[19] 中华人民共和国国家知识产权局

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

[12] 发明专利说明书

专利号 ZL 200510071690.0

[45] 授权公告日 2009 年 7 月 8 日

[11] 授权公告号 CN 100511224C

[22] 申请日 2005.4.13
[21] 申请号 200510071690.0
[30] 优先权
[32] 2004.4.15 [33] US [31] 10/826,161
[73] 专利权人 微软公司
地址 美国华盛顿州
[72] 发明人 B·张 D·B·库克 G·希施勒
洪小文 H-J·曾 K·弗里斯
K·塞缪尔森 马维英 陈 正
[56] 参考文献
CN1170908A 1998.1.21
WO97/49048A1 1997.12.24
US5845278A 1998.12.1
WO99/48028A3 1999.9.23

Image retrieval by hypertext links. V. Harmandas, M. Sanderson, M. D. Dunlop. Annual ACM Conference on Research and Development in Information Retrieval, Proceedings of the 20th annual international ACM SIGIR conference on Research and development In information retrieval. 1997
审查员 杨 薇
[74] 专利代理机构 上海专利商标事务所有限公司
代理人 李 玲

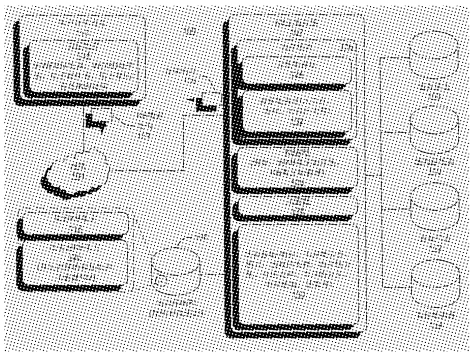
权利要求书 6 页 说明书 26 页 附图 10 页

[54] 发明名称

用于改进文档检索的内容传播的方法和计算设备

[57] 摘要

描述了为改进的文档检索提供内容传播的系统和方法。在一个方面中，识别针对一个或多个文档的参考信息。所述参考信息是从一个或多个数据源中识别出来的，所述一个或多个数据源与包括所述一个或多个文档的数据源无关。从一个或多个数据源中提取被接近地定位到所述参考信息的元数据。对于所述一个或多个文档中相关文档的内容，计算所述元数据的各个特征之间的相关性。对于所述一个或多个文档的每个文档，将所述元数据的相关部分用各自部分的所述特征相关性索引到所述文档的原始内容。所述索引产生了一个或多个改进文档。



1、一种为改进的文档检索提供计算机可执行的内容传播的方法，所述方法包括：

识别针对一个或多个文档的参考信息，所述参考信息是从一个或多个数据源中识别出来的，所述一个或多个数据源与包括所述一个或多个文档的数据源无关；

提取被接近地定位到所述参考信息的元数据，其中所述元数据接近所述参考信息并且在语义上和/或上下文上涉及所述参考信息；

对于所述一个或多个文档中相关文档的内容，计算所述元数据的各个特征之间的相关性；

对于所述一个或多个文档的每个文档，将所述元数据的相关部分用各自部分的所述特征相关性索引到所述文档的原始内容；

其中，所述索引步骤产生了一个或多个改进的文档；

基于搜索查询分析所述一个或多个改进的文档以定位相关信息；

基于相关性分数对所述一个或多个改进的文档进行排序；以及

基于所述搜索查询，发送排序的结果以及对于所述一个或多个改进的文档的片断描述；

其中，所述一个或多个数据源包括一搜索查询日志，并且，计算相关性进一步包括：

从所述搜索查询日志中识别出搜索查询，其中，为了搜索所述数据源，所述搜索查询具有相对高的出现频率FOO；

确定一终端用户从搜索查询结果中选择的条文，所述条文来自所述数据源；以及

确定遗漏的终端用户选择，其中一个遗漏的终端用户选择就是所述搜索查询结果中一个没有选择的条文；

其中，确定遗漏的终端用户选择进一步包括利用层内连接分类不同类的对象，以便为所述不同类对象确定重要性测量，所述不同类对象包括第一类相似查询以及第二类相关文档，所述相似查询已经在所述搜索查询日志中被识别出，所述相似查询与包括所述一个或多个文档的搜索结果相关，不管终端用户是否

在所述搜索结果中选择了所述相关文档的各个文档，所述相关文档都从所述搜索结果中识别出。

2、如权利要求 1 所述的方法，其中，所述参考信息包括一个连接和/或实际上唯一的文档 ID，所述文档 ID 与所述一个或多个文档中的一个文档相关。

3、如权利要求 1 所述的方法，其中，所述一个或多个文档是知识库条文、产品帮助、任务和/或开发者数据。

4、如权利要求 1 所述的方法，其中，所述一个或多个数据源包括服务请求和/或新闻组投递。

5、如权利要求 1 所述的方法，其中，所述元数据在语义上或上下文上涉及所述一个或多个文档中相关的文档。

6、如权利要求 1 所述的方法，其中，所述元数据包括文档标题、产品问题上下文和/或产品问题解决信息。

7、如权利要求 1 所述的方法，其中，对于所述一个或多个改进文档的每个改进文档，都具有产生所述改进文档的相应原始文档。

8、如权利要求 1 所述的方法，其中，计算所述相关性基于所述一个或多个文档的一个特定文档在所述元数据的上下文中被识别出的次数。

9、如权利要求 1 所述的方法，其中，所述元数据包括条文标题、产品问题上下文和/或产品问题解决信息，并且计算相关性进一步包括加权所述条文标题和/或产品问题上下文，以便指示比任何产品问题解决信息大的相关性。

10、如权利要求 1 所述的方法，其中，计算相关性进一步包括给在所述数据源中出现频率较大的元数据的特征赋予较大的相关性，所述频率较大是相对于所述内容中的其他元数据特征的出现频率的。

11、如权利要求 1 所述的方法，其中，计算相关性进一步包括给在一个或多个文档中的一个文档中找到的所述元数据的特征赋予较大的相关性以用作所述文档的寿命的函数。

12、如权利要求 1 所述的方法，其中，所述特征用所述第一和第二类中的各个节点表示，并且每个所述节点的所述重要性测量都基于一个相似性函数，该相似性函数测量了在第一和第二类的对象之间的距离。

13、一种为改进的文档检索提供内容传播的计算设备，所述计算设备包括：

用于识别针对一个或多个文档的参考信息的装置，所述参考信息是从一个

或多个数据源中识别出来的，所述一个或多个数据源与包括所述一个或多个文档的数据源无关；

用于提取被接近地定位到所述参考信息的元数据的装置，其中所述元数据接近所述参考信息并且在语义上和/或上下文上涉及所述参考信息；

用于对于所述一个或多个文档中相关文档的内容，计算所述元数据的各个特征之间的相关性的装置；

用于对于所述一个或多个文档的每个文档，将所述元数据的相关部分用各自部分的所述特征相关性索引到所述文档的原始内容的装置；

其中，所述索引产生了一个或多个改进文档；

用于基于搜索查询分析所述一个或多个改进的文档以定位相关信息的装置；

用于基于相关性分数对所述一个或多个改进的文档进行排序的装置；以及

用于基于所述搜索查询，发送排序的结果以及对于所述一个或多个改进的文档的片断描述的装置；

其中，所述一个或多个数据源包括一搜索查询日志，并且，用于计算相关性的装置进一步包括：

用于从所述搜索查询日志中识别出搜索查询的装置，其中，为了搜索所述数据源，所述搜索查询具有相对高的出现频率 FOO；

用于确定一终端用户从搜索查询结果中选择的条文的装置，所述条文来自所述数据源；以及

用于确定遗漏的终端用户选择的装置，其中一个遗漏的终端用户选择就是所述搜索查询结果中一个没有选择的条文；

其中，用于确定遗漏的终端用户选择的装置进一步包括：用于用层内连接分类不同类的对象，以便为所述不同类对象确定重要性测量的装置，所述不同类对象包括第一类相似查询以及第二类相关文档，所述相似查询已经在所述搜索查询日志中被识别出，所述相似查询与包括所述一个或多个文档的搜索结果相关，不管终端用户是否在所述搜索结果中选择了所述相关文档的几个文档，所述相关文档都从所述搜索结果中被识别出。

14、如权利要求 13 所述的计算设备，其中，所述参考信息包括一个连接和/或实际上唯一的文档 ID，所述文档 ID 与所述一个或多个文档中的一个文档相

关。

15、如权利要求13所述的计算设备，其中，所述一个或多个文档是知识库条文、产品帮助、任务和/或开发者数据。

16、如权利要求13所述的计算设备，其中，所述一个或多个数据源包括服务请求和/或新闻组投递。

17、如权利要求13所述的计算设备，其中，所述元数据在语义或上下文上涉及所述一个或多个文档中相关的文档。

18、如权利要求13所述的计算设备，其中，所述元数据包括文档标题、产品问题上下文和/或产品问题解决信息。

19、如权利要求13所述的计算设备，其中，对于所述一个或多个改进文档的每个改进文档，都具有产生所述改进文档的相应原始文档。

20、如权利要求13所述的计算设备，其中，计算所述相关性基于所述一个或多个文档的一个特定文档在所述元数据的上下文中被识别出的次数。

21、如权利要求13所述的计算设备，其中，所述元数据包括条文标题、产品问题上下文和/或产品问题解决信息，并且用于计算相关性的装置进一步包括用于加权所述条文标题和/或产品问题上下文，以便指示比任何产品问题解决信息大的相关性的装置。

22、如权利要求13所述的计算设备，其中，用于计算相关性的装置进一步包括用于给在所述数据源中出现频率较大的元数据的特征赋予较大的相关性的装置，所述频率较大是相对于所述内容中的其他元数据特征的出现频率的。

23、如权利要求13所述的计算设备，其中，用于计算相关性的装置进一步包括用于给在一个或多个文档中的一个文档中找到的所述元数据的特征赋予较大的相关性以用作所述文档的寿命的函数的装置。

24、如权利要求13所述的计算设备，其中，所述特征用所述第一和第二类中的各个节点表示，并且每个所述节点的所述重要性测量都基于一个相似性函数，该相似性函数测量了在第一和第二类的对象之间的距离。

25、一种为改进的文档检索提供内容传播的计算设备，所述计算设备包括：

识别部件，其识别针对一个或多个文档的参考信息，所述参考信息是从一个或多个数据源中识别出来的，所述一个或多个数据源与包括所述一个或多个文档的数据源无关；

提取部件，其提取被接近地定位到所述参考信息的元数据，其中所述元数据接近所述参考信息并且在语义上和/或上下文上涉及所述参考信息；

计算部件，其对于所述一个或多个文档中相关文档的内容，计算所述元数据的各个特征之间的相关性；

索引部件，其对于所述一个或多个文档的每个文档，将所述元数据的相关部分用各自部分的所述特征相关性索引到所述文档的原始内容；

其中，所述索引产生了一个或多个改进文档；

分析部件，其基于搜索查询分析所述一个或多个改进的文档以定位相关信息；

排序部件，其基于相关分数对所述一个或多个改进的文档进行排序；以及

发送部件，其基于所述搜索查询，发送排序的结果以及对于所述一个或多个改进的文档的片断描述；

其中，所述计算部件进一步包括分类部件，其利用层内连接分类不同类的对象，以便为所述不同类对象确定重要性测量，所述不同类对象包括第一类相似查询以及第二类相关文档，所述相似查询已经在搜索查询日志中被识别出，所述相似查询与包括所述一个或多个文档的搜索结果相关，不管终端用户是否在所述搜索结果中选择了所述相关文档的各个文档，所述相关文档都从所述搜索结果中被识别出。

26、如权利要求25所示的计算设备，其中，所述参考信息包括一个连接和/或实际上唯一的文档ID，所述文档ID与所述一个或多个文档中的一个文档相关。

27、如权利要求25所示的计算设备，其中，所述一个或多个文档是知识库条文、产品帮助、任务和/或开发者数据。

28、如权利要求25所示的计算设备，其中，所述一个或多个数据源包括服务请求、新闻组投递和/或搜索查询日志。

29、如权利要求25所示的计算设备，其中，所述元数据在语义上或上下文上涉及所述一个或多个文档中相关的文档。

30、如权利要求25所示的计算设备，其中，所述元数据包括条文标题、产品问题上下文和/或产品问题解决信息，并且用于计算相关性的所述计算部件进一步包括加权部件，其加权所述条文标题和/或产品问题上下文，以便指示比任

何产品问题解决信息大的相关性。

31、如权利要求25所示的计算设备，其中，用于计算相关性的所述计算部件进一步包括赋值部件，其给在所述数据源中出现频率较大的元数据的特征赋予较大的相关性，所述频率较大是相对于所述内容中的其他元数据特征的出现频率的。

32、如权利要求25所示的计算设备，其中，用于计算相关性的所述计算部件进一步包括赋值部件，所述赋值部件给在一个或多个文档中的一个文档中找到的所述元数据的特征赋予较大的相关性以用作所述文档的寿命的函数。

33、如权利要求25所示的计算设备，其中，所述一个或多个数据源包括一搜索查询日志，并且用于计算相关性的所述计算部件进一步包括：

标识部件，其从所述搜索查询日志中识别出搜索查询，其中，为了搜索所述数据源，所述搜索查询具有相对高的出现频率FOO；

确定部件，其确定一终端用户从搜索查询结果中选择的条文，所述条文来自所述数据源；以及

计算遗漏的终端用户选择的部件，其中一个遗漏的终端用户选择就是所述搜索查询结果中一个没有选择的条文。

用于改进文档检索的内容传播的方法和计算设备

相关申请

本申请涉及下列专利申请，其中每个申请通常都被指定为本申请的受让人并在此引用结合：

美国专利申请 no.10/826,159，其名称为“用于搜索项建议的多类型数据对象的补充分类”，04 年 4 月 15 日申请；以及

美国专利申请 no.10/427,548，其名称为“利用内层联系（Inter-Layer Links）的对象分类”，2003 年 5 月 1 日申请。

技术领域

本发明的实施属于数据挖掘。

背景技术

如今的高科技公司通常都会提供某些方面的产品支持，以确保使消费者和合伙人能获得技术投资的最大值。例如，可提供各种消费和商业支持提供（offering）以及战略 IT 咨询服务，以满足消费者和合伙人的需求。支持提供可包括电话、站点、基于 Web 的支持等等。但不幸的是，产品支持服务不仅在财务成本方面特别昂贵，而且在寻求解决方案所花费的时间方面也是特别昂贵的。例如，网络咨询服务通常会很昂贵，以至于非公司消费者没有能力进行单个产品咨询或故障诊断。

另外，当自动服务时，例如在线搜索包括产品“如何”（帮助）和故障诊断条文的知识库时，消费者识别相关条文所需的时间就是不能令人接收的。一个原因就是知识库产品故障诊断条文通常是由专业作者、供货商等等提供的，而不是由寻求支持的产品用户提供的。在这种情况下，如果用户不利用知识库（KB）创建者所采用的项目形成搜索查询，用户在定位相关知识库故障诊断信息上时非常困难的并要花费大量时间。

