(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 104160390 A
(43)申请公布日 2014.11.19

(21)申请号 201380013249.X
(22)申请日 2013.02.22
(30)优先权数据
13/413,179 2012.03.06 US
(85)PCT国际申请进入国家阶段日
2014.09.09
(86)PCT国际申请的申请数据
PCT/US2013/027203 2013.02.22
(87)PCT国际申请的公布数据
WO2013/133985 EN 2013.09.12
(71)申请人 微软公司
地址 美国华盛顿州
(72)发明人 K·K·盖加姆 K·查卡拉巴蒂
M·A·亚考特 S·乔德里
(74)专利代理机构 上海专利商标事务所有限公司 31100
代理人 胡利鸣
(51)Int.Cl.
G06F 17/00(2006.01)
G06F 17/30(2006.01)

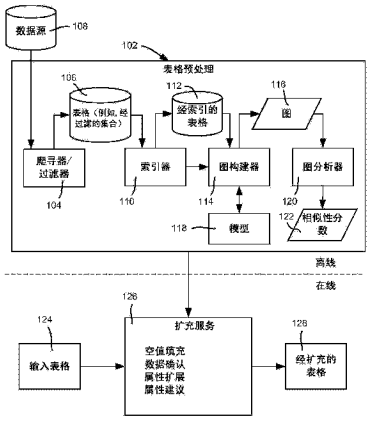
权利要求书2页 说明书15页 附图6页

(54)发明名称

来自潜在关系数据的实体扩充服务

(57)摘要

本发明涉及提供用于扩充与实体-属性-相关的任务的数据。对从web提取的实体-属性表格执行预处理，例如以提供可被访问来寻找完成扩充任务的数据的索引。索引基于表格之间的直接映射和间接映射两者。示例扩充任务包括基于属性名称或示例对被扩充的数据的查询，或寻找用于扩充的同义词。通过访问索引来返回与任务相关的被扩充的数据以高效地处理在线查询。



CN 104160390 A

1. 一种在计算环境中至少部分地在至少一个处理器上执行的方法，包括：处理扩充任务，包括访问与从至少一个语料库中挖掘的实体和属性之间的包括至少一个间接关系在内的关系相对应的基于关系的数据以及使用所述基于关系的数据来寻找完成所述扩充任务的数据。

2. 如权利要求 1 所述的方法，其特征在于，还包括：

a) 接收作为对被扩充的数据的查询的所述扩充任务，其中所述查询标识实体集和属性名称，并且其中使用所述基于关系的数据来完成所述扩充任务包括查找针对所述实体集的每一个实体的至少一个属性名称，

b) 接收作为对被扩充的数据的查询的所述扩充任务，其中所述查询标识实体集和一个或多个属性示例，并且其中使用所述基于关系的数据来完成所述扩充任务包括基于所述一个或多个属性示例来查找针对所述实体集的至少一个实体的至少一个属性名称，或

c) 接收作为对同义词数据的查询的所述扩充任务，其中所述查询标识名称，并且其中使用所述基于关系的数据来完成所述扩充任务包括查找针对所述名称的至少一个同义词。

3. 如权利要求 1 所述的方法，其特征在于，访问所述基于关系的数据包括标识种子表格并经由偏好矢量和表格的存储的矢量来计算主题敏感页面排名分数。

4. 如权利要求 1 所述的方法，其特征在于，使用所述基于关系的数据来寻找完成所述扩充任务的数据包括从最终预测中聚集并选择值。

5. 一种系统，包括：服务，所述服务被配置成处理与领域无关的实体扩充任务，包括将从至少一个语料库获得的关系表格预处理成多个索引，其中所述索引包括基于所述表格的至少一些之间的间接映射的数据；以及用经由所述索引获得的数据对对应于实体扩充任务的查询进行响应，包括经由所述索引标识种子表格，计算每个种子表格的分数，基于所述种子表格的矢量来计算偏好矢量，基于所述偏好矢量和与所述关系表格中的至少一些相关联的矢量来计算预测分数，聚集所述分数，并基于对所述分数的聚集来返回最终预测以完成实体扩充任务以对所述查询进行响应。

6. 如权利要求 5 所述的系统，其特征在于，所述服务被进一步配置成通过访问从包括基于间接映射在内的所述关系表格中构建的与同义词相关的索引来处理对同义词的请求。

7. 如权利要求 6 所述的系统，其特征在于，所述间接映射基于从所述表格中构建的图上的主题敏感页面排名。

8. 一个或多个具有计算机可执行指令的计算机可读介质，所述计算机可执行指令在被执行时执行以下步骤，包括：

将从语料库中提取的实体属性关系表格预处理成被用于实体扩充的索引，包括执行表格之间的整体匹配，所述整体匹配包括计算针对所述表格的至少一些之间的直接关系和间接关系的值；以及

访问所述索引来处理实体扩充任务。

9. 如权利要求 8 所述的一个或多个计算机可读介质，其特征在于，执行所述整体匹配包括使用与所述表格的至少两个相关联的上下文。

10. 如权利要求 8 所述的一个或多个计算机可读介质，其特征在于，进一步包括计算机可执行指令包括，接收作为对被扩充的数据的查询的所述扩充任务，其中所述查询标识实体集和属性名称；接收作为对被扩充的数据的查询的所述扩充任务，其中所述查询标识实

