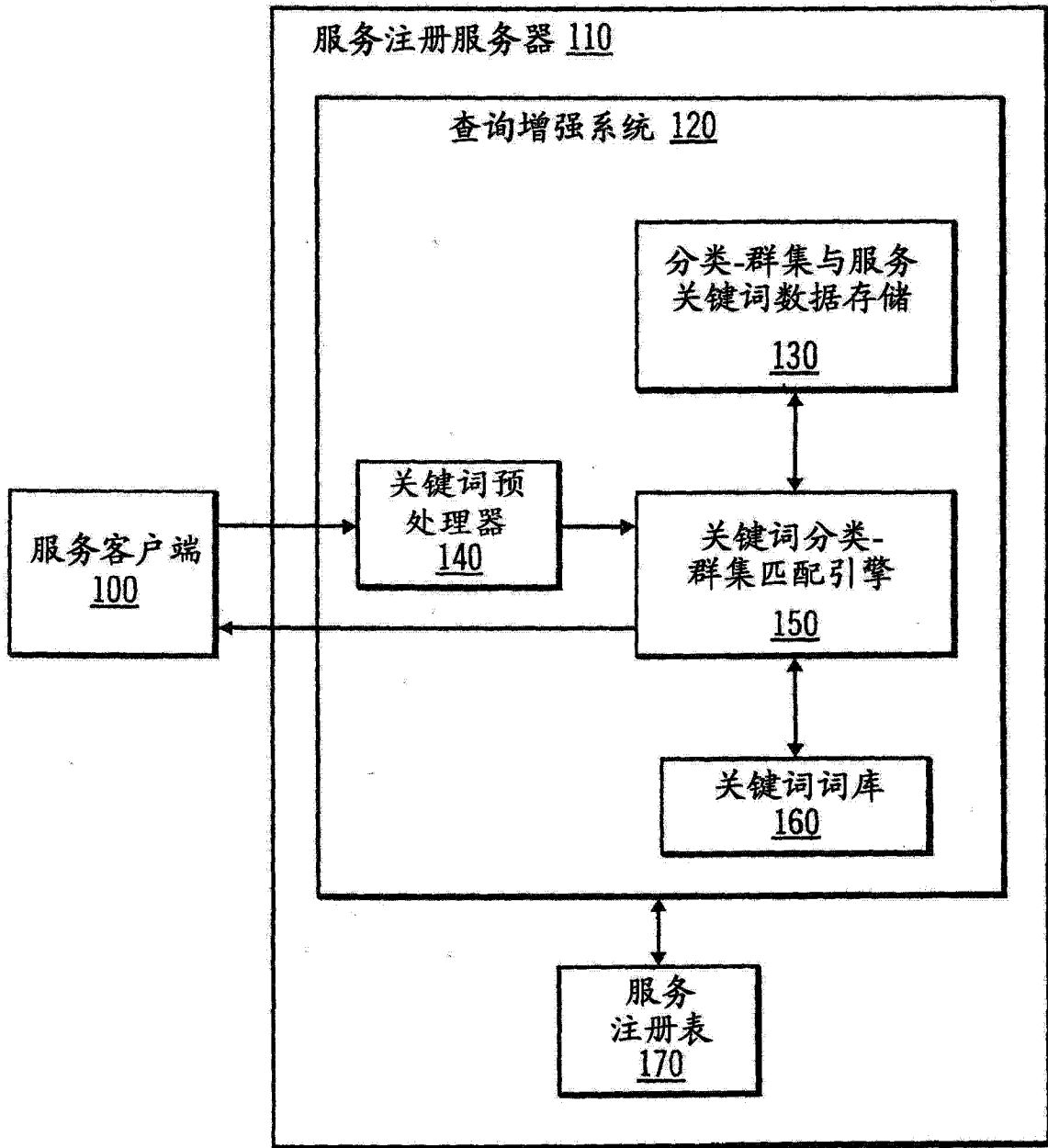


图 1