发明内容

这里描述了提供计算机执行的改进文档检索的内容传播的系统和方法。在一方面，识别出针对一个或多个文档的参考信息。该参考信息是从一个或多个

数据源中标识出的，该一个或多个数据源与包括一个或多个文档的数据源无关。元数据被接近地定位到参考信息，该元数据是从一个或多个数据源中提取的。计算所述元数据的相应特征之间的相关性，所述元数据组成任何一个所述一个或多个文档。对于所述一个或多个文档中的每个文档，利用特征的相关性该元数据的相关部分从各自部分被索引到文档的原始内容。该索引生成了一个或多个改进的文档。

附图说明

在附图中，部件附图标记最左边的数字表示该部件首次出现特定附图。

图 1 示出了为改进的文档检索提供内容传播的示例性系统。

图 2 示出了为改进的文档检索提供内容传播的典型过程。

图 3 表示了一个适合的计算环境，在该环境中可以完全或者部分执行随后将描述的为改进的文档检索提供内容传播的系统、装置和方法。

图 4 是可用于分类的计算机环境的一个实施例的方框图。

图 5 是用于将不同类对象分类的体系的一个实施例的方框图。

图 6 是混合网络模型的一个实施例的方框图。

图 7 是直接与 Internet 相连的计算机环境的另一个实施例的方框图。

图 8 是分类算法的一个实施例的流程图。

图 9 是分类算法的另一个实施例的流程图。

图 10 是用于将包含隐蔽层的不同类对象分类的体系的另一个实施例的方框图。

图 11 是分类算法的另一个实施例的流程图。

具体实施方式

概述

为了解决产品问题（故障）和/或另外的搜索一产品，要创建 KB 条文来辅助消费者定位“如何”(帮助)条文。研究表明终端用户越容易搜索到并获得直接定位消费者查询的在线 KB 条文，消费者对该产品及其相关支持设施就越满意。因此，下面描述的系统和方法就通过分析存储在多种数据源中的信息来定位相关信息的 KB 条文（KBARI），而提供内容传播和改进的文档检索。这样的数据源例如包括：请求存储库、在线产品和开发者支持组新闻组投递（posting）、搜索查询点击通过日志，和/或等等。

KBARI 包括，例如，实际上唯一的 PS 文档（例如，KB 条文）ID，到特定 PS 条文的超文本链接、特定 PS 条文的通用资源标识符（URI）、文档标题等等。当在产品服务请求中发现 KBARI，和/或来自产品开发者到的投递支持新闻组时，临近 KBARI 的文本包括在语义和/或上下文上对由 KBARIPS 参考的故障诊断条文有意义的信息，这是可能的。另外，这样的文本很可能是在实时问题解决情况下由终端用户和/或产品支持服务发（PSS）引擎所产生的——而不仅仅是由专业作者或者具有证明产品任务的卖主产生。

例如，服务请求的 PSS 日志中的服务请求（SR）是一个存档文档（例如一个或多个相关联的电子邮件），这些存档文档包括最初由终端用户提交给 PSS 引擎的信息。也就是说，SR 将涉及例如故障诊断情况的问题的产品引导到 PSS 引擎。PSS 引擎长生一个 SR 汇总，以便清楚标识下列信息的一些合并：产品、定位的问题、故障现象（例如，行为和结果）、原因和/或解决方案。结果，SR 包括很可能包含对 KB 条文 106 有实际意义的参考的数据，和/或与在实时问题解决情况下由终端用户和 PSS 引擎产生的信息相关的产品。

对于新闻组投递（posting），机构和公司通常具有与新闻组有关的产品和/或开发者，以便向终端用户提供在线讨论产品开发和故障诊断提交的机会。例如，如果一个终端用户遇到一个特殊产品的一个问题，用户就可以向服务器投递一个相应的条文，该条文标识了问题和请求援助。在这种情况下，新闻组阅读者，其一般可包括从业者或者与该产品有关的服务专业人员，就可以向请求者投递一个答复。在具有请求时，新闻组投递可以包括与一个或多个 KB 条文直接或者由前后关系的内容（例如，一个连接，参考等等）。当投递参考一个 KB 条文时，该投递就提供了实质上对 KB 条文有用的元数据。

对于请求日志，终端用户经常提交对一搜索引擎的搜索查询，例如，通过一 Web 站点，查找与特殊产品有关的 KB 条文，对产品行为进行故障诊断，等等。一台服务器，例如一台宿主了搜索引擎和/或 KB 数据库的服务器，记录了终端用户请求以及任何随后的终端用户点击通过（click一thru）行为。如果一查询经常涉及一个 KB 条文，那么次查询就特别可能是该 KB 条文的适宜元数据。

为了调节这些来自多个数据源的与语义和/或上下文有关的信息，该系统和方法提取到定位的 KBARI（例如，接近）的近似文本。分析提取的文本，以便产生关于相关 PS 条文的特征（关键词）重要性加权值。（提取的文本与 PS 条文

有关，而该PS条文由与KBARI相似的文本所指示)。提取的文本（下面一般称为“原数据”）和相应的特征重要性加权值由相关PS条文的原始内容索引，以便产生新的或改进的PS条文。在这种实施中，原始和改进的PS条文之间是一对一对应。例如，对于每个改进的PS条文，都有一个相应的未改进的或者原始PS条文。在另一个实施中，不具有这样的一对一对应，并且原始PS条文可以用改进的PS条文替换。

响应于从一个终端用户接收到搜索查询，为改进的文档检索提供内容传播的系统和方法检索任何的PS条文（原始和/改进的），该PS条文包括搜索查询项。结合查询项相似性（proximity）和通用标准，确定检索到的原始和/或改进的PS条文的相关性。随后根据相关性分数，将搜索结果排序。从所搜结果中产生摘录说明，以便清楚的向终端用户指示返回文档的相关性。排序结果与摘录说明一起传输给终端用户。

在一个实施中，为改进的文档检索提供内容传播的系统和方法也有助于为自动PS条文的产生标识新的PS内容。现在将更详细描述为改进的文档检索提供内容传播的系统和方法的这些和其它方面。

一个典型系统

在附图中，相似的附图标记涉及相似的元件，该系统和方法将被描述，并且实施在所示的适当的计算环境中。尽管不需要，但是总的来说，该系统和方法也将在由个人计算机执行的计算机可执行指令的环境中描述，例如，程序模块。程序模块通常包括例程，程序、对象、组件、数据结构等等，它们执行特殊的任务或者特殊的抽象数据类型。虽然在前述的环境中描述该系统和方法，但是此后描述的动作和操作也可以在硬件中实施。

图1示出了为改进的文档检索提供内容传播的典型系统100。在此实施中，系统100包括KB主服务器102和客户计算设备116，该主服务器102通过网络104与KB条文106（一个数据库）数据源108－114相连。网络104可包括局域网（LAN）和通用广域网（WAN）通信环境的任意组合，例如，办公室、企业内计算机网络、内联网和Internet内的常用通信环境。KB主服务器102挖掘存储在数据源108－114中的信息，并将这些信息添加到KB条文118中，以便产生新的或改进的KB条文120。在此实施中，数据源108－114包括，例如，服务请求108、新闻组投递110、查询日志112和/或其它数据源114。响应于从

一个客户计算设备 116 的终端用户那里接收到一个关于 KB 的搜索查询 122，KB 主服务器 102 检索原始 KB 条文 118 和/或包括搜索查询 122 的项目的改进的 KB 条文 120。客户计算设备 116 是任何类型的计算设备，例如，个人计算机、便携式电脑、服务器、移动计算设备（例如，移动电话、个人数字助理或掌上电脑）等等。

从多个数据源挖掘 PS 文档连接、ID 等等

更具体地，KB 主服务器 102 的元数据提取 124 挖掘存储在数据源 108-114 中的信息，以便识别出与各自的 KB 条文有关的信息。为了讨论和说明，这些识别的信息信息被称为 KB 条文相关信息（KBARI）126。KBARI126 包括，例如，实际上唯一的 KB 条文 ID（例如，GUID）、到特殊 KB 条文的超文本链接、特殊 KB 条文的通用资源标识符（URI），等等。当元数据提取 124 在来自一 PSS 的服务请求 108 中和/或来自以产品开发者支持新闻组的新闻组投递 110 中定位到 IBARI126 时，与 KBARI126 相似的文本很可能包括在语义上和/或在上下文上对由 KBARI126 参考的原始 KB 条文 118 有意义的信息。例如，这样的信息可包括条文标题、条文关键字、产品问题描述以及解决数据，等等。另外，这样的文本很可能是在实时问题解决的情况下，由终端用户和/或 PSS 引擎所产生的——而不仅仅是由专业作者或者具有证明产品任务的卖主产生。

特征提取和重要性加权

为了调剂来自数据源 108-114 的与语义和/或上下文有关的 KB 条文信息，元数据提取 124 提取到定位到的 KBARI126（例如，接近）的近似文本。为了描述方便，此提取的文本被表示为元数据 128。为了向终端用户提供实际上与搜索查询 122 的项目最相关的 KB 条文 106，元数据提取 124 分析元数据 128，以便产生与一个相关 KB 条文 106 的特征重要性（相关性）加权值。（由相应 KBARI126 指示提取的元数据 128 与 KB 条文 106 相关）。

更特别地，元数据提取 124 使用全文搜索技术来给元数据特征分配不同的相关性加权。在此实施中，并且对于服务请求 108，与分配给其它服务请求信息相比，例如，问题解决方案，标题和故障现象将赋予更大的加权值。这是由于用户更倾向于用故障现象表述一个搜索查询，而不是用问题解决方案信息表述。特征加权也可以反应特殊 KB 条文参考在其上下文中被识别出的次数，其是参考寿命（age）的一个功能。这样的特征加权也可用于新闻组投递 110。

对于从查询日志中 112 提取的元数据 128，元数据提取 124 进行特征分析以及通过识别出下列信息的某些合并而初次加权：（a）终端用户搜索 KB 条文 106 而多次产生的搜索查询，（b）随后选择的 KB 条文 106，和/或（c）与所选的 KB 相关的任何其它 KB 条文 106。然后，通过产生相似查询的分类（查询分类）和相关 KB 条文 106 的分类（即，条文分类），元数据提取 124 定位与（a）、（b）和/或（c）相联系的少量点击通过数据。如果用户选择搜索引擎返回的少数几个（例如，一个或多个）文档，通常会产生少量点击通过数据。为了进行说明，查询分类和条文分类也被各自表示为“其它数据”130 的一部分。在下面的附录 A 中详细描述了一个用于表述相似查询和相关 KB 条文分类的典型分类技术，其名称为“不同类对象的典型分类”。

为了继续并管理元数据 128 的加权特征，元数据提取和分析模块 124 用相关原始 KB 条文 118 的原始内容索引元数据 128 和相应的特征重要性加权值，以便产生新的或改进的 KB 条文 120。（元数据 128 包括从一个或多个数据源 108-114 中挖掘的数据，而该数据源 108-114 被确定补充了一个或多个各个 KB 条文 106）。在此实施中，标记元数据 128 的加权特征，这样，例如 XML 的置标语言就可以用于参考和检索索引的内容。在一个实施中，元数据 128 在改进的 KB 条文 120 中被索引为一个逆索引。在此实施中，改进的 KB 条文 120 和原始 KB 条文 118 之间一对一对应。例如，对于每个改进的 KB 条文，都有一个相应的未改进或者原始 KB 条文 118。此一对一对应意味着原始 KB 条文 118 的至少一个子集将具有一个相应的改进 KB 条文 120。在另一个实施中，不具有这样的一对一对应关系。例如，原始 KB 条文可以被改进的 KB 条文 120 替换。

改进的 KB 条文检索

搜索提供器 132 从客户计算设备 116 的终端用户那里接收到一个关于 KB 的搜索查询 122。搜索查询 122 的项目涉及产品调查和故障诊断询问。在一个实施中，搜索查询 122 包括用扩展标记语言（XML）指定的信息。终端用户使用任何不同的可用应用程序 134 通过网络 104 向 KB 主服务器 102 发送搜索查询 122。应用程序 130 包括，例如，Web 浏览器、字处理器、电子邮件和/或其它类型的计算机编程应用程序。

在此实施中，搜索提供器 132 提供了一个到 KB 主服务器 102 的远程应用入口点和搜索引擎功能。该入口点允许在 KB 服务器 102 和应用程序 134 的任

何不同可用的结构实施之间通信。例如，在一个实施中，入口点支持来自一个实施为 Web 浏览器的应用程序 134 的超文本传输协议（HTTP）命令。在另一个实施中，入口点支持基于例如简单对象访问协议（Simple Object Access Protocol SOAP）的消息协议的 XML。另一个入口点实施也可能是应用程序 134 和搜索提供器 132 之间所需的通信支持的特殊类型的功能。

响应于接收到搜索查询 122，搜索提供器 132 根据一模式分析并执行搜索查询 122 的数据格式，该模式被各自表示为“其它数据” 130 的一部分。在一个实施中，该模式也可以例如被客户计算设备 116 上载到 KB 主服务器 102 中。接着搜索提供器 132 进行在 KB 条文 106 中进行全文搜索，以便识别和检索到相关/有关的原始 KB 条文 118 和/或改进的 KB 条文 120。为了描述和讨论，如此检索到的文档被各自表示为“其它数据” 130 的一部分。

检索到的文档相关性和排列操作

然后根据查询项相似性和通用标准，确定检索到的文档的相关性。对于项目相似性，搜索 KB 条文 106 的搜索查询 122 的长度可以比其它类型的查询（例如，一般 Web 搜索的查询模型）要长。这是由于描述产品故障诊断和/或研究问题一般要使用更多的词/项。据此，为了在可包括多个项的查询中定位到一个覆盖尽可能多查询项片段的 KB 条文 106，搜索提供器 132 使用项目相似性加权搜索查询 122 中的项。相似性值通过一条如下的曲线被转换成用于全文检索模块输出的相似值的加权因数：

$$Sim = Sim_{orig} * proximity,$$

$$\text{proximity} = \frac{\log(1+\alpha(\beta * Hit + (1-\beta) * (1 - EditDis\tan ce)))}{\log(1-\alpha)}$$

其中α, β用于控制搜索查询 122 的各个部分的相关加权的参数。参数 *Hit* 表示在一个文档中出现搜索查询 122 的项占搜索查询 122 的所有项的百分比，参数 EditDistance 是查询和文档之间“混乱”（misorder）的度，对于“混乱”术语，认为是例如，一个查询包括下列关键词：“信息检索和数据挖掘”，而文档是“检索信息并从数据中挖掘”。这个例子中的关键词“信息”、“检索”、“数据”和“挖掘”就是混乱的。为了解决这个问题，当计算查询和相应的文档的相似性时，我们对混乱的关键词提出了一种处理。为了讨论方便，项目相似和近似值被表示为“其它数据” 130 中“相关分数”。

搜索提供器 132 根据基于相关分数的查询项相似性排列检索到的文档。在一个实施中，通过确定标识的 KB 条文 106 的寿命（age），以及由于较新的条文 106 比较旧的 KB 条文“更流行”，给较新的条文 106 赋予较大的加权，实现排列。在另一种实施中，其中，KB 条文 106 的普及性实际上很难确定标识的 KB 条文 106 的普及性，其被确定为该条文出现的次数的函数，该条文由服务请求 108 和/或新闻组投递 110 参考。一个条文被参考的次数越大，与没有被参考那么多次的条文相比，该条文就越普及并且其排序就越高。对于新闻组投递 110，KB 条文 106 的普及性是条文参考频率的函数，和/或由新闻组中新闻投递者普遍性确定的——特定用户投递的越多，用户普遍性就越大。