体集和一个或多个属性示例；或接收作为对同义词数据的查询的所述扩充任务，其中所述查询标识名称。

来自潜在关系数据的实体扩充服务

[0001] 背景

[0002] 信息工作者（用户）与包括提及各种实体的文档、电子表格、数据库等一起工作。例如，电子表格用户可具有相机型号列表，或者数据库用户可具有公司表格。用户可能想要与一个实体或多个实体有关的附加信息。

[0003] 作为示例，用户可能需要填充附加信息以完成任务。作为更具体的示例，为了帮助作出决策，用户可被分配通过填充每个相机型号的各个属性（诸如品牌、分辨率、价格和光学变焦）来扩充包括相机型号的电子表格的任务。当今为了完成这样的实体扩充任务，用户手动地尝试寻找包括所需信息的 web 源并将找到的对应数据值与现有的数据合并来组装完整的数据集。

[0004] 也通常会产生与扩充数据有关的其他任务。由此一般地，用户可从协助用户执行这样的任务的自动化解决方案中受益。然而，现有的方式一般在以下方面不是令人满意的：它们的数据精确水平（扩充的数据经常是错的）以及回调（扩充的数据经常不能被找到，例如，由于差的覆盖）。由此，对这些和其他这样的任务提供在合理的所需程度上对用户进行帮助的自动化解决方案是合乎需要的。

发明内容

[0005] 提供本概述以便以简化形式介绍将在以下的详细描述中进一步描述的一些代表性概念。本概述不旨在标识出所要求保护的主题的关键特征或必要特征，也不旨在以限制所要求保护的主题的范围的任何方式来使用。

[0006] 简而言之，在此描述的主题的各方面涉及通过其基于从至少一个语料库中挖掘的实体和属性之间的直接和间接关系来处理扩充任务的技术。访问被处理成索引的基于关系的数据来寻找完成扩充任务的数据。

[0007] 示例扩充任务包括对被扩充的数据的查询，诸如其中查询标识实体集（或更多实体）和属性名称的一个查询，并且其中使用基于关系的数据来完成扩充任务包括查找针对该实体集的每一实体的至少一个属性值，和 / 或其中查询标识实体集和一个或多个属性值示例的一个查询，并且其中使用基于关系的数据来 完成扩充任务包括基于该一个或多个属性示例来查找针对该实体集的至少一个实体的至少一个属性值。另一任务查询同义词数据，其中查询标识实体名称，并且其中基于关系的数据被查找来寻找针对属性名称的至少一个同义词。

[0008] 在一个方面，服务被配置成处理与领域无关的实体扩充任务，包括将从至少一个语料库获得的关系表格预处理成多个索引。索引包括基于表格的至少一些之间的间接映射的数据。服务用经由索引获得的数据对对应于实体扩充任务的查询进行响应，包括经由索引标识种子表格，计算每个种子表格的分数，基于种子表格的矢量来计算偏好矢量，基于偏好矢量和与关系表格中的至少一些相关联的矢量来计算预测分数，聚集分数，并基于对分数的聚集来返回最终预测以完成实体扩充任务来对查询进行响应。

[0009] 结合附图阅读以下具体实施方式，本发明的其他优点会变得显而易见。

附图说明

[0010] 作为示例而非限制，在附图中示出了本发明，附图中相同的附图标记指示相同或相似的元素，附图中：

[0011] 图 1 是表示根据一个示例实现的实体属性扩充服务的示例组件的框图。

[0012] 图 2 是表示根据一个示例实现的在实体属性扩充中使用的示例查询表格、种子表格和相关 web 表格之间映射的框图。

[0013] 图 3A-3C 包括根据一个示例实现的由实体属性扩充服务完成的实体属性扩充任务的表示。

[0014] 图 4 是根据一个示例实现的在实体属性扩充中使用的示例查询表格以及各表格之间的直接和间接映射的表示。

[0015] 图 5 是根据一示例实现的显示用于提供实体属性扩充服务的示例离线和在线组件的框 / 流程图。

[0016] 图 6 是表示其中可实现在此处所描述的各实施例的一个或多个方面的示例性、非限制性计算系统或操作环境的框图，例如在移动电话设备的例子中。

具体实施方式

[0017] 在此描述的技术的各个方面一般涉及自动实体扩充服务，该自动实体扩充服务对于被扩充的数据具有相对高的精确度和覆盖 / 回调以及快速的（例如交互的）响应时间。实体扩充技术可被应用于任何任意实体领域。

[0018] 一般而言，服务是基于从非常大量的数据源（例如，web 表格语料库）中收集的数据。例如，存在可从 web 爬寻的大约数以亿计量级的这样的实体 - 属性表格（也称为关系表格和二维表格）。在此描述了可基于主题敏感页面排名的整体匹配框架和聚集来自多个匹配的表格的预测的扩充框架，该多个匹配的表格除了直接匹配表格之外还包括间接匹配表格。这使得预测对虚假匹配的表格更加鲁棒。

[0019] 应当理解的是，此处的任何示例均是非限制的。例如，尽管描述了具有实体 - 属性表格形式的关系表格，但是其他数据结构并且甚至是未被结构化的数据也可被处理来得到与扩充有关的数据。此外，尽管 web 是这样的数据的一个源，但是可访问诸如企业数据库、以主题为中心的语料库（例如，与医疗有关的源、与金融有关的源等）等其他源来获得与实体有关的数据。