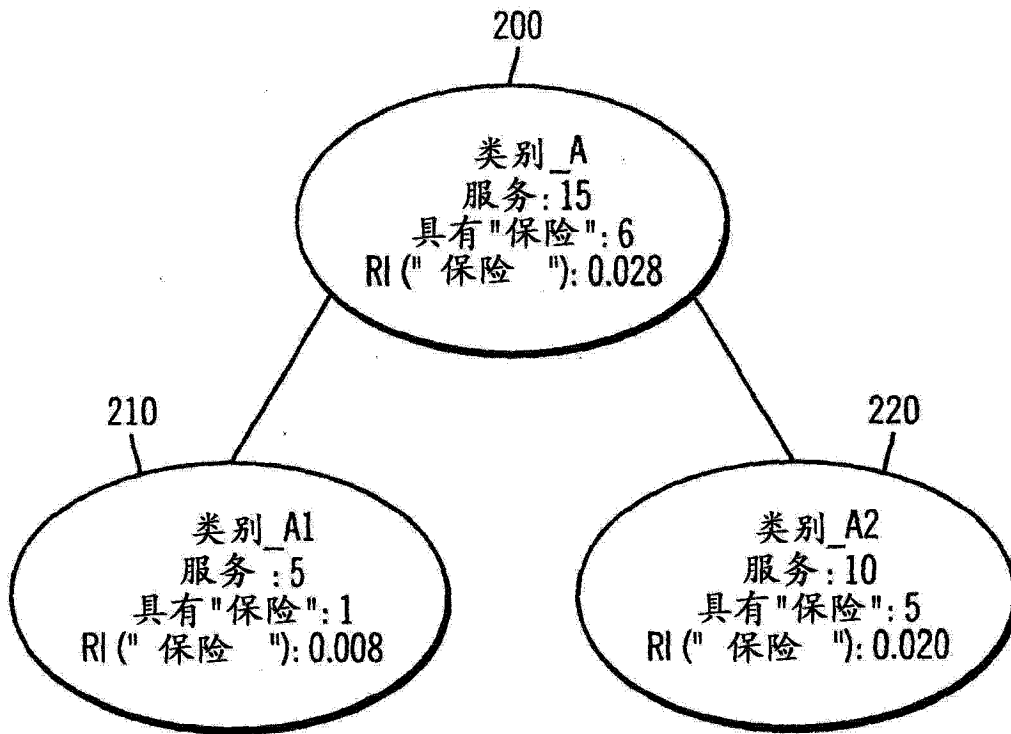


图 2