小参考频率的相对新的 KB 条文表示相对小的普及性。但是，新条文对终端用户来说非常重要。因此，在一个实施中，搜索提供器 132 合并参考频率和次数的因数，并根据如下公式将具有不同次数的 KB 条文的普及性：

$$popularity = \frac{\log(1+\alpha(\beta * I_{ref} + (1-\beta) * (1-I_{age})))}{\log(1+\alpha)}$$

I_{ref} 表示根据参考频率的重要性（参考频率越高，其重要性值就越大）。I_{age} 表示根据释放时间（条文寿命）的重要性。参数 α 和 β 表示根据参考频率的重要性和根据释放时间的重要性之间的相对加权，其可以由先前的知识指定和/或在处理输出中学习。越新的 KB 条文 126，其计算的条文重要性就越高。

$I_{ref}=0.5+0.5^{freq(ref)}/maxfreq(ref)$

$$I_{age}=\frac{1}{1+e^{age}}$$

用搜索查询 122 的项搜索 KB 条文 106 的搜索结果被排列，或者被相关认为是计算的重要性值的一个函数，每个结果都被各自表示为“其它数据”130 的一部分。

搜索结果片段产生/高亮显示

在一个实施中，并且为了实际上最小化显示给终端用户的与查询有关的信息，搜索提供器 132 产生最高排行的一个和多个检索到的文档的片断描述，以便清楚地向用户指示检索到的与搜索查询 122 的项有关的文档相关性（例如，向终端用户清楚指示标识的材料（条文）如何）。为了描述，片断描述被各自表示为“其它数据”130 的一部分。为了产生片段描述，搜索提供器 132 为片段描述，从被确定为与搜索查询 122 相关的检索到 KB 条文 106 中定位一个或多个

块，然后高亮显示一个或多个块中的搜索查询122的任何项目。搜索提供器132用具有可调尺寸的滑动窗口标识一个或多个块，该滑动窗口用于检索到的文档的各部分。在一个实施中，滑动窗口的尺寸是UI空间的函数，该UI空间对于在客户计算设备116上显示片段描述是可靠的。

对于检索到的文档的一部分的滑动窗口的每种应用程序，搜索提供器132测量由滑动窗口文本描述记载的涉及查询的信息的数量。此测量被各自表示为“其它数据”130的一部分。该测量包括基于定量标准的值，例如字频率、相对改进的查询项的字相似性、字位置等等。搜索提供器132使用学习分类模块（参见“其它数据”130）合并这些不同标准，以便得到片段描述的最丰富信息块。在这种方式中，片段描述清楚地向终端用户表示了标识的KB条文106的相关性。

该学习分类模型通过线性回归学习，该线性回归是统计学中的一种经典学习方法。线性回归旨在用适合该数据的直线解释矢量x和值y之间的关系。线性回归公设了：

$y = b_0 + \sum_{j=1}^{p} b_j x_j + e$ 其中，“残差”e是平均值为零的随即变量。系数b_j由尽可能小的残差的平方的总和的情况确定。变量x_j直接来自输入或输入的某种转换，例如日志或多项式。

搜索提供器132将最高排行的检索文档的至少一个子集与其相应的片段描述一起打包成查询响应136。为了显示和终端用户解决产品研究和/或故障诊断调查所使用，，搜索提供器132将查询响应136传输到客户计算设备116。

一个典型过程

图2示出了用于改进的文档检索的内容传播的典型过程200。为了进行讨论，针对图1的部件讨论该过程的操作。（所以的附图标记数字以首次介绍部件的图号打头）。在方框202，元数据提取124（图1）从多个各自数据源108-114中标识出与特殊KB条文106相关的信息 — 基于知识的条文相关信息（KBARI126）。在方框204，元数据提取124提取与方框202中标识出的信息相似（proximity）的特征。在方框206，元数据提取124分析提取的特征（元数据128），以便产生相关基于知识的条文106的相应条文的相关重要性测量。在方框208，元数据提取124将提取的特征与相应的相关性分数一起索引到相关的单个基于知识的条文106。这就产生了新的或改进的基于知识的条文120。

在方框210，搜索提供器132，响应于接收到搜索查询122，检索包括搜索查询122的项目的原始KB条文118和/或改进的KB条文120。在方框212，搜索提供器132根据搜索查询122的项到各自的文档/条文的相关性分数，排列检索到的文档/条文。在方框214，搜索提供器132为检索到的基于知识的条文106产生片段说明。在方框216，搜索提供器132将排序的结果和片段描述传输给终端用户。

典型操作环境

图3表示了适合的计算环境300的一个例子，在该环境中可以完全或者部分执行为改进的文档检索提供内容传播的图1的系统100和图2的方法。图3表示了适合的计算环境300的一个例子，在该环境中可以执行（完全或者部分）为改进的文档检索提供内容传播的系统、装置和方法。典型计算环境300仅仅是可用的计算环境的一个例子，其并不意于限制这里所描述的系统和方法的使用范围或功能。该计算环境300也不应当被解释为依靠或需要计算环境300中所示的任何一个部件或其组合。

这里描述的方法和系统可使用更多的其它目和特定用途的计算系统环境或配置运行。可用的公知计算系统、环境和/或配置的例子包括，但不局限于，个人计算机、服务器计算机、多处理器计算机、基于微处理器的系统、网络PC、微型计算机、大型计算机、可包括任意上述系统或设备的分布式计算环境，等等。在限制资源的客户机中也可实施压缩或子集版的构架，例如，掌上电脑或其它计算设备。本发明一般实施在分布式计算环境中，该分布式计算环境中，任务由通过通信网络连接的远程处理设备执行任务。在分布式计算环境中，程序模块可位于本地和远程存储设备中。

参照图3，为改进文档检索提供内容传播的典型系统包括计算机310形式的通用计算设备。计算机310的下述方面是客户计算设备116（图1）和KB主服务器102（图1）的的典型实施。计算机310的部件可包括，但不局限于，处理器单元320。系统存储器330和系统总线321，该系统总线321将包括将系统存储器的各种系统部件连接到处理器单元320。系统总线321可以是任何类型的总线结构，其包括使用任何总线结构的存储器总线或存储器控制器、外围总线和局域总线。举例来说，但不限制，这样的结构包括工业标准结构（ISA）总线、微通道结构（MCA）总线、扩展ISA（EISA）总线、视频电子标准协会（VESA）

局域总线以及也被称为 Mezzanine 总线的外设部件互连（PCI）总线。

计算机 310 一般包括各种计算机可读介质。计算机可读介质可以是任何可由计算机 310 访问的介质，并包括易失和非易失介质、可移动和不可移动介质。举例来说，但不限制，计算机可读介质可包括计算机存储介质和传播介质。。计算机存储介质包括易失和非易失介质、可移动和不可移动介质，为了存储例如计算机可读指令、数据结构、程序模块或其它数据的信息，其可以在任何方法或技术中运行。计算机存储介质包括，但不局限于，RAM、ROM、EEPROM、闪存或其它存储技术、CD-ROM、数字通用盘（digital versatile disk DVD）或其它光盘存储，盒式磁带、磁带、磁盘存储器或其它类型的磁存储设备、或其它任何可用于存储需要信息并可由计算机 310 访问的介质。

传播介质一般是计算机可读介质、数据结构、程序模块或调制的数据信号中的其它数据，例如，载波或其它传送机制，并包括认可传递介质的任何信息。词语“调制的数据信号”是指具有一个或多个特征集的信号或者在信号中将信息编码而改变的信息。举例来说，但不限制，传播介质包括有线介质，例如，有线网络或直接有线连接，并且包括无线介质，例如音频、射频、红外和其它无线介质。任何上述介质的结合可应当包含在计算机可读介质的范围内。

系统存储器 330 包括易失和/或非易失存储器的计算机存储器介质，例如只读存储器（ROM）331 和随机存取存储器（RAM）333。基本输入/输出系统 333（BIOS）包含有助于在例如启动时在计算机 310 内的元件件传输信息的基本程序，基本输入/输出系统 333 一般存储在 ROM331 中。RAM332 一般包含处理器单元 320 立即访问和/或当前操作的数据和/或程序。举例来说，但不限制，图 3 示出了操作系统 334、应用程序 335、其它程序模块 36 和程序数据。在一个实施中，结合参考图 1，计算机 310 示一个 KB 主服务器 102。在这种情况下，应用程序 335 包括图 1 的程序模块 138，并且程序数据 337 包括图 1 的 KB 条文相关信息（KBARI）126、元数据 128 和/或“其它数据”130。

计算机 310 也可包括其它可移动/不可移动、易失/非易失计算机存储介质。仅仅举例来说，图 3 示出了一个硬盘驱动器 341、磁盘驱动器 351 和光盘驱动器 355，该硬盘驱动器 341 读取或写入不可移动、非易失磁介质，磁盘驱动器 351 读取或写入可移动、非易失磁盘 352 ，光盘驱动器 355 读取或写入可移动、非易失光盘 356，例如 CD-ROM 或其它光学介质。其它可用在典型操作环境中的

可移动/不可移动、易失/非易失计算机存储介质包括，但不局限于，盒式磁带、闪存卡、数字通用盘、数字视频磁带、固态 RAM、固态 ROM 等等。硬盘驱动器 341 一般通过例如接口 340 的不可移动存储器接口与系统总线 321 连接，磁盘驱动器 351 和光盘驱动器 355 一般通过例如接口 350 的可移动存储器接口与系统总线 321 相连。

上述以及图 3 所示的驱动器及其相关的计算机存储介质为计算机 310 存储了计算机可读指令、数据结构、程序模块和其它数据。在图 3 中，例如，硬盘驱动器 341 被表示为存储了操作系统 344、应用程序 345、其它程序模块 346 以及程序数据 347。注意，这些部件可以与操作系统 334、应用程序 335、其它程序模块 336 和程序数据 337 相同，也可以与它们不同。这里给了操作系统 344、应用程序 345、其它程序模块 346 和程序数据 347 不同的数字，是为了表示它们至少是不同的复本。

用户可通过输入设备将命令和信息输入到计算机 310 中，例如，键盘 362 和指示设备 361，其一般指鼠标、跟踪球或触摸板。其它输入设备（未示出）也可包括麦克风、操纵杆、游戏垫、圆盘式卫星电视天线、扫描仪等等。这些和其它输入设备一般通过与系统总线 321 连接的用户输入接口 360 连接到处理单元 320，但是也可以通过其它接口和总线结构连接，例如，并行端口、游戏端口或通用串行总线（USB）。

监视器 391 或其它类型的显示设备也可以通过例如视频接口 390 的接口和系统总线 321 连接。除监视器外，计算机还可以包括其它外围输出设备，例如，扬声器 397 和打印机，其可以通过外设输出接口 395 连接。

计算机 310 利用与一个或多个例如远程计算机 380 的远程计算机的逻辑连接，在网络环境中运行。远程计算机 380 可以是一台个人计算机、服务器、路由器、网络 PC、同级设备（peer device）或其它公共网络节点，并且尽管在图 3 中仅仅示出了存储器设备 381，但是其通常包括上述计算机 310 的所有元件。图 3 描述的逻辑连接包括局域网（LAN）371 和广域网（WAN）373，但是也可以包括其它网络。此网络环境一般用在办公室、企业内计算机网络、内联网和 Internet 中。

当用在 LAN 网络环境中时，计算机 310 通过网络接口和适配器 370 连接到 LAN371。当用在 WAN 中时，计算机 310 一般包括一个调制解调器 372 或用于

在例如 Internet 的 WAN373 上建立通信的其它装置。调制解调器 372 可以是内置的或外置的，其可以通过用户输入接口 360 或其它适当的机制连接到系统总线 321。在网络环境中，相对 310 描述的程序模块或其它部分可存储在远程存储器设备中。举例来说，但并不限制，图 3 示出了驻留在存储器 381 上的远程应用程序读 385。所示的网络连接是典型的，可以使用在计算机间建立通信连接的其它手段。

结论

尽管用限定了结构特性和/或方法或动作的语言描述了为改进文档检索的提供内容传播的系统和方法，但是应当理解，在所附权利要求中限定的实施不需要限制到描述的特殊特征和动作。例如，尽管在改建 KB 条文 106 检索的数据源内容传播方面描述了图 1 的系统 100，但是所描述的系统和方法也可用于将从一个或多个独立数据源挖掘的数据，传播到任何类型的参考文档，并不局限于 KB 或产品支持条文。例如，对于其它类型的文档，系统 100 可用于为改进的文档检索在数据源中提供内容传播，该数据源包括连接、参考、标题、文档 ID 等等。因此，文字说明的特殊特征和动作的公开只是所要求保护的主题的典型实施方式。

附录A

不同类对象的分类

典型分类系统和方法的背景

分类包括将多个对象分组，并以用于如搜索引擎和数据挖掘的应用程序中。分类算法根据对象相似性将对象分组。例如，Web 页面对象的分类是根据其内容、连接结构或者其用户的访问日志的。用户的分类是根据用户所选择的项目的。用户对象的分类是根据用户访问的历史的。与用户有关的项目的分类通常是根据选择那些项目的用户的。多种分类算法都是公知的。已有的分类算法包括基于划分的分类、分层分类以及基于密度的分类。

用户访问的 Web 页面的内容或者访问模式经常用于构建用户简介，以便分类 Web 用户。然后使用传统的分类技术。在合作过滤中，为了更好地建议/预测，分类也用于将用户或项目分组。

一般，这些在先分类算法的使用具有某些局限性。传统的分类技术可具有数据稀缺的问题，其中对象的数量或不同类对象之间连接的数量也很少，这样就不能有效地分类对象。对于同类分类，分析的数据组包含相同类型的对象。例如，如果基于 Web 页面和用户进行同类分类，那么 Web 页面对象和用户对象每个都将分别分类。如果基于项目和用户进行同类分类，那么项目对象和用户对象每个都将分别分类。在此同类分类的实施例中，相同类型的对象被一起分类，而不考虑其他类型对象。

现有技术的不同类对象分类分别分类对象组。不同类对象分类仅仅将（flat）使用为表示每个对象节点的平特征。在现有的不同类分类中，所有的内部和层之间的连接结构都不考虑，或者将爱你跟其简单地作为独立的特征。

典型的分类系统和方法

图 4 示出了可有利于使用分类的计算机环境 400 的一个实施例（即通用计算机）。计算机环境 400 包括一个存储器 402、一个处理器 404、一个分类部分 408 以及支持电路 406。支持电路包括例如显示器和输入/输出电路部的设备，该输入/输出部允许计算机环境 400 的确定部件传输信息（即，数据对象）。

在分类部分 408 中执行分类。分类部分 408 可集成在计算机环境的存储器 402 和处理器 404 部分中。例如，处理器 404 处理分类不同对象的分类算法（器从存储器中获得）。存储器 402（例如数据库）负责存储分类的对象和相关程序