[0020] 因此，本发明不限制于在此描述的任何具体的实施例、方面、概念、结构、功能或示例。相反，此处所描述的实施例、方面、概念、结构、功能或示例中的任一个都是非限制性的，并且本发明一般能够以在计算和实体扩充方面提供好处和优点的各种方式来使用。

[0021] 图 1 是示出一个示例实现中各种组件的框图。一般地，在可被离线执行的表格预处理阶段 102 中，包括爬寻器 104，该爬寻器 104 从数据源 108（例如，网页和 / 或数据市场）提取关系表格 106。对于网页，爬寻器 104 可对 HTML 和 / 或文本数据是否实际上是表格并且不仅仅被用于页面的格式化或布局进行分类。各种技术可被用于滤除格式化表格和看上去本质上不相关的其他表格，由此图 1 显示了爬寻器 / 过滤器。取决于大量的网页，映射 - 减少 (map-reduce) 计算体系结构可被用于跨大量计算节点并行地提取表格。

[0022] 索引器 110 输入由爬寻器 104 产生的表格 106 并对它们进行索引（经索引的表格 112）以实现对那些类似于给定查询表格的 web 表格的快速标识。索引器 110 返回具有与查询表格记录的充分记录重叠的 web 表格。该重叠可允许模糊匹配以允许不同的值表示，例如，查询记录中的“Microsft Corp”（在此有意地拼错以表示实际的用户错误）可被考虑为与特定 web 表格中的“Microsoft Corporation”的匹配。

[0023] 图构建器 114 生成图 116，其中图 116 的顶点对应于每个 web 表格并且边 被加权以对应于两个表格的相似性。在一个示例实现中，使用基于特征的测量来计算表格相似性，该基于特征的测量合并了两个表格的许多特征，包括：记录重叠的程度、列名相似性、在其中出现表格的周围网页上下文的相似性、URL 相似性、网站的文档和域的预先计算的静态页面排名 (PageRank) 值以及每个表格中单词集之间的单词包相似性。这个测量可使用在已经被知晓为高度相似的表格上进行训练的模型 118 来计算。注意，通过寻找具有与两个其他表格的高度重叠的桥梁表格并将这其他两个表格用作正面训练示例，模型训练数据收集过程可被执行到不需要用户标记的数据的程度。

[0024] 图分析器 120 处理表格图 116 并生成每对 web 表格之间的相似性分数 122。一个方式传播表格之间的相似性来为每个 web 表格计算其他表格的个性化的页面排名 (PPR)，例如被结构化为矩阵。一般地，这个方式传播表格之间的本地成对相似性并允许具有许多共同的邻居但不直接链接的两个表格具有更高的相似性。由于完整的查询记录可能不具有与扩充不完整的查询记录所需要的表格的直接重叠，这个属性在数据扩充的上下文中是有用的。

[0025] 作为示例，图 2 例示了在其中对查询表格 220 的关系被传播以获得合乎需要的预测和覆盖的整体方式。该整体方式允许与查询表格不共享直接重叠或其他直接关系的表格对扩充任务作出贡献。在一个实现中，扩充服务利用由图分析器 120 产生的 PPR 矩阵。服务 126 接收来自客户端的扩充任务请求（例如，经由 API 集等），执行实体扩充处理来完成任务并返回经扩充的结果。

[0026] 在一个方面，空值填充被执行为一个任务，其中输入是可能具有缺失的属性值的实体名称表格。一个步骤寻找具有与输入表格共同的完整记录的 web 表格，其在此被称为“种子”表格。图 2 显示了示例种子表格 221-223。计算输入表格 220 和每个种子表格 221-223 之间的相似性强度并形成矢量。PPR 矩阵乘以这个矢量来产生最终矢量，该最终矢量包括针对给定输入表格的每个 web 表格的分数。web 表格接着被用于通过以下方式预测缺失的属性值：通过取得与每个 web 表格记录相关联的与非缺失的输入属性值相匹配的属性值，并聚集 web 表格排名来达到每个值的最终概率。最高排名的值，任选的高于用户指定的阈值，被用于填充空值。

[0027] 如果输入或 web 表格包括多于两个的列，则过程可尝试标识类似键 (key-like) 的列并接着将表格划分成若干个两列表格。

[0028] 其他可能的扩充任务包括通过属性名称的扩充，其类似于空值填充，除了属性名称与表格一起提供。在通过属性名称的扩充中，目标属性值被当作空值并且用户提供目标属性名称。利用与用于空值填充的过程相同的过程来相对于输入表格对 web 表格进行排名。种子表格具有与输入表格的重叠并且与目标属性名称匹配。预测过程与用于空值填充的过程相同。

[0029] 另一任务是数据确认，其类似于空值填充，除了如果头个经预测的值不匹配给定输入记录中的实际值，则产生确认警报。另一可能的扩充任务是通过值的属性扩展（通过示例的扩充），其基于与表格一起提供的几个种子示例值来填充目标属性的空值。