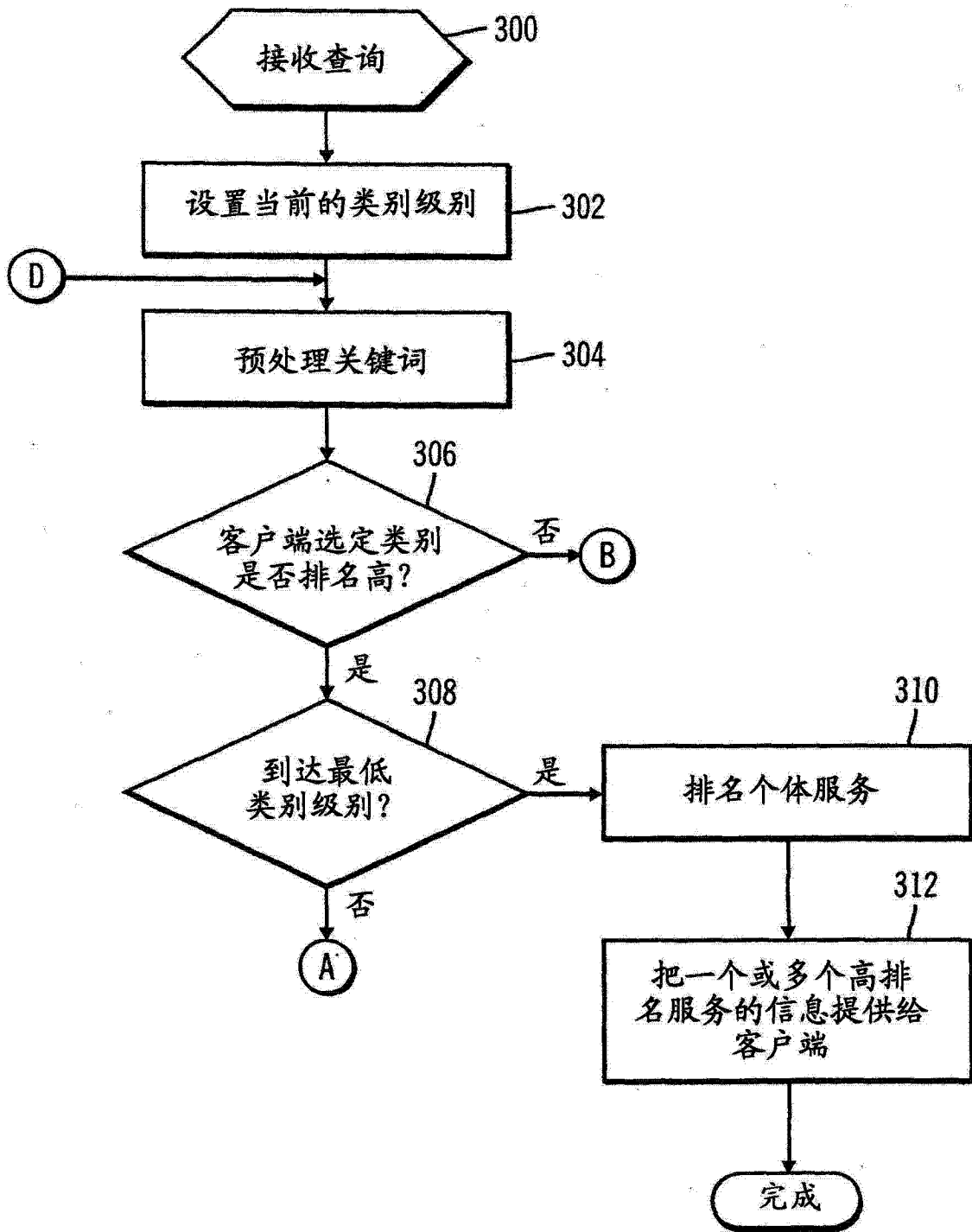


图 3A

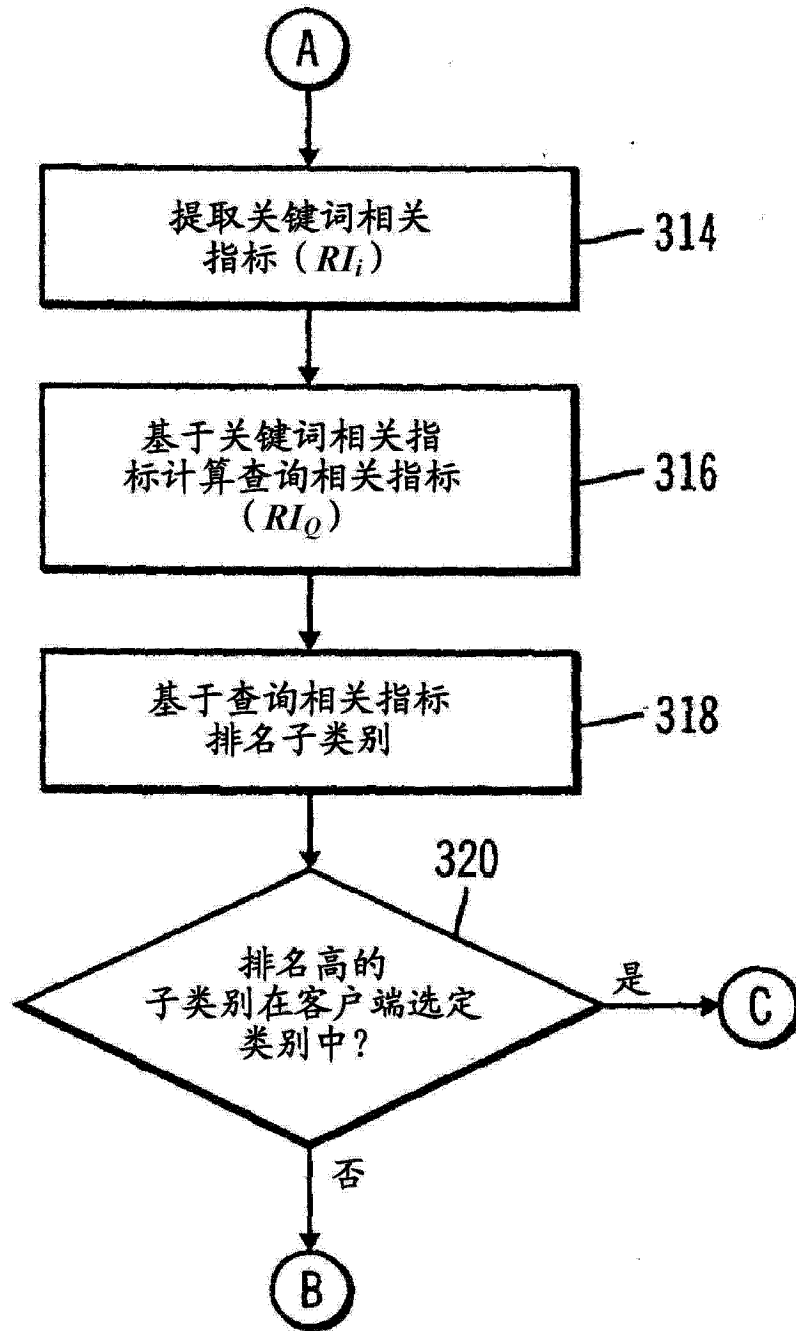