以及分类算法，这样如果需要就可以获得分类的对象。计算机缓建 400 可被配置为单机计算机、网络计算机系统、大型计算机或任何已知的各种计算机系统。在此公开的某些实施例描述了计算机环境应用（计算机从 Internet 下载 Web 页面）。可以想象，这里描述的概念用于任何已知类型的计算机环境 400。

此文字说明提供了一种分类机制，在这种机制下，提高了被认为是可靠的返回结果的百分比（即，有利于用户的查询）。分类可用于例如搜索工具、信息挖掘、数据挖掘、合作过滤等技术领域。搜索工具由于其服务不同信息需求的能力而具有接收注意信号，并且改进了检索性能。搜索工具与例如 Web 页面、用户、查询等计算机方面有关。

本文字说明表了各种用于分类数据对象的分类算法实施例。数据对象的分类是一种技术，通过该技术，数据对象的大组被分组为大量的数据对象组或类（数据对象每个类都具有较少的数据对象）。数据对象的分类组中的每个数据对象都具有某些相似性。因此分类的一个方面可以被认为是将多个数据对象分组。

此文字说明中描述的一个分类机制涉及一个体系图 550，图 5 中示出了体系图的一个实施例。提供了一体分类机制的某些实施例，其中，在不同等级或图 5 的体系图 550 中所示的节点组 P 和 U 之间分类不同类型对象。可以想象，此文字说明中描述概念可用于三或更多层，而不是在文字说明中描述的两层。每个 P 和 U 节点组都可以被认为是一层。在此文字说明中，术语“一体”分类用于一种分类不同类数据的技术。节点组 P 包括多个数字对象 P_1、P_2、P_3...，P_i，每个都具有近似的数字类型。节点组 U 包括多个数字对象 U_1、U_2、U_3...，U_j，每个都具有近似的数字类型。每个节点组（P 或 U）分类对象的数据类型是同样的，因此每个节点组（P 或 U）中数据对象都是同类的。节点组 P 中的数字对象 P_1、P_2、P_3...，P_i 的类型与节点组 U 中的数字对象 U_1、U_2、U_3...，U_j 的类型不同。这样，不同节点组 P 和 U 中的数据对象的类型是不同的，或者是不同类的。此文字说明的某些方面提供了利用对象的同类和不同类数据类型的输入（基于连接）而分类。

通过在数据对象对之间的连接延伸在此文字说明中说明了连接。连接表示分类中数据对象对之间的关系。在一个例子中，连接可以从一个 Web 页面对象延伸到一个用户对象，并表示了选择某一 Web 页面的用户。在另一个例子中，连接可以从一个 Web 页面对象延伸到另一个 Web 页面对象，并表示了不同 Web

页面见到关系。在分类的某一特定实施例中，“连接”是指“边缘”。广义的术语“连接”在本文字说明中用于描述连接、边缘和描述对象间关系的一个对象到另一个对象的任何联系。

连接具有各种不同的类型（如在本文字说明中所描述的），其中该连接涉及分类在体系图 550 中提出的不同对象的不同类型。连接可以被分类为层间连接或层内连接。层内连接 503 或 505 是体系图 505 中的连接的一个实施，其描述了相同类型的不同对象之间的关系。层级关系 504 是体系图 505 中的连接的一个实施，其描述了不同类型的对象之间的关系。如图 5 所示，在某些特定数据对象 U_1、U_2、U_3...，U_j 之间延伸有多个层内连接 503。在图 5 所示的实施例中，在特定数据对象 P_1、P_2、P_3...，P_i 之间也延伸有多个层内连接 505。在图 5 所示的实施例中，节点组 U 的某些特定数据对象 U_1、U_2、U_3...，U_j 和节点组 P 的某些特定数据对象 P_1、P_2、P_3...，P_i 之间也延伸有多个层间连接 504。层间连接意的利用识到了一个对象类型的分类要由另一个对象类型而进行。例如，Web 页面对象的分类要由用户对象的配置、状态和特征而进行。

由于数据对象之间的关系可以指向任一方向，所以连接方向（图 5 中的连接 503、504 或 505 的箭头，以及图 6 的箭头）是双向的。该箭头是用于说明的，并不限制本发明的范围。体系图 550 的图中某些连接指向一个方向是更恰当的，箭头的方向一般不影响体系的运行。体系图 550 由节点组 P、节点组 U 和连接组 L 组成。在体系图 550 中，p_i 和 u_j 表示两种类型的数据对象，其中 $p_i \in P$（$i=1,\ldots,I$）并且 $u_j \in U$（$j=1,\ldots,J$）。I 和 J 分别是节点组 P 和 U 的基数。

连接（p_i，u_j）$\in L$ 是层间连接（其是 2 元组），该层间连接由不同类型的对象之间的附图标记 504 表示。分别由 505HE 503 指示的连接（p_i，p_j）$\in L$ 及（u_i，u_j）$\in L$ 是在同类对象间延伸的层内连接。简而言之，层间连接组（504）和层内连接组（503，505）使用不同的参考特征。

利用一体分类，对象间更完全的使用连接改进了分类。提高效率的分类改进了不同层中的不同类型的对象的分类。如果正确分类对象，那么分类结果就更合理。分类可以提供分析数据使用的结构化数据。

体系图 550 表示了多种类型的对象的分类，在这些对象中每种类型的对象都是相同的（即，一种类型从属于 Web 页面组，用户组和文档组，等等）。每组对象的类型通常与体系图 550 中的其他对象组的类型不同。

公开的分类技术在分类时考虑并接收不同对象类型的输入。此文字说明的一方面是根据一种内在共有关系，在这种关系中提供的分类对象具有与其他对象的连接。某些与每个对象联系的连接可以用不同的重要性加权，以便反映与该对象的关联性。例如，可以向与正在分类的对象同类的对象提供比不同类型的对象更大的重要性。此文字说明提供了一种机制，通过该机制可以向不同对象和不同对象类型赋予不同的重要性等级。对不同对象（或不同对象类型）赋予不同重要性等级在这里是指利用重要性分类。不同对象的各种重要性等级通常会改进分类结果和效率。

在图 5 所示的用于不同类对象分类的体系图 550 的实施中，不同节点组 P 或 U 表示了每个都包括不同对象类型的不同层。体系图 550 的多个节点组（表示为 P 和 U）提供了分类的基础。两层指向图 550 包括一组要被分类的数据对象。每种对象类型的对象（其将根据分类算法分类）可以被认为是一个例子“潜在”类。延伸在某些对象节点之间连接 503、504、505 反映了分类提供的对象节点间的固有关系。用于分类的迭代投射（projecting）方案，在本文字说明中描述了其几个实施例，该技术能够独立分类具有独立数据类型的对象，以便进行分类处理。

通过使用这里所描述的迭代分类技术，加强了对象的不同类型（以及他们的关系连接）。迭代分类投射技术依靠从独立层中排列的独立数据类型中获得的分类信息，其中每个独立层都包含同类的对象。与连接信息结合的节点信息用于迭代投射，并传播分类结果（在层间执行分类算法），直到分类收敛。一种对象类型的迭代分类结果到另一种对象类型的迭代分类结果可以减少与数据稀少有关的分类难题。在这种迭代投射下，根据分类，而不是另一类型的分类单个组，计算一层分类中的相似性测量。

检验不同类节点和连接的每种类型，以便获得用于分类的结构信息。举例来说，考虑到与不同数据对象相连的连接的类型（即，连接是层间连接或层内连接），可获得结构信息。如图 5 所指示的，每个对象的类型都由其节点组 P 或 U 指示。

产生的图 5 的体系图 550 可用于特殊的分类应用。也就是，体系图 550 可表示与一组用户有关的 Internet 上的一组 Web 页面。Web 页面层被分组为 P 节点组。用户层对象被分组为节点 U。体系图 550 将用两层体系图 550 表示的多

个 Web 页面对象与多个用户对象结合成一体。体系图 550 使用连接（例如边缘）关系 503、504、505 方便了不同类型的对象的分类（如图 5 所描述的体系图）。在分类过程中，检验整个数据组的连接结构，以便知道节点的不同重要性等级。在文类过程中，根据其重要性给节点加权，以便保证更合理地分类重要节点。

在文字说明的某些实施例中，连接类中的连接 503、504、505 被保留。保留的连接是在对象类间的连接，而不是对象本身间的连接。例如，一个保留的连接位于 Web 页面类和用户类之间（而不是如其原始连接那样位于 Web 页面对象和用户对象之间）。在某些实施例中，保留的连连接是为各种将进行应用而保留的，例如体系图 550 中的建议。即，具有保留连接的 Web 页面/用户分类的分类结果可被表示为用户定位（hit）行为的总图，其提供了用户定位的预测。

各个节点 p_i 和 u_j 的内容用各自矢量 f_i 和 g_j 表示(图 5 中未示出)。根据此应用，每个单独节点 p_i 和 u_j 可具有（不具有任何）内容特征。已有的分类技术独立将节点 p_i 从节点 u_j 中分离出。相反，在此文字描述中说明的分类体系 550 中，根据其相关重要性相互依赖地分类节点 p_i 和节点 u_j。在此所述的分类算法使用相似性函数为每个分类类型测量对象间的距离，以便产生分类。公式（1）所示的 cosine-相似性的函数可用于分类：

$$s_c(x,y)=\cos(f_x,f_y)=\frac{\sum_{i=1}^{kx} f_x(i)\cdot\sum_{j=1}^{ky} f_y(j)}{\sqrt{\sum_{i=1}^{kx} f_x^2(i)}\cdot\sqrt{\sum_{j=1}^{ky} f_y^2(j)}}) \quad (1)$$

$$s_c(x,y)=\cos(f_x,f_y)=\frac{f_x \bullet f_y}{\|f_x\|\|f_y\|}=\frac{\sum_{k f_x(k)=f_y(k)} f_x(k) f_y(k)}{\sqrt{\sum_{i-1}^{kx} f_x^2(i)}\cdot\sqrt{\sum_{j=1}^{ky} f_y^2(j)}} \quad (2)$$

afx▪ fy 是两个特征矢量的点积。它等于 fx 和 fy 中的相同分量的加权积的总和。s_c 表示相似性基于内容特征；$f_x(i)$和 $f_y(y)$是特征矢量 f_x 和 f_y 的第 i 个和第 j 个分量。k_x 是各个特征 f_x 中项的数量；k_y 是特征 f_y 中项的数量。

在此文字说明中，节点组 P 用作描述节点的层内连接 504 和层间连接 503 和 505 的一个例子。假设所有的数据都包括一列节点对，例如由连接 503 或 505 所连接的用于层内节点对（$p^{(1)}$，$p^{(1)}$），（$p^{(2)}$，$p^{(2)}$），…[其中 $p^{(1)}$和 $p^{(2)}$与 p_i 相同，而对（$p^{(1)}$，$p^{(1)}$），（$p^{(2)}$，$p^{(2)}$）都表示同类层中的一个节点]；和连接 504 所拦截

的层间对 （$p^{(1)}$，$u^{(1)}$），（$p^{(2)}$，$u^{(2)}$），...。节点对（p_i，p_j）或（p_i，u_j）间连接表示在数据序列中出现一个或多个等同对。连接加权涉及其出现频率。

在此文字说明中，两个独立矢量表示每个特定节点的层间连接 504 和层内连接 503、505 的特征。例如，利用其分量相应于相同层中的其他节点的矢量表示层内连接 503、505 特征。作为比较，用其分量相应于另一层中的节点的矢量表示层间连接 504 特征。每个分量是一个表示从（或到）该相应节点的连接的加权。例如，节点 p_1 和 p_2 的层间连接 504（如图 5 所示）可分别表示为 $[1,0,0\ldots,0]^T$ 和 $[1,1,1\ldots,0]^T$。

这样，相应的近似函数就可以被定义为上述的 cosine-相似性函数。用于层内连接 503、505 特征的相似性函数 $s_{lx}(x,y)$ 确定了节点 p_1 和 p_2 间的相似性，其由如下的公式（3）所示：

$$s_{l1}(x,y)=\cos(l_x,l_y)=\frac{l_x \bullet l_y}{\| l_x \| \| l_y \|} \qquad (3)$$

作为比较，层间连接 5034 特征的相似性函数 $s_{lx}(x,y)$ 确定了节点 p_1 和 u_2 间的相似性，其由如下的公式（4）所示：

$$s_{l2}(x,y)=\cos（h_x，h_y） \qquad （4）$$

其中，s_{l1} 和 s_{l2} 分别表示相似性是基于各自的层内和层间特征的；l_x 和 l_y 是节点 x 和节点 y 的层内连接特征矢量；而 h_x 和 h_y 是节点 x 和节点 y 的层间连接特征矢量。

也可以使用其他连接特征的表示和其他相似性测量，例如将每个节点的连接表示为一组，并使用 Jaccard 系数。这里描述的实施例具有多个优点。一个优点就是，分类算法的某些实施例建议了加权连接。此外，诸如 k-means 分类算法的分类算法方便了计算分类的矩心。在进一步计算以指示分类对象的一般值和特征中，此矩心是有用的。

节点 x 和节点 y 的所有相似性函数都可以如公式（5）那样被定义为包括三个加权值 α、β 和 γ 的三个相似性的加权总和。有两个公知技术可以指定三个加权值：启发法和学习法。例如，如果没有调整数据，就手工地将加权赋值为需要的值（即，alpha=0.5，beta=0.25 而 gamma=0.25）。如果通过比较有一些额外的调整数据，那么可利用贪婪算法、登山算法或局域或全球改进或优化程序的某些其他类型计算加权。贪婪算法是指异类优化算法，其在每一步中查找

以便改进每个矢量，这样可以最终改进（和某些实施例中优化）解决方案。

$$s(x,y)= \alpha s_c(x,y)+ \beta s_{l1}(x,y)+ \gamma s_{l2}(x,y) \quad (5)$$

其中 $\alpha+\beta+\gamma=1$。

利用这些计算，确定了节点的内容和节点相似性。根据此应用，可以修改这三个变量，从而为分类算法提供了三个不同的信息值。因此节点的这些内容和相似性可以用作检索的基础。

许多不同类分类问题经常共享节点不同样重要的资源。不同类分类的例子包括 Web 页面/用户分类、用于合作过滤的项目/用户分类，等等。对于这些应用，在获得更合理的分类结果中，重要的对象同样具有重要的作用。在此文字描述中，整个数据组的连接结构都用于获知节点的重要性。对于节点组 P 和 U 中的每个节点，例如，p_i 和 u_j，重要性加权 ip_i 和 iu_j 利用连接结构计算，并被用于分类过程。

一个分类方面涉及连接分类算法，在本文字说明中提供了其多个实施例。在连接分析算法的一个实施例中，构造了如图 6 所示的混合网络模型 600。利用混合网络模型 600，用户和 Web 页面可被用作两个节点类型。图 6 中包括 Web 页面和用户对象类型的混合网络模型的实施例特别用于包括 Internet、内联网或其他网络的类型分类。连接包括连接 605 所示的 Web 页面超文本链接/交互作用、连接 604 所示的用户-Web 页面超文本链接/交互作用，以及连接 603 所示的用户-用户超文本连接/交互作用。图 6 的混合网络模型 600 通过指示用连接 603、604 和 605 表示用户和 Web 页面之间的关系而说明了这些超文本链接/关系。