[0030] 属性建议是另一任务，其取得输入表格并建议表格可被其扩展的前 K 个最感兴趣的属性。该过程通过首先相对于输入表格的键列对 web 表格进行排名来与通过名称的属性扩展类似地进行。通过利用 PPR 矩阵，较高排名的 web 表格被群集到适当的团集中。每个群集包括彼此相互类似的 web 表格。由于在一个实现中，每个 web 表格已经被转换到一个或多个两列表格中，所以可从每个表格群集中取得最频繁的非输入匹配列名以导出该群集的名称。对群集进行排名并将前 K 个群集名称返回到用户。

[0031] 考虑正在搜索产品或股票的用户或正在执行竞争者分析的分析者。这种任务的最费力的子任务之一是聚集关于感兴趣实体的信息。两个这种子任务包括寻找一个或多个实体的属性值，以及寻找实体类型的相关属性。这些子任务基于提取的 web 表格是自动化的。在一个实现中，这些子任务可使用在此描述的操作，即通过属性名称的扩充、通过示例的扩充和属性发现。可提供其他操作。

[0032] 通过属性名称的扩充用在此被称为扩充属性的值 / 属性来自动化上述的示例任务，例如，在给定相机型号的情况下，寻找诸如品牌、分辨率、价格和光学变焦等各种属性的值。图 3 显示了用于这个操作的被应用到具有一个扩充属性（品牌）的一些示例型号实体的示例输入和输出。

[0033] 图 3B 例示了通过示例的扩充，其提供缺失的实体的一个或多个扩充属性的值，而非提供一个或多个扩充属性的名称。如可以看到的，从提供的已知示例中确定实体 - 属性关系，该实体 - 属性关系促进定位缺失的属性值。

[0034] 图 3C 中例示了属性发现。考虑用户可能不知晓关于实体域的足够信息；在这种情况下，用户想要知道给定实体集的最相关属性，例如来选择特定的所需的一些属性并请求针对这些所选择的属性的扩充值。通过使用自动地确定相关属性的服务，用户节省了用于尝试手动发现它们的时间和努力。

[0035] 为了提供更加有用的服务，分别对于实体被正确地扩充、被扩充实体的数量以及实体的数量，高精确度和高覆盖是所需要的。合乎需要的服务还提供快速的（例如，交互的）响应时间并应用到任何任意领域的实体。

[0036] 为简明起见，在示例中仅考虑一个扩充属性，其中合适的扩展是直截了当的。如图 3A 中显示的，输入可被视为二元关系，其中第一列对应于实体名称而第二列对应于扩充属性。第一列可用要被扩充的实体名称来填充，而第二列是空的。这个表格在此被称为查询表格（或简单地称为查询）。基线技术使用模式匹配技术，例如使用一对一映射，来标识语义上与查询表格匹配的 web 表格。在 web 表格中查找每个实体以获得它对扩充属性的值。

[0037] 考虑查询表格 Q（一个示例查询表格在图 4 中被标记为 440）。为简明起见，在这个示例中，考虑类似于查询表格，web 表格是类似地两列实体 - 属性二元关系，其中第一列对应于实体名称而第二列对应于实体的属性。通过使用传统的模式匹配技术，如果第一列中的数据值与查询表格 Q 的第一列中的那些数据值重叠并且第二列的名称与扩充属性的名称一致，则 web 表格匹配于查询表格 Q。这样的匹配在此被称为“直接匹配”并且该方式被称为“直接匹配方式”。

[0038] 在图 4 中，只有 web 表格 441-443 与查询表格 440 直接匹配（使用实线箭头显示）。分数可基于值重叠的程度和列名称匹配的程度与每个直接匹配相关联；在图中通过靠近箭头的值来显示示例分数。如果仅使用直接匹配，则简单地查找 web 表格 441-443 中的实体；对于型号“D3150”，web 表格 441 和 443 都包括它，然而，值是不同的（分别为“ABCDCo”和“NGD”）。可任意地选择或从具有较高分数的 web 表格中选择值，即，来自 web 表格 443 的“NGD”。对于型号“S-456”，可选择“WXYZCo”或“EFGcorp”，因为它们具有相等的分数。对于“N444”，只有“WXYZCo”。查找不能扩充 V199，因为没有匹配的表格包括该实体。

[0039] 由此可容易地理解，直接匹配通常遭受低的精确度；考虑例如表格 443 可 包括蜂窝电话型号和品牌而非所需的实体（诸如相机）。表格 443 中的蜂窝电话型号中的一些的名称与在查询表格 440 中的相机型号的名称相同，从而表格 443 得到高的分数。这导致了（三个中的）至少一个并且可能两个（如果当解决冲突时，从表格 443 中选择，则三个中的两个）错误的扩充。这种实体名称的模糊性实际上在所有领域中都存在，而这个可通过提升匹配的阈值来缓减，但这么做导致差的覆盖。

[0040] 使用仅直接匹配技术的另一问题是低覆盖；以上示例，过程没能扩充 V199，并且覆盖由此是百分之七十五。注意，这个数量比实践中低的多，尤其对于尾域，并且趋向于发生，这是因为可提供理想值的表格要么不具有列名称要么不使用与用户提供的扩充属性名称相同的列名称。扩充属性的同义词可有所帮助，但是这些同义词是手动生成的（自动生成导致差的质量），这在实体可能来自任何任意领域的情况下不是可行的。