图 3B

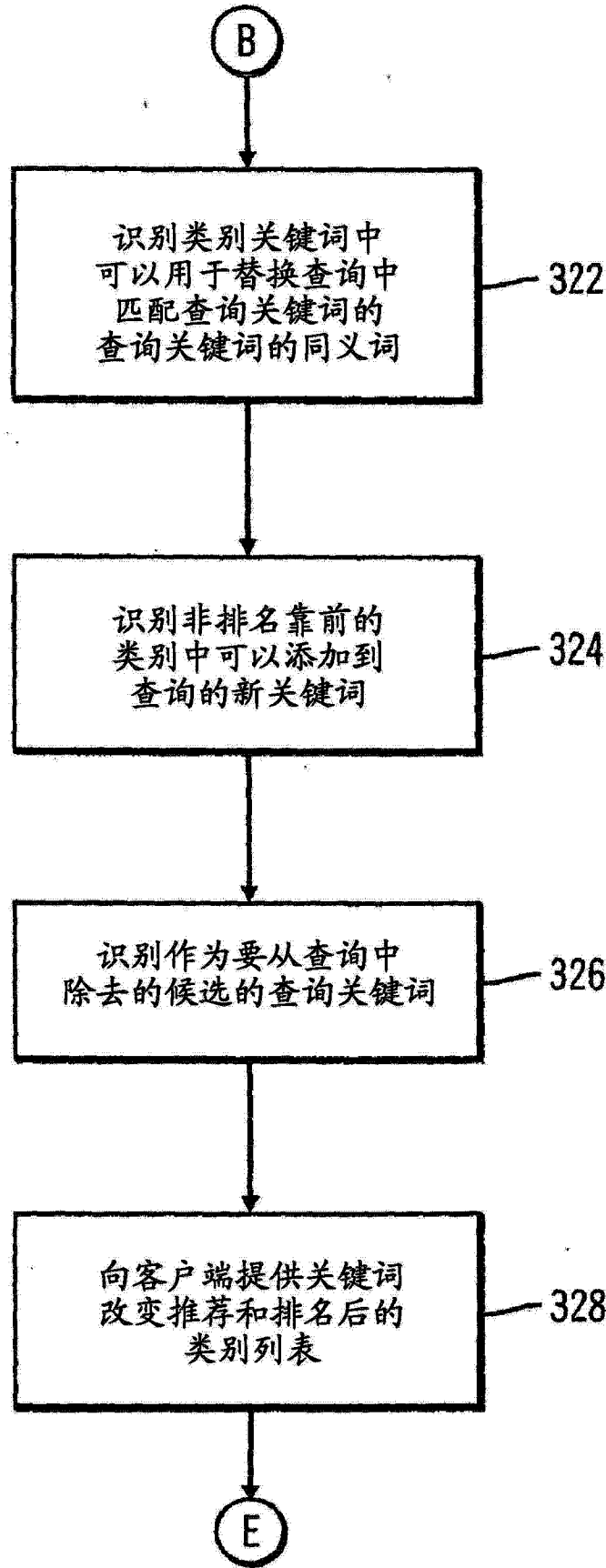