假设用户组 610 中包括一个特定用户组 608，用户组 610 访问的任何节点的所有 Web 页面形成了 Web 页面组 612。通过向搜索引擎发送根 Web 页面，Web 页面组 612 被确定了，并获得了一个基本 Web 页面组。由图 6 的箭头表示的三个连接具有不同的含义。箭头 605 表示的包含在 Web 页面组 612 中的连接指示 Web 页面见到超文本链接。箭头 603 表示的包含在用户组 610 中的连接指示用户的社交关系。在用户组 610 和 Web 页面组 612 之间的由箭头 604 表示的连接指示了用户对 Web 页面的访问行为。箭头 604 表示的连接指示了每个特定 Web 页面的用户评价，这样 Web 页面的权限/集中分数就更可信。由于不同类型的连接 603、604 和 605 表示不同的关系，每个连接都可以根据例如该连接的访问频率或者该连接所连接的每个节点对如何被联系，用不同的重要性加权。

图 7 示出了被配置为利用 Internet 进行分类的计算机环境 400 的一个实施例。此分类的一个方面包括根据用户分类 Web 页面（包括相关层间连接的层内连接）。该计算机环境包括多个 Web 站点 750、搜索引擎 752、服务器/代理部 754、建模模块 756、计算模块 758 以及建议/参考部 760。计算环境 400 例如通过图形用户接口（GUI）于用户 762 接口。计算模块 758 包括迭代计算部件 780，该部件 780 执行该分类算法（依靠迭代计算的特定实施例）。建模模块 756 收集数据并跟踪数据（例如与对象有关的数据）。搜索引擎根据用户查询返回搜索结果。Web 站点 750 原样向用户显示数据。服务器/代理将查询等传输给执行大量分类的服务器。建议/参考部 760 允许用户修改和选择分类算法。

建模模块 756 包括在先的格式化部件 770、web 页面提取部 772 和用户提取部 774。部件 770、772 和 774 被配置为提供和/或跟踪由 770 在先格式化的数据，从 Web 页面中提取的数据或从用户 762 中提取的数据。如图 7 所示的计算环境的一个实施例被配置为提供一个连接分析算方，在本文字描述中描述了该算法的一个实施。

分类算法的一个实施例可以通过查找两种类型的页面：集中（hubs）、权限（authorities）和用户，而分析一个 Web 图形。集中是页面，该页面与大量的其他页面相连，这些其他页面提供对特定主题有用的相关信息。权限页面被认为是与许多集中相关的页面。用户访问每个权限和集中。因此，集中、权限和用户中的每一对都显示了相互加强的关系。分类算法依赖于在本连接分析算法的每些实施中使用的三个矢量：web 页面权限加权矢量 a，集中加权矢量 h 以及用户矢量 u。在本文字说明中描述了这些使两大某些方面。

下列加权计算中所涉及的几个下列项没有在图 7 的图形中示出，并且其涉及到该计算。在一个实施中，对于一给定用户 i，用户加权 u_i 说明了他/她的知识水平。对于一 Web 页面 j，项 a_j 和 h_j 表示权限加权和集中加权。在一个实施中，三个矢量（表示用户矢量 u、web 页面权限加权矢量 a 和集中加权矢量 h）的每一个都分别初始为某值（例如 1）。然后根据下列公式（6）、（7）和（8）的计算，基于 Internet 的使用，迭代更新全部的三个矢量 h、a 和 u：

$$\begin{cases} a(p) = \sum_{q \to p} h(q) + \sum_{r \to p} u(r) & (6) \\ h(p) = \sum_{p \to q} a(q) + \sum_{r \to p} u(r) & (7) \\ u(r) = \sum_{r \to p} a(p) + \sum_{r \to q} h(q) & (8) \end{cases}$$

其中，p 和 q 表示特定的 web 页面，r 表示特定用户。在公开的网络的某些实施中，具有两类连接：不同页面间的连接（超文本链接）以及用户和页面之间的连接（浏览模式）。使 A=[a_{ij}]表示三个矢量 h、a 和 u 的基本集的相邻矩阵。注意，如果页面 i 连接到页面 j，那么 a_{ij}=1，否则相应的 a_{ij}=0。V=[v_{ij}]是用户组到 Web 页面组的访问矩阵。假设如果用户 i 访问了页面 j，那么 v_{ij}=1，否则 v_{ij}=0。同样，如公式（9）、（10）以及（11）为：

$$\begin{cases} a = A^T h + V^T u & (9) \\ h = Aa + V^T u & (10) \\ u = V(a + h) & (11) \end{cases}$$

在一个实施中，公式（9）、（10）和（11）中的矢量 a、h、u 的计算进行几个迭代，以提供有意义的结果。某些实施中，在迭代前，给每个矢量 a、h 和 u 设者为随机值。接着在每个迭代中，a、h、u 的值都会改变并且为给下一迭代提供基础，而被规格化。接着在每个迭代中，a、h 和 u 的每个迭代值都试图收敛到各自的某值。可以报告具有高用户加权 u_i 的用户以及具有高权限加权 a_j 和/或集中加权 h_j 的 Web 页面。在一优选的实施中，特定的各个用户或 web 页面对象可被赋予一个比其他的各个用户或 web 页面对象都高的值。值越高，该对象就越重要。

因此，在可分类的本文字说明中描述的连接分析算法的实施依赖于来自 web 页面和用户的迭代输入。这样，用户的加权输入就应用于 web 页面的分类算法。分类中加权的用户输入的使用增加了搜索结果的精度和分类算法的执行速度。

虽然这里描述的连接分析算法基于用户用于分类 web 页面的分类算法，但是可以预见，连接分析算法可以用于任何不同类分类算法。正如这里所描述的，加权部分地提供了具有重要性的分类。

已经描述了可用于分类对象类型的分类算法的各种实施例。分类算法意于根据要分类的数据对象间的某些相似性，查找数据类型的自然组。这样，分类

算法对数据对象执行分类操作。分类算法的某些实施也查找一组数据集的矩心，其表示了一个点，该点的参数值是分类中所有点的参数值的平均值。为了确定分类关系，大多数的分类算法都评估点和分类矩心间的距离。分类算法的输出基本上是每个分类具有几个分量的分类矩心的统计学描述。

在此文字说明中描述了分类算法的多个实施。双向 k-means 分类算法是基于分类过程的共有加强的。双向 k-means 分类算法是一个迭代分类算法。在双向 k-means 分类算法中，首先通过公式（6）-（8）或（9）-（11）计算对象重要性，接着该结果就用于随后的迭代分类过程。分类算法根据限定的相似性函数将每层中的对象分类。尽管要使用大量的分类算法，例如，k-means、k-medoids 和会聚分层方法，但是本文字说明表述了 k-means 分类算法的应用。

计算的节点重要性分数有几种技术。一种技术包括将基本的 k-means 分类算法修改为“加权的” k-means 分类算法。在修改的 k-means 算法中，利用具有加权设置的特性的加权总和，计算给定分类的矩心，其中，该加权设置用于确定重要性分数。因此在为内容和连接特征形成分类矩心时，更关注具有较高重要性或加权的节点。另一种实施包括用其重要性分数修改节点的连接加权，然后利用在相似性公式中利用该加权的连接特性。这种方法中，在分类过程中，仅仅在连接特征中反应了节点重要性。

图 8 和图 9 中示出了分类算法的输入/输出的一个实施。分类算法的输入包括两层体系图 550（其包括节点的内容特征 f_i 和 g_i）。分类算法的输出包括反映分类的新的体系图 550。在新体系图的某些实施中，可表示出已经改变到其新节点位置的各个老节点的变化。

图 8 和图 9 中示出了说明分类算法 850 一个实施例的流程图。分类算法 850 包括 851，其中输入原始体系图（在每个分类迭代之前）。在 852 中，利用公式（6）-（8）或（9）-（11）确定或计算考虑的每个节点的重要性。在 854 中，为了进行分类，选择任意层。注意，在 855 中，以适合的方式（例如，根据内容特征）分类选择的层。在某些实施中，可利用需要的过滤算法过滤节点。在 856 中，将每类中的节点合并成一个节点。例如，如果在过滤后存在两个候选节点，例如通过平均这两个候选节点的向量值，将最接近的两个候选节点合并。此合并允许组合单个节点，以便减少要考虑的节点的数量。因此，合并操作可用于减少重复和类似重复（near-duplicates）的出现。

根据857中的合并结果，更新相应的连接。在858中，分类算法改变（从任选的层）到分类的第二层。在960中，根据其内容特征，分类第二层的节点，并更新其特征。在961中，每类节点都被合并成一个节点。

在节点962，重新存储其他层的原始连接结构和原始节点。在964中，第二层的每类节点都被合并，并更新相应的连接。在966中，在计算环境中继续进行此迭代分类过程。在968，输出修订后的体系图550。

在最初的分类中，仅仅使用了内容特征。由于在大多数的情况下，开始时连接特征太少而不能用于分类。在随后的分类过程中，要合并内容特征和连接特征，以便改进分类效率。通过合并内容特征和连接特征，加权被确定为用不同的值，并可比较结果，提供了具欧改进精度的分类。

参照图8和9描述的分类算法可以用于许多分类实施。更具体地，根据用户如何访问Web页面的Web页面分类的一个实施是新描述的。那些在用户层的一个节点到Web页面层节点间的那类连接中，如果具有从u_j到p_i的一个连接，用户u_j之前就已经访问过了Web页面p_i。连接加权意味着在特定时候，用户u_j访问Web页面p_i的可能性，并被表示为$Pr(p_i|u_j)$。如公式(12)所示，通过计算已观察到数据数量，其可以简单计算。

$$\Pr(p_i \mid u_j) = \frac{C(p_i, u_j)}{\sum_{i \in p(u_j)} Cz(p_i, u_j)} \qquad (12)$$

其中，$p(u_j)$是用户u_j之前访问的页面组。$C(p_i,u_j)$表示之前用户u_j之前访问页面p_i的计数。

如图10的体系图550的实施所示，分类算法的一个实施包括一个概念层或隐蔽层。在图10中，对于相似性，图5的体系层中的层内连接503和505是隐含的。但是，可以想象，图10所示的体系图550的实施可以依赖层内连接和层间连接任何组合，并且仍然保留在本文字说明概念内。

隐含层1070（如图10中所示的体系图550的实施中）位于web页面和用户层间。隐含层550提供了提取的附加层（连接从该层延伸到每个节点组P和U），该提取附加层允许用改进的实体（realism）建模，该改进是与原始节点组P和U之间的延伸连接相比的。例如图5的体系图550实施中（其不具有隐含层）的一个层间连接504可以被建模为体系图550实施的一对隐含层内连接，

例如图 10 所示的。一个隐含层间连接延伸在包含节点组 P 的 web 页面层和隐含层 1070 之间，而一个隐含层间连接延伸在用户层和隐含层 1070 之间。图 10 所示的每个隐含层间连接的箭头方向是任意的，就按照各自节点组 P 和 U 中的特定 web 页面和用户的原样，其由到隐含层的隐含层间连接所连接。

包含节点组 P 的 web 页面层和隐含层 1070 间的连接（即，隐含层间连接）指示了 web 页面 p_1、p_2 等如何可能属于隐含层 1070 中特定概念节点 $P(c_1)$、$P(c_2)$。用户层和隐含层 1070 将的连接（即，隐含层内连接）指示了用户节点 u_1、u_2 等如何可能对隐含层 1070 中的特定概念节点 $P(c_1)$、$P(c_2)$等有兴趣。

因此，web 页面层和概念层间的每个连接都表示 web 页面 p_i 被分类为概念种类 c_k 的可能性，其被表示为 $Pr(p_i|c_k)$。由体系图实施的此模型共享了 Narve Bayesian 分类所用的设想，其中一般认为不同的词是独立的。这样，概念 c_k 可被表示为正态分布，例如期望矢量 $\overrightarrow{\mu_k}$ 和协方差向量 $\overrightarrow{\sigma_k}$。值 $Pr(p_i|c_k)$可按公式(13)导出。

$$E(\Pr(p_i \mid c_k)) = \frac{\Pr(p_i \mid c_k)}{\sum_i \Pr(pi \mid c_k)} = \frac{\prod_l \Pr(w_{l,i} \mid c_k)}{\sum_i \prod_l \Pr(w_{l,i} \mid c_k)} = \frac{e^{-\sum \frac{1}{2\sigma_{l,k}}(w_{l,i}-\mu_{k,k})^2}}{\sum_l e^{-\sum \frac{1}{2\sigma_{l,k}}(w_{l,i}-\mu_{k,k})^2}} \quad (13)$$

其中，$w_{l,i}$ 是第 l 个字上的 web 页面 p_i 的加权。

那些位于用户层一个节点和隐含层一个节点之间那些连接反应了用户在由该概念反应的种类上的兴趣。这样，一个矢量（$I_{j1}, I_{j2}, \ldots I_{jn}$），$I_{jk}=Pr(c_k|u_j)$就响应于每个用户，其中 n 是隐含概念的数量。图 10 所示的连接可被看作是用户的矢量模型。该矢量由用户的使用数据所约束，如公式（14）所示：

$$\Pr(p_i \mid u_j) = \sum_l \Pr(p_i \mid c_l, u_j)\Pr(c_l \mid u_j) \approx \sum_l \Pr(p_i \mid c_l)\Pr(c_l \mid u_j) \quad (14)$$

这样，值 $Pr(c_k|u_j)$可从（13）中通过查找解法而获得。

为了简化，$Pr(p_i|u_j)=R_{i,j}$，$Pr(p_i|c_k)=S_{i,k}$ 而 $Pr(c_k|u_j)=T_{k,j}$。如公式 15 所示，用户 j 可被认为是独立的。

$$\begin{bmatrix} R_{1,j} \\ R_{2,j} \\ \cdots \\ R_{|Page|,j} \end{bmatrix} = \begin{bmatrix} S_{1,1} & S_{1,2} & \cdots & S_{1,|Concept|} \\ S_{2,1} & S_{2,2} & & \\ & & \cdots & \\ S_{|Page|,1} & & \cdots & S_{|Page|,|Concept|} \end{bmatrix} \times \begin{bmatrix} T_{1,j} \\ T_{2,j} \\ \cdots \\ T_{|Concept|,j} \end{bmatrix} \quad (15)$$

其中“|Page|”是 Web 页面的总数，而“|Concept|”是隐含概念的总数。由

于|Page|>>|Concept|，所以 T_{kj} 的最少平方解可利用公式(15)或可选择的公式(16)解答。

$$[R_{i,1} \quad R_{i,2} \quad \cdots \quad R_{i,|User|}] = [S_{i,1} \quad S_{i,2} \quad \cdots \quad S_{i,|Concept|}] \times \begin{bmatrix} T_{1,1} & T_{1,2} & \cdots & T_{1,|User|} \\ T_{2,1} & T_{2,2} & & \\ \cdots & & \cdots & \\ T_{|Concept|,1} & & & T_{|Concept|,|User|} \end{bmatrix} \quad (16)$$