[0041] 在此描述的是对例如经由其他 web 表格间接匹配查询表格的表格的进一步使用。通过使用这样的间接匹配表格，结合直接匹配表格，一般提升覆盖和精确度两者。作为提升的覆盖的示例，在图 4 中，表格 444 包括 V199 的理想属性值 (NGD)，但它不能使用仅直接匹配到达。通过使用模式匹配技术，表格 444 与表格 441（即，存在两个关系的两个属性之间的一对一映射）以及表格 442 匹配（由于它具有与表格 442 共同的两个记录以及与表格 442 共同的一个记录）。这样的在 web 表格之中的模式匹配被虚线箭头表示；每个这样的匹配具有表示匹配程度的分数。由于表格 441 和 / 或表格 442（大致）与查询表格 440 匹配（使用直接匹配）并且表格 444（大致）与表格 441 和表格 442 匹配（使用模式匹配），可以得出表格 444（大致）与查询表格 440 匹配。表格 444 在此称为间接匹配表格；通过使用它，V199 可被正确地扩充。在这个示例中，覆盖经由间接匹配被提升到百分之百。

[0042] 然而，许多间接匹配表格包括虚假的匹配，从而使用这些表格来预测值导致了错误的预测。为了对这种虚假匹配的鲁棒性，在此描述了基于以下观察来使用整体匹配：真正匹配的表格要么直接地要么间接地相互匹配并且要么直接地要么间接地与直接匹配表格匹配，而虚假匹配的表格则不是。例如，表格 441、442 和 444 相互直接匹配，而表格 444 仅微弱地与表格 442 匹配。如果，例如， 通过聚集直接匹配以及间接匹配来计算表格的整体匹配分数，则真正匹配的表格收到更高的分数，其是用于在此描述的整体匹配的基础。在图 4 的示例中，相比于表格 443，表格 441、442 和 444 得到更高的分数；这导致通过不使用表格 443 的正确扩充，得到了百分之百的精确度。此外，对于每个实体，预测可从多个匹配的表格中获得并聚集，从而“前”一个（或 k 个）值可被选择为最终所预测的一个（或多个）值。

[0043] 注意在实践中，这导致技术挑战，例如使用 573M×573M 表格对来计算 web 表格对之间的模式匹配 (web 表格或 SMW 图之中的模式匹配）需要是精确的。此外，整体匹配需要

被建模，使得模型将与 SMW 图中的边相关联的分数以及与直接匹配相关联的那些分数考虑在内。此外，实体需要在查询时被高效地扩充。

[0044] 为此，提供了基于对图的主题敏感页面排名 (TSP) 的整体匹配框架。还提供了充分利用预处理（例如，在 MapReduce（映射减少）技术中）来实现查询时的极度快速的（交互的）响应时间的系统体系架构。

[0045] 在一个实现中，基于匹配学习的技术被用于基于使用与 web 表格相关联的特征（例如，包括文本）来确定是否两个 web 表格匹配来构建 SMW 图。此外，虽然可使用手动产生的经标记的数据，但是代替于或附加于手动产生的经标记的数据，用于该学习任务的训练数据可如在此描述的那样被自动地生成。

[0046] 转向示例整体匹配框架和数据模型，为简明起见，假设查询表格是实体 - 属性二元关系，例如，查询表格 Q 具有 Q(K,A) 的形式，其中 K 表示实体名称属性（在此也称为查询表格“键”）而 A 是扩充属性。如在图 4 的示例查询表格 440 中显示的，键列被填充而属性列是空的。此外，假设 web 表格也是实体 - 属性二元关系，如在图 4 的 web 表格 441-444 中。

[0047] 对于每个 web 表格 $T \in \mathcal{T}$，关系是 $T_R(K,B)$，其中 K 表示实体名称属性（在此称为 web 表格键属性）并且 B 是实体的属性，从中提取表格的网页的 URL T_U 以及从中提取表格的网页中的它的上下文 T_C（例如，表格周围的文本）。为简明起见，当从上下文中清晰时，$T_R(K,B)$ 可被表示为 T(K,B)。

[0048] 对于通过属性名称的扩充，给定查询表格 Q(K,A) 和 web 表格集合 $(T(A,B), T_U, T_C \in \mathcal{T}$，操作是要预测每个查询记录 $q \in Q$ 在属性 A 上的值。注意，不是所有 web 表格都具有实体 - 属性二元关系，然而在此描述的框架被通用于 n 元 web 表格。此外，查询表格可具有多于一个的扩充属性；在一个实现中，属性被考虑为独立的并且一次可针对一个属性执行预测。

[0049] 在一个实现中，通用扩充框架标识“匹配”查询表格的 web 表格，并使用每个匹配的 web 表格来提供对于恰好在查询和 web 表格之间重叠的特定键的值预测。为了标识匹配的表格，一般而言，如果 Q.K 和 T.K 指代相同的实体类型并且 Q.A 和 Q.B 指代实体的相同属性，则 web 表格 T(K,B) 匹配查询表格 Q(K,A)；（为简明起见，描述了一对一映射）。每个 web 表格被分配表示对于查询表格 Q 的匹配分数的分数 S(Q,T)；由于 Q 是固定的，所以标记可被表示为 S(T)。存在各种获得查询表格和 web 表格之间匹配分数的方式；以下描述示例。