图 3C



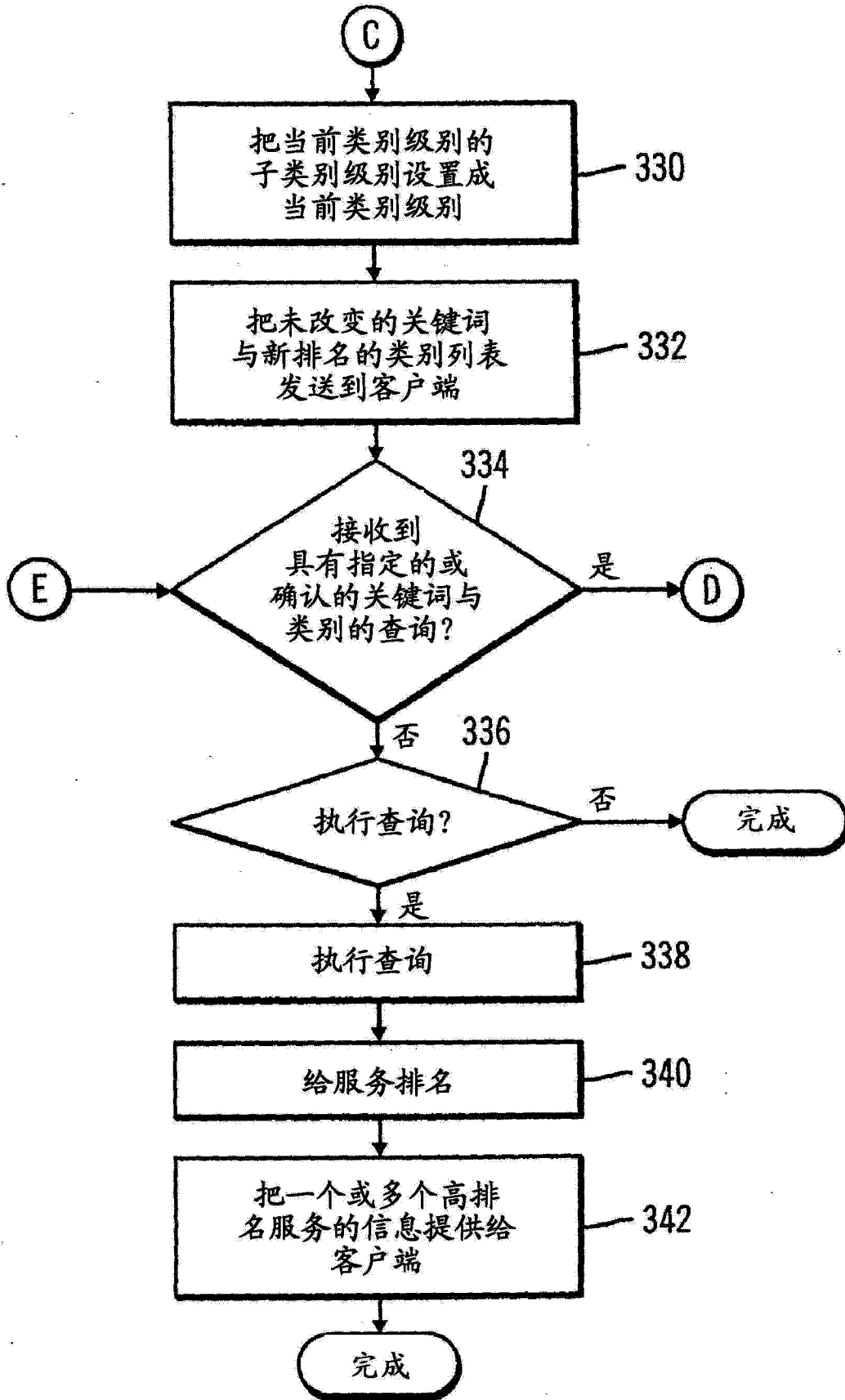