其中“|User|”是用户的总数。

由于|User|>>|Concept|，所以我们可以给出 $S_{i,k}$ 的最少平方解，如公式（17）所示。

$$\overrightarrow{\mu_j} = \sum_l \overrightarrow{P_l} \Pr(p_l \mid c_k) = \sum_K S_{l,k} \overrightarrow{P_L} \quad (17)$$

在获得了期望 $\overrightarrow{\mu_j}$ 的矢量后，就可以计算协方差 $\overrightarrow{\sigma_j}$ 的新矢量。虽然图 10 所述的体系图 550 的实施位于节点组 P 和节点组 U 之间，但是可以想象，节点组的特殊内容自然说明的，并且可以用于任何节点组。

在分类算法的一个实施中，web 页面对象可以根据用户对象分类，该分类算法的实施可以相对于图 1 中的 1100 所示的 web 页面分类算法而描述如下：

1. 如 1102 所示，收集用户日志组。

2. 计算在公式（12）示出的特定时间 $\Pr(p_i|u_j)$用户 u_j 访问 web 页面 p_i 的可能性，如图 11 的 1104。

3. 在图 11 的 1106 中，限定隐含概念层（图 10 所示的 670）的节点数量|Concept|，并且在图 11 的 1108 中为随机给期望矢量 $\overrightarrow{\mu_j}$ 和协方差矢量 $\overrightarrow{\sigma_j}$ 的初始参数赋值。

4. 计算 $\Pr(p_i|c_k)$的值，其表示 Web 页面被归为概念种类的可能性，如公式（13）所示以及图 11 的 1110 所示。

5. 计算 $\Pr(c_k|u_j)$，其表示用户对一用户节点和一隐含层节点间的连接感兴趣，其可由公式（15）导出，并如图 11 的 1112 所示。

6. 更新 Web 页面被归为概念种类的 $\Pr(p_i|c_k)$可能性，其在概述步骤 14 中用公式（13）确定，如图 11 的 1114 所示。

7. 利用公式（13）所示的 $\Pr(p_i|c_k)$重新设置每个隐含概念节点的参数。

8. 几次迭代执行（13）和（15），以便提供节点组值的基础（或者至少直到模型结果显示稳定的节点组矢量）。

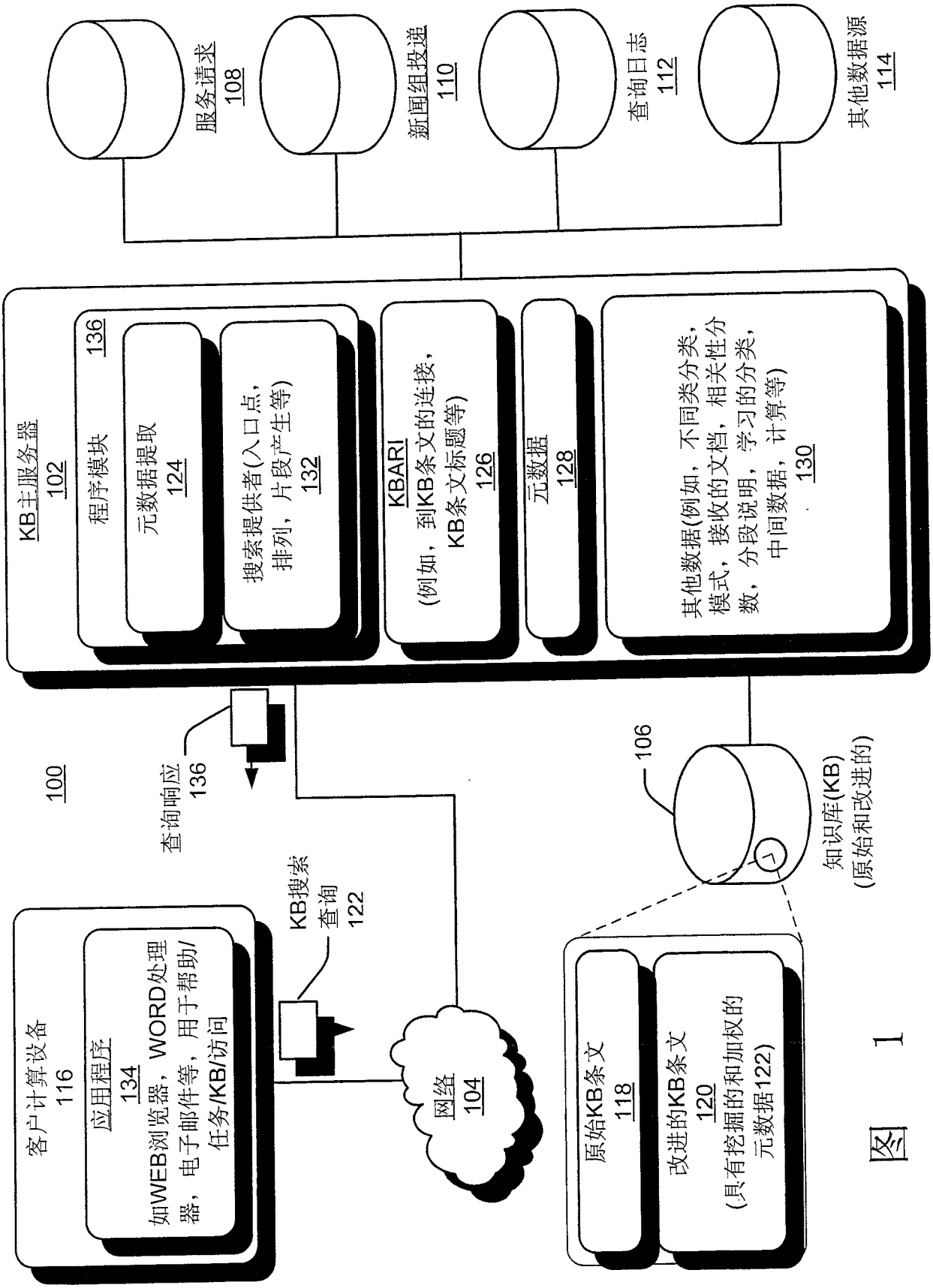

服务请求
108
新闻组投递
110
查询日志
112
其他数据源
114
KB主服务器
102
程序模块
136
元数据提取
124
搜索提供者(入口点，
排列，片段产生等)
132
KBARI
(例如，到KB条文的连接，
KB条文标题等)
126
元数据
128
其他数据(例如，不同类分类，
模式，接收的文档，相关性分
数，分段说明，学习的分类，
中间数据，计算等)
130
100
查询响应
136
106
知识库(KB)
(原始和改进的)
客户计算设备
116
应用程序
134
如WEB浏览器，WORD处理
器，电子邮件等，用于帮助/
任务/KB/访问
KB搜索
查询
122
网络
104
原始KB条文
118
改进的KB条文
120
(具有挖掘的和加权的
元数据122)

图 1

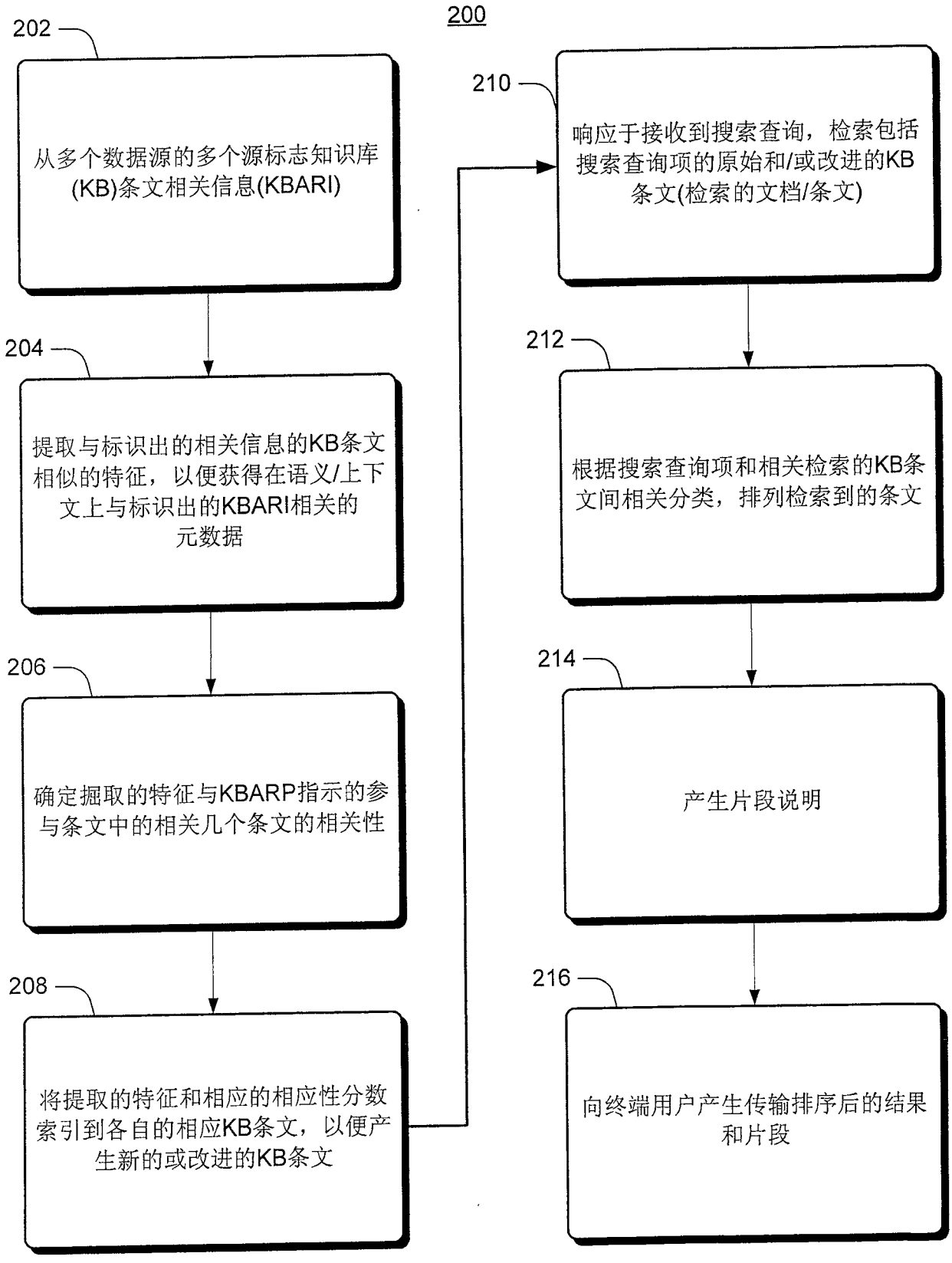

200
202
从多个数据源的多个源标志知识库(KB)条文相关信息(KBARI)
204
提取与标识出的相关信息的KB条文相似的特征，以便获得在语义/上下文上与标识出的KBARI相关的元数据
206
确定掘取的特征与KBARP指示的参与条文中的相关几个条文的相关性
208
将提取的特征和相应的相应性分数索引到各自的相应KB条文，以便产生新的或改进的KB条文
210
响应于接收到搜索查询，检索包括搜索查询项的原始和/或改进的KB条文(检索的文档/条文)
212
根据搜索查询项和相关检索的KB条文间相关分类，排列检索到的条文
214
产生片段说明
216
向终端用户产生传输排序后的结果和片段

图 2

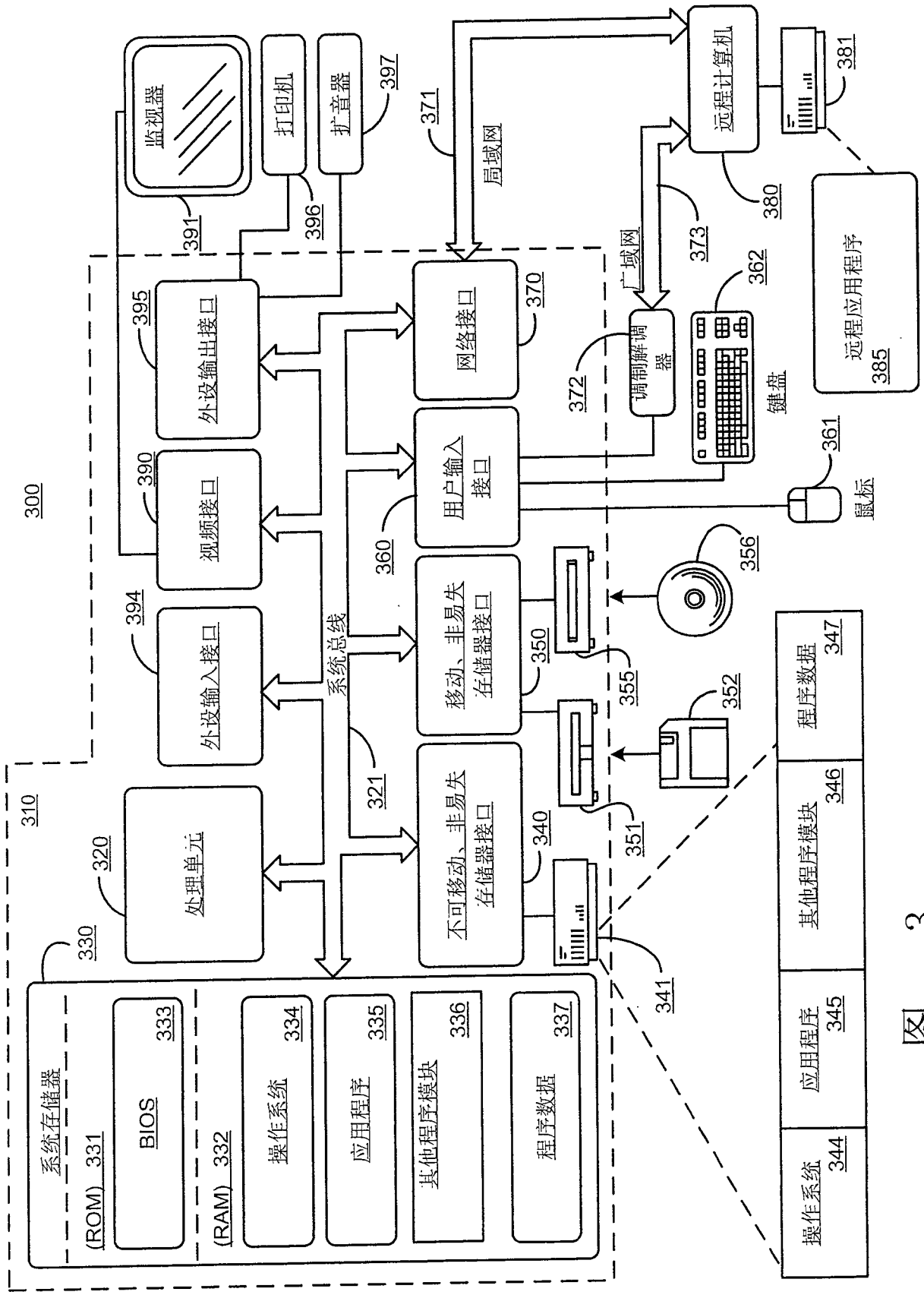

300
310
320
330
处理单元
系统存储器
(ROM) 331
BIOS 333
(RAM) 332
操作系统 334
应用程序 335
其他程序模块 336
程序数据 337
系统总线 321
外设输入接口 394
视频接口 390
外设输出接口 395
网络接口 370
用户输入接口 360
移动、非易失存储器接口 350
不可移动、非易失存储器接口 340
监视器 391
打印机 396
扩音器 397
局域网 371
广域网 373
调制解调器 372
远程计算机 380
远程应用程序 385
381
键盘 362
鼠标 361
356
352
355
351
341
操作系统 344
应用程序 345
其他程序模块 346
程序数据 347