[0050] 为了预测值，对于每个记录，通过将查询表格 Q(K,A) 与每个匹配的 web 表格 T(K,B) 在键属性 K 上联结来预测来自匹配的 web 表格的记录 q 在属性 Q.A 上的值 q[Q.A]。如果存在记录 $t \in T$ 使得 $q[Q.K] \approx t[T.K]$（其中≈表示要么值的精确相等要么值的近似相等），则 web 表格 T 对 q[Q.A] 预测了具有预测分数 $S_T(v) = S(T)$ 的值 $v = t[T.B]$，并且 $(v, S_T(v))$ 被返回。在处理了匹配的表格后，存在针对 q[Q.A] 的所预测的值连同它们对应的预测分数的集合 $P_q = \{(x_1, S_{T1}(x_1)), (x_2, S_{T2}(x_2)), \ldots\}$。对于每个不同的所预测的值 $v \in P_q$，通过聚集针对 v 获得的预测分数来计算最终预测分数：

[0051]

$$S(v) = \mathop{\mathcal{F}}_{(x_i, S_{T_i}(x_i)) \in \mathcal{P}_q | x_i \approx v} S_{T_i}(x_i) \qquad (1)$$

[0052] 其中$\mathcal{F}$是聚集函数。可在这个框架中使用替换的聚集函数，诸如求和 (sum)、取最大 (max) 或取最大的前 d 个 (max_top_d)。针对 q[Q.A] 的最终所预测的值是具有最高最终预测分数的一个值：

[0053] $$q[Q.A] = \underset{v}{\operatorname{argmax}} \quad S(v) \tag{2}$$

[0054] 在对实体扩充 k 个值的实现中（例如，实体是音乐家，并且所需要的是用他或她的专辑来扩充音乐家姓名），选择 k 个最高最终预测分数。由此，在图 4 的示例中，通过使用显示的表格匹配分数，对于查询记录 D3150 而言，P_q = {(ABCDCo, 0:25), (NGD, 0:5)}（分别由表格 441 和 443 预测）。由此，最终所预测的值是分别具有分数 0:25 和 0:5 的 ABCDCo 和 NGD，所以所预测的值是 NGD。

[0055] 一种计算匹配的方法是如以上描述的直接匹配方式，例如，通过使用传统的模式匹配技术，直接匹配当且仅当在 T.K 中的数据值与 Q.K 中的那些数据值重叠并且 T.B ≈ Q.A 的情况下才认为 web 表格 T 与查询表格 Q 匹配。直接匹配将 Q 和 T 之间的匹配分数 $S_{DMA}(T)$ 计算为：

[0056] $$S_{DMA}(T) = \begin{cases} \frac{|T \cap_K Q|}{\min(|Q|,|T|)} & \text{如果} Q.A \approx T.B \\ 0 & \text{其他} \end{cases} \tag{3}$$

[0057] 其中$|T \cap_K Q| = |\{t | t \in T \,\&\, \exists q \in Q \; s.t. \; T[T.K] \approx q[Q.K]\}|$。例如，在图 4 中，表格 441、442 和 443 的分数分别是$\frac{1}{4}$,$\frac{2}{4}$和$\frac{2}{4}$，这是因为它们分别具有 1 个、2 个和 2 个匹配键，min(|Q|, |T|) = 4 并且 Q.A ≈ T.B；

[0058] 表格 444 的分数是 0，因为 Q.A 不近似等于 T.B。预测步骤与以上描述的在通用扩充框架中的步骤一致。

[0059] 转到整体匹配方式，考虑加权有向图 G(V,E)，其中边上的权重被 $\alpha_{u,v}$ 表示为 (u,v) ∈ E。页面排名 (pagerank) 是在 G 上的随机走查的静止分布，在每一步，使用概率 ε（通常称为传输概率）跳到随机节点，并且使用概率 (1-ε)，沿着来自当前节点的随机传出边。个性化的页面排名 (PPR) 与页面排名相同，除了随机跳在返回到被表示为“源”节点的相同节点时完成，针对该“源”节点页面排名被个性化。正式地，相对于源节点 u 的节点 v 的被 $\Pi_u(v)$ 表示的 PPR 被定义为以下等式的求解：

[0060] $$\pi_u(v) = \epsilon \delta_u(v) + (1-\epsilon) \sum_{\{w|(w,v)\in E\}} \pi_u(w) \alpha_{w,v} \tag{4}$$

[0061] 其中当且仅当 u = v 时 $\delta_u(v)$ = 1，否则为 0。节点 v ∈ V 相对于 u 的 PPR 值 $\Pi_u(v)$ 在此被称为 u 的 PPR 矢量。“主题”被定义为导出对 V 的概率分布的偏好矢量$\vec{B}$；对于节点 v ∈ V 的$\vec{B}$的值被表示为 B_v。主题敏感页面排名 (TSP) 与页面排名相同，除了随机跳在返回到用概率 B_u 选择的具有 B_u>0 的节点 u 之一时完成。正式地，针对主题$\vec{B}$的节点 v 的 TSP 被定义为以下等式的求解：