图 3D

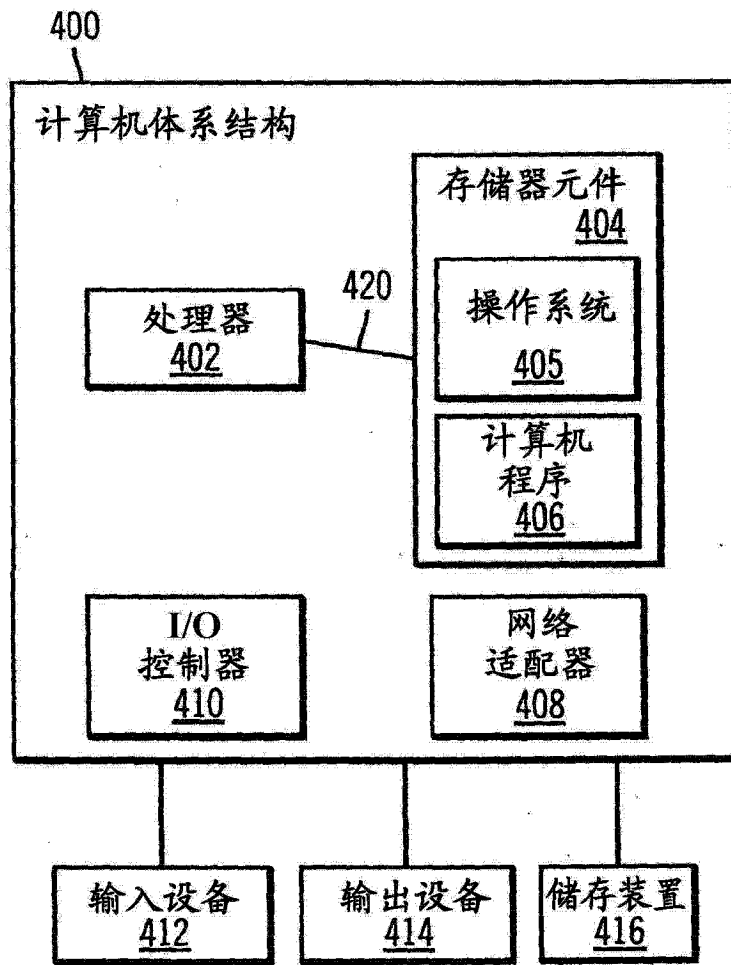


图 4