图 3

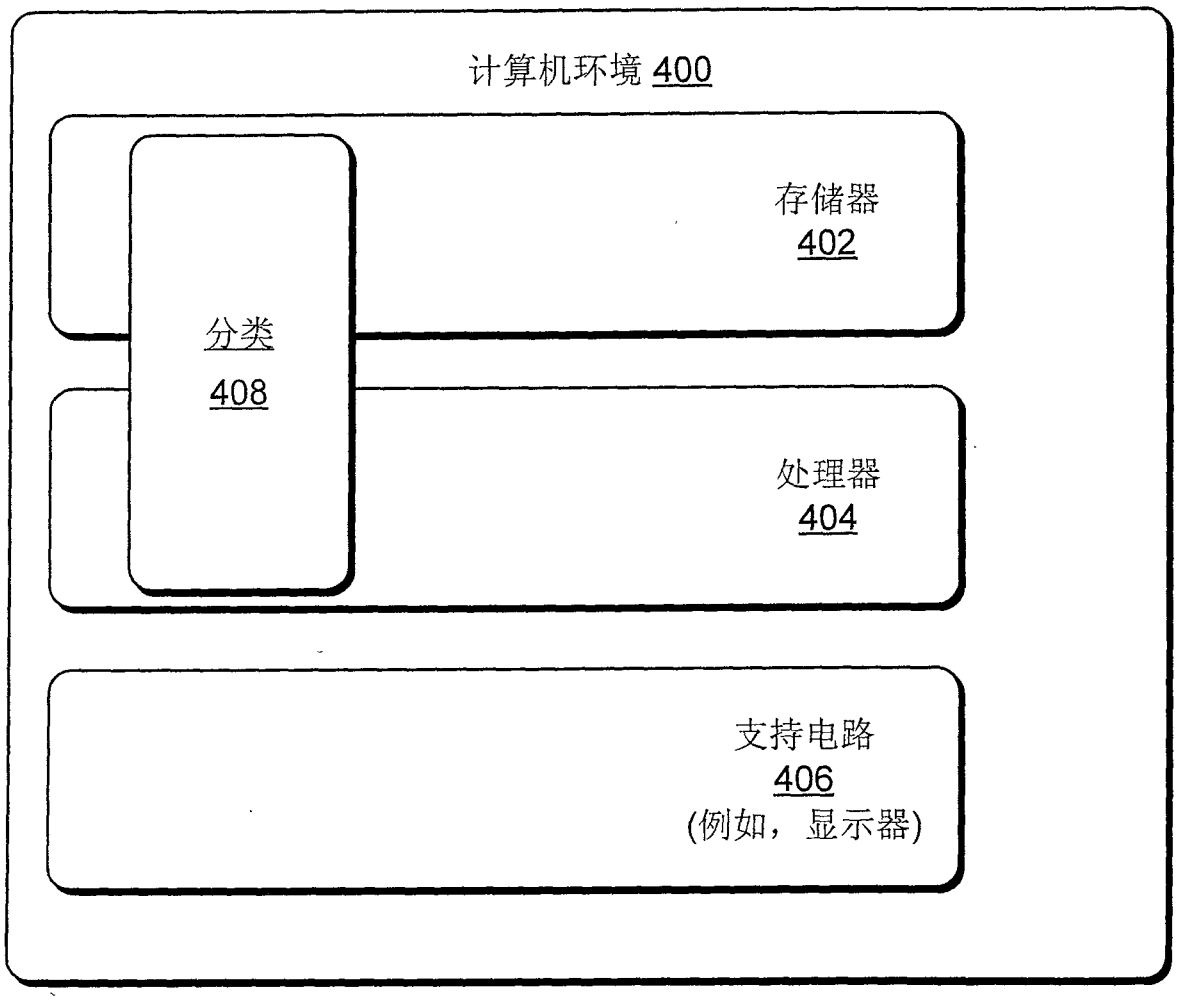


图 4

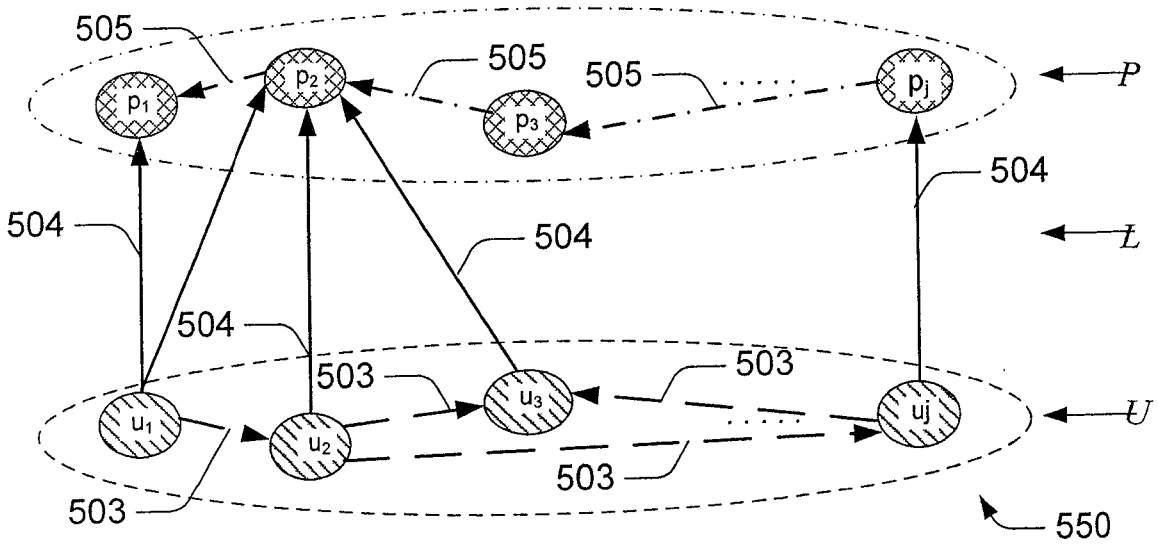

505
505
505
P1
P2
P3
Pj
P
504
504
504
504
L
503
503
503
503
u1
u2
u3
uj
U
550

图 5

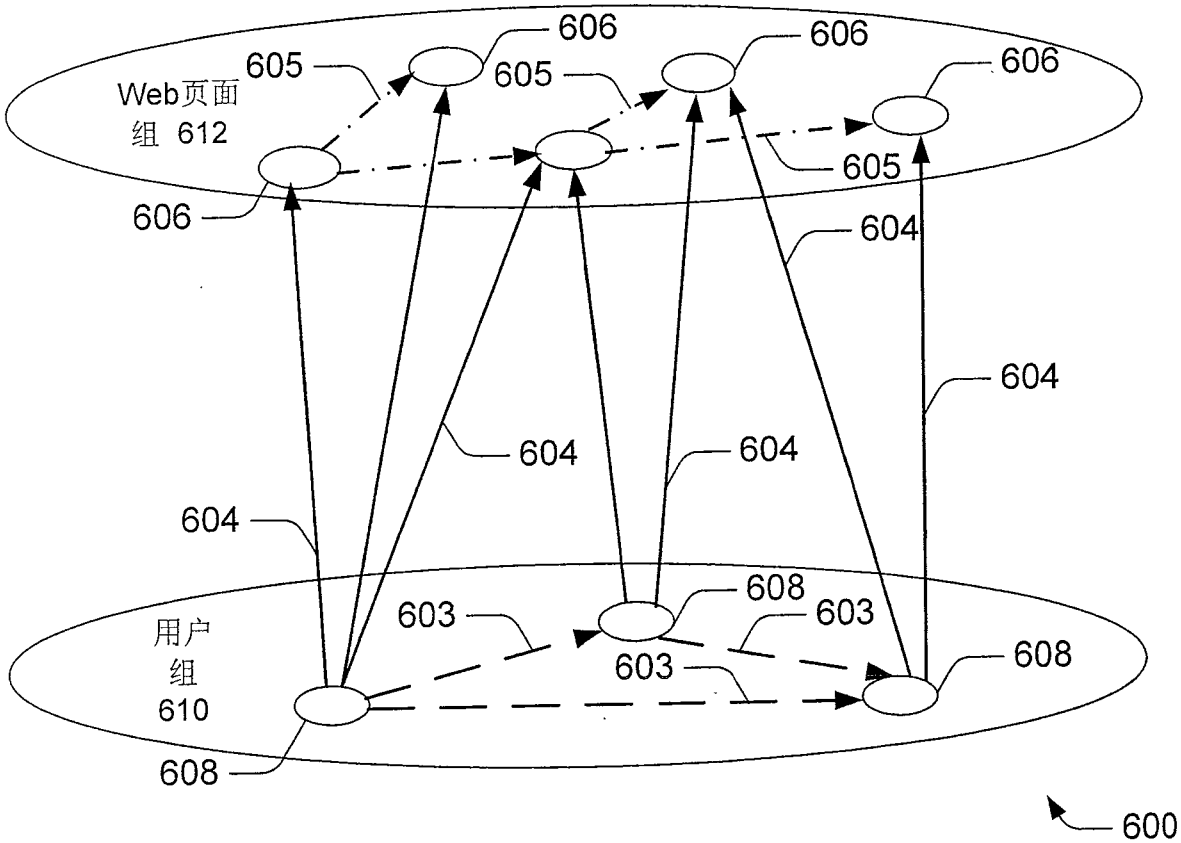

606
606
606
605
605
605
Web页面
组 612
606
604
604
604
604
604
603
608
603
603
608
用户
组
610
608
600

图 6

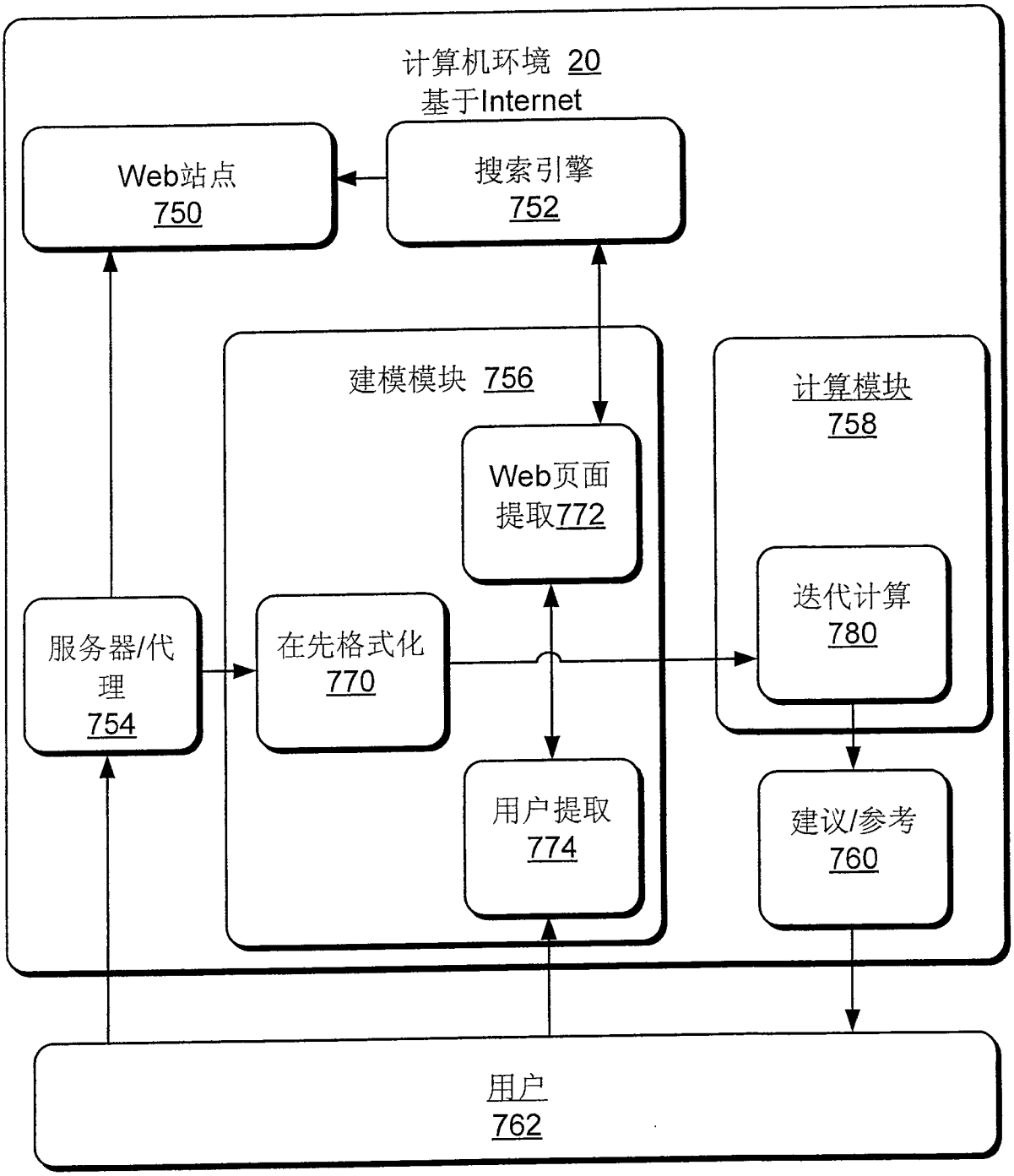

计算机环境 20
基于Internet
Web站点
750
搜索引擎
752
建模模块 756
计算模块
758
Web页面
提取772
迭代计算
780
服务器/代
理
754
在先格式化
770
用户提取
774
建议/参考
760
用户
762