[0062] $$\pi_{\vec{\beta}}(v) = \epsilon\vec{\beta} + (1-\epsilon)\sum_{\{w|(w,v)\in E\}} \pi_{\vec{\beta}}(w)\alpha_{w,v} \quad (5)$$

[0063] 关于使用TSP对整体匹配进行建模，在两个web表格之间的整体匹配和节 点相对于源节点的PPR之间绘制连接。在给定web表格集合的情况下，构建加权的SMW图G(V,E)，其中每个节点v ∈ V对应于web表格，并且每个边(u,v) ∈ E表示对应于u和v的web表格之间的语义匹配关系。每个边(u,v) ∈ E具有表示web表格u和v之间匹配程度（由模式匹配技术提供）的权重 $\alpha_{u,v}$。

[0064] 考虑任意加权有向图G(V,E)的两个节点u和v。v相对于u的PPR $\Pi_u(v)$ 表示v对v的整体关系，其中E表示直接的成对的关系，例如，它考虑从u到v的所有直接以及间接的路径并聚集它们的分数来计算整体分数。当直接的成对的关系是网页之间的超链接时，$\Pi_u(v)$ 是来自u的对v的整体重要性授予（经由超链接）。替换地，当直接的成对的关系是社交网络中的直接朋友关系时，$\Pi_u(v)$ 是来自u的对v的整体朋友关系。在此描述的是将这样的技术应用到SMW图。在此E表示直接的、成对的语义匹配关系，例如对SMW图上的v相对于u的 $\Pi_u(v)$ 的PPR对v对u的整体语义匹配进行建模。如果u是查询表格，则 $S_{Hol}(T) = \Pi_u(v)$，其中v对应于T。注意，查询表格Q通常不与web表格一致；然而，通过将Q考虑成“主题”和模型，如对应于T的节点v的TSP的匹配对应于主题。在其中关系是重要授予的web上下文中，关于主题的最重要的页面被用于对主题进行建模（例如，在开放目录项目中包括在该主题下的那些）；在其中关系是语义匹配的这个上下文中，排名靠前的那些匹配的表格被用于对Q的主题进行建模。与Q直接匹配的web表格集合S（在此被称为种子表格）被用于对它进行建模，即 $S = \{T | S_{DMA}(T) > 0\}$。此外，直接匹配分数 $S_{DMA}(T) | T \in S$ 可被用作偏好值 $\vec{\beta}$：

[0065] $$\beta_v = \begin{cases} \frac{S_{DMA}(T)}{\sum_{T\in S} S_{DMA}(T)} & \text{如果} T \in S \\ 0 & \text{其他} \end{cases} \quad (6)$$

[0066] 其中v对应于T。例如，B_v 针对表格441、442和443分别是 $\frac{0.25}{1.25}$、$\frac{0.5}{1.25}$ 和 $\frac{0.5}{1.25}$，针对所有其他表格是0。由于网页的TSP分数表示整体计算的页面对于主题的重要性，在SMW图上v相对于以上主题 $\vec{\beta}$ 的 $\Pi_{\vec{\beta}}(v)$ 对v对Q的整体语义匹配进行建模。由此，可使用 $S_{Hol}(T) = \Pi_{\vec{\beta}}(v)$，其中v对应于T。

[0067] 考虑SMW已经被提前构建。一种为每个web表格计算整体匹配分数 $S_{Hol}(T)$ 的方法是在扩充时对G运行TSP计算算法，然而这导致极其高的响应时间。相 反，通过相对于SMW图中的每个其他节点来预先计算每个节点的PPR（在此称为完全个性化页面排名(FPPR)计算），可针对任意查询表格高效地计算整体匹配分数，其产生在查询时非常快速的响应时间。

[0068] 图5提供了具有框和流程图形式的类似于关于图1描述的体系架构的附加细节。离线预处理可包括web爬寻550来提取web表格552，构建SMW图（框554和556）和为每

个 web 表格计算 FPPR 558(完全个性化的页面排名),例如,缩放到大约数以亿计表格的量级;映射减少 (MapReduce) 框架可被用于此目的。如以上描述的,基于已知技术,离线部分从 web 爬寻中提取 web 表格并使用分类器来从其他类型的 web 表格中区分关系 web 表格,(例如格式化表格、实体属性表格等)。此外,web 表格被索引来促进对表格更快速的访问。一种实现使用两个索引(框 560),包括对 web 表格的键属性的索引 (WIK)。给定查询表格 Q,WIK(Q) 返回在各键的至少一个上与 Q 重叠的 web 表格集合。第二个包括对于 web 表格完整记录的索引(那是组合的键和值)(WIKV)。WIKV(Q) 返回包括来自 Q 的至少一个记录的 web 表格的集合。

[0069] 离线部分还基于在此描述的模式匹配技术来构建 SMW 图,计算 FPPR 并存储每个 web 表格的 PPR 矢量(在一个实现中仅存储非零项),在此被称为 T2PPV 索引 562。对于任意 web 表格 T,如也在此描述的,T2PPV(T) 返回 T 的 PPR 矢量。框 560 表示为每个 web 表格 T(K,B) 发现属性 B 的同义词,其也在以下与属性发现操作一起描述的,并在此被称为 T2Syn 索引 566。对于任意 web 表格 T,T2Syn(T) 返回表格 T 的属性 B 的同义词。索引 (WIK、WIKV、T2PPV 和 T2Syn) 可例如要么是磁盘驻留的要么是驻留在存储器中以供更快速的访问的。