图 7

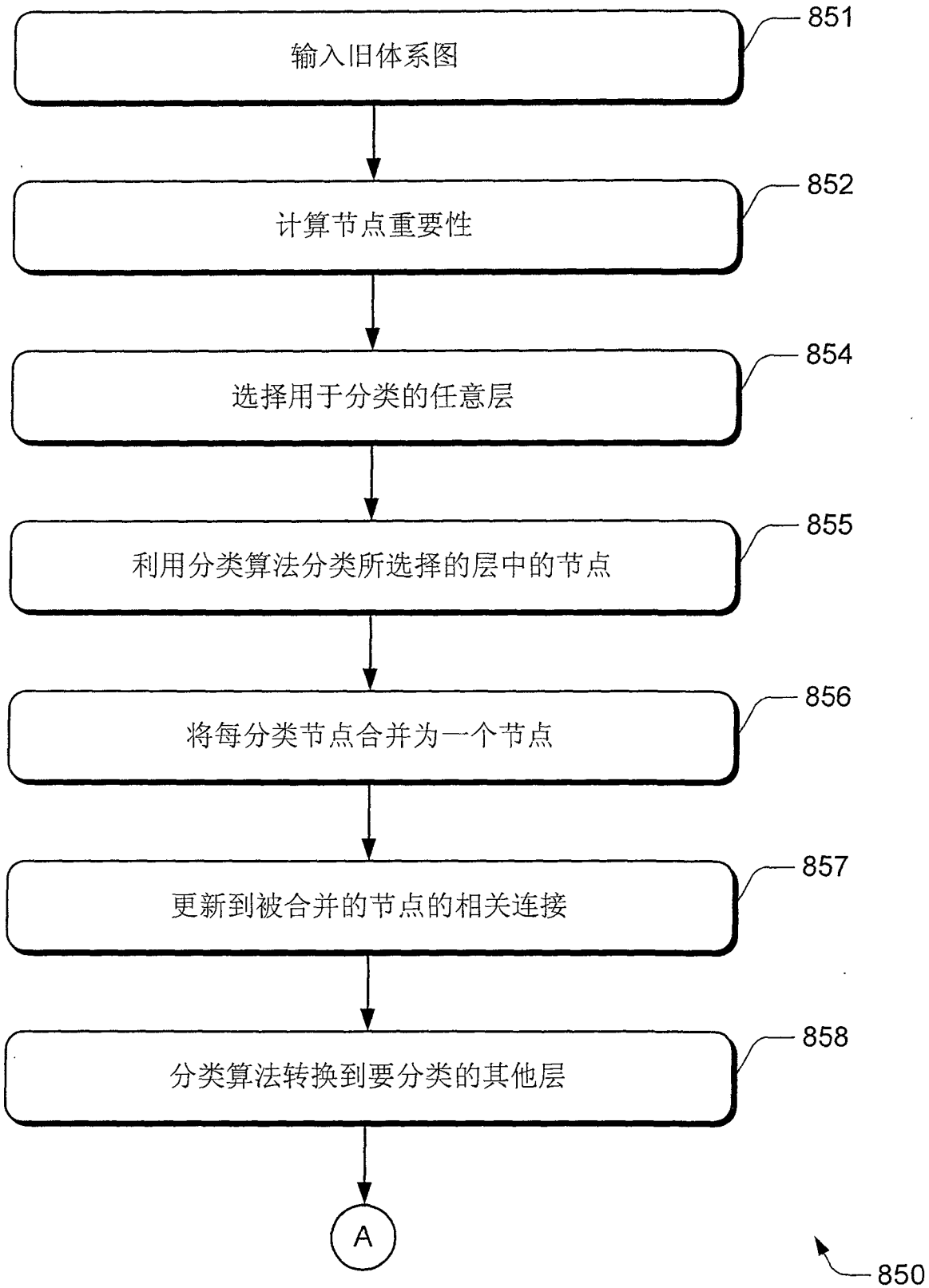

输入旧体系图
851
计算节点重要性
852
选择用于分类的任意层
854
利用分类算法分类所选择的层中的节点
855
将每分类节点合并为一个节点
856
更新到被合并的节点的相关连接
857
分类算法转换到要分类的其他层
858
A
850

图 8

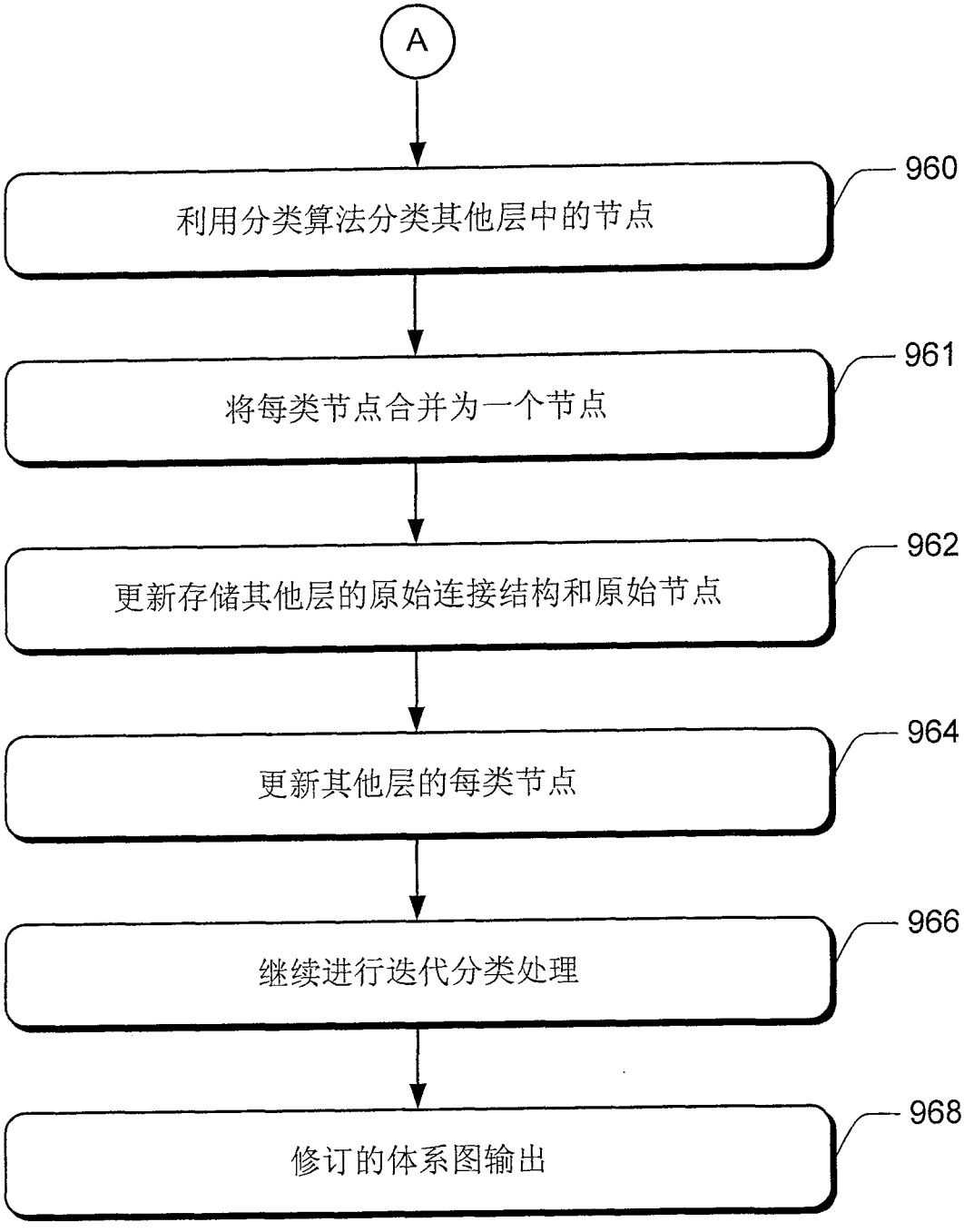

A
利用分类算法分类其他层中的节点
960
将每类节点合并为一个节点
961
更新存储其他层的原始连接结构和原始节点
962
更新其他层的每类节点
964
继续进行迭代分类处理
966
修订的体系图输出
968
850

图 9

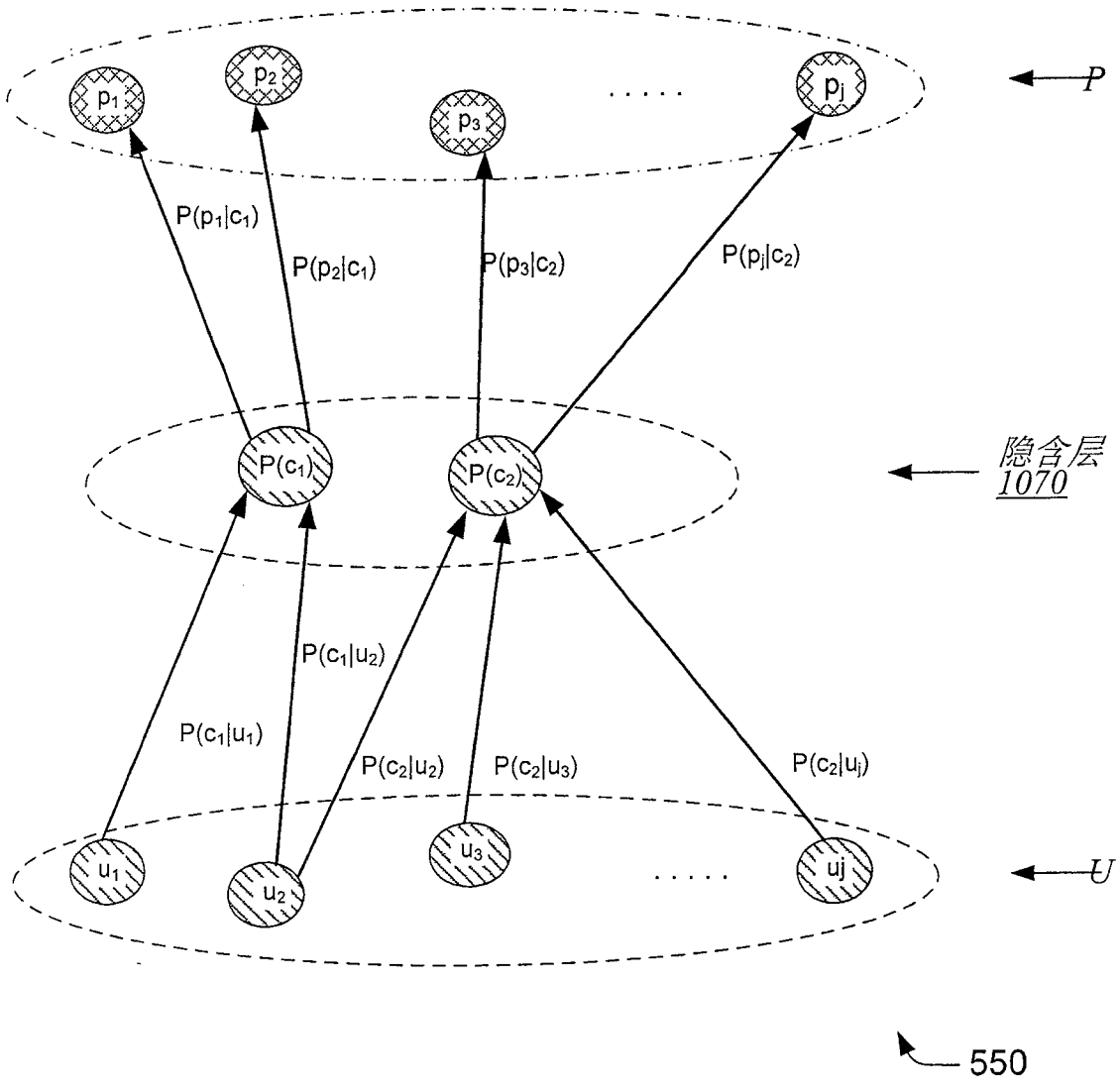

P
p1
p2
p3
pj
P(p1|c1)
P(p2|c1)
P(p3|c2)
P(pj|c2)
隐含层
1070
P(c1)
P(c2)
P(c1|u2)
P(c1|u1)
P(c2|u2)
P(c2|u3)
P(c2|uj)
U
u1
u2
u3
uj
550

图 10

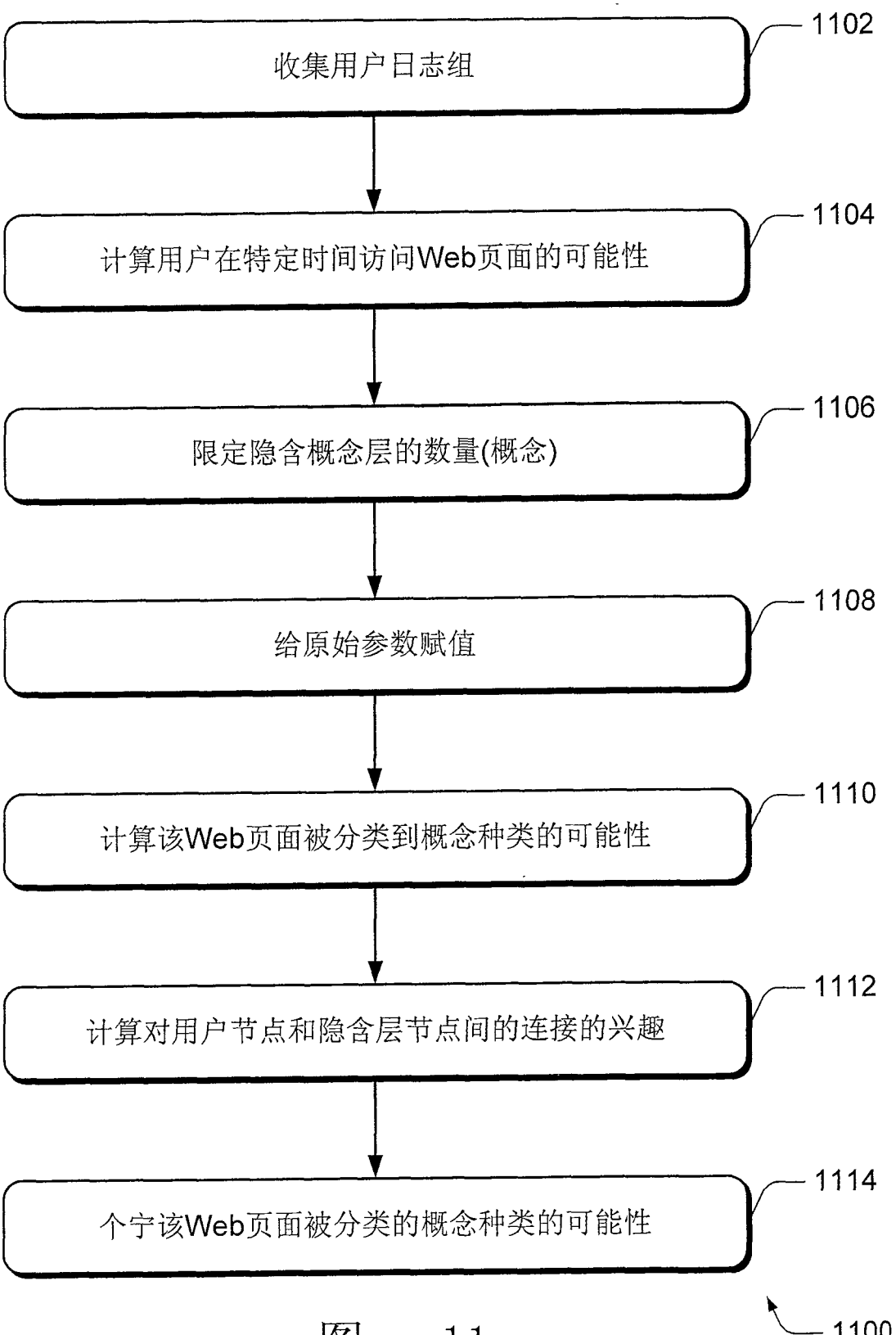

收集用户日志组
1102
计算用户在特定时间访问Web页面的可能性
1104
限定隐含概念层的数量(概念)
1106
给原始参数赋值
1108
计算该Web页面被分类到概念种类的可能性
1110
计算对用户节点和隐含层节点间的连接的兴趣
1112
个宁该Web页面被分类的概念种类的可能性
1114
1100

图 11