[0070] 给定查询表格 568,在线查询时间处理部分计算 web 表格的 TSP 分数 570 并聚集来自 web 表格的预测 572。查询时间处理可在各个示例操作中被抽象。一个操作步骤通过充分利用 WIK 和 WIKV 索引(框 560) 来标识种子表格并计算它们的直接匹配分数来标识种子表格。另一操作步骤通过使用偏好矢量和每个表格的存储的 PPR 矢量来计算每个 web 表格的 TSP 分数来计算 TSP 分数 570,例如通过插入等式 (6) 中的直接匹配分数来计算偏好矢量。注意,只有种子表格具有矢量中的非零项,并且由此只有使用 T2PPV 索引的种子表格的 PPR 矢量需要被检索。注意,不需要对所有 web 表格计算 TSP 分数,只对可被用在聚集步骤中的表格计算 TSP 分数,例如,具有至少一个与查询表格重叠的键的表格,在此被称为关系表格;这些可通过调用 WIK(Q) 来高效地标识。

[0071] 另一操作步骤通过收集由相关 web 表格 T 提供的预测以及分数 $S_{Hol}(T)$ 来聚集并选择值。根据该操作,预测 572 被处理,分数被聚集并且最终预测被选择。

[0072] 转向构建 SMW 图和计算 FPPR,在此描述了匹配一对 web 表格以及可缩放性挑战。在 SMW 图中,如果 T 匹配于 T',则在一对 web 表格 (T(K,B)),(T'(K',B')) 之间有一个边,即 T.K 和 T.K' 指代相同类型的实体并且 T.B 和 T'.B' 指代那些实体的相同属性。模式匹配可被用于确定 T 是否匹配于 T'以及映射的分数,但仅在一定程度上,例如,相机和蜂窝电话的表格可以是使用模式匹配的匹配,并且没有找到间接匹配。

[0073] 在此描述了与网页中的 web 表格相关联的上下文(例如,文本),其趋向提供关于表格的有价值的信息。例如,如果表格 443 的上下文是“移动电话或蜂窝电话,范围是……”或“与……兼容的电话”,则关于表格的信息是可获得的。类似地,对于表格 442 和 444,文本可分别包括“用于相机和透镜的预先制好的纠正数据”和“相机包兼容性列表”,其指示表格 442 和 444 更可能关于相机,而表格 443 是关于电话的。

[0074] 在此描述了经由上下文相似性特征来捕捉基于上下文的概念,其可使用表格周围文本的 tf-idf 余弦相似性来计算(基于 IR 检索)。表格的上下文可包括与另一 web 表格内部的值重叠的关键词。这提供以下证据:包括表格的网页可能关于类似的主题并且由此,表格也可能关于类似的主题。可使用第一表格周围文本和第二表格内部的文本的 tf-idf

余弦相似性来计算表格对上下文相似性特征。

[0075] 此外，包括表格的网页的URL可有助于与另一表格进行匹配。例如，有时候网站在若干个网页中列出来自同一原始大的表格中的记录，例如，网站可在若干个网页中按年、按首字母等列出电影。在这种情况下，对网页的URL进行匹配是web表格的匹配的良好信号。可使用例如通过使用URL项的余弦相似性来计算的URL相似性特征等。

[0076] 以上相似性（以及可能的许多其他特征）可被用作分类模型中的特征。示 例特征包括：

[0077]

特征名称	描述/文档
上下文	web表格周围文本中具有idf权重的项。
表格对上下文	作为文本的表格内容和具有idf权重的上下文文本。
URL	URL中的具有从URL集合中计算的idf权重的项。
属性名称	在列名称提到的具有相等权重的项。
列值	列中来自文档的具有相等权重的不同值。
表格到表格	作为文本的具有对它们之间边的加权的 idf 权重的表格内容（类似于单词包）。
列宽	列宽方面的相似性。

[0078] 给定各特征，模型使用概率来预测两个表格之间的匹配，该概率被用作对它们之间的边的加权。注意，可针对非常大量（数以亿计）的web表格的平方来计算这些特征，其可能计算上过于昂贵。然而，在一个实现中，对于示例特征中的每一个，web表格可被考虑为单词包（或文档）。接着可充分利用用于在大的文档集合上计算成对文档的相似性的可缩放技术。可使用用于计算大的文档集合的文档相似性矩阵的已知技术，例如，使用映射减少(MapReduce)，例如，每个文档d包括项集合并可被表示为项权重Wt,d的矢量Wd。两个文档之间的相似性是项权重的内积；如果项t在两个文档中均出现，则该项t将对两个文档的相似性作出贡献。通过使用倒排索引I，获得包括特定项t的文档I(t)是直截了当的。通过处理项，计算完整的相似性矩阵。

[0079] 这可被直接实现为映射减少(MapReduce)任务，包括索引，其中映射器处理每个文档并针对每个项发出，并且减少器输出作为键的项以及包括该键的文档的列表。另一任务是相似性计算，其中映射器用它的文档列表处理每个项并对每对文档发出相似性分数。减少器执行求和来输出分数。为了更加高效，df-切(df-cut)概念可被用于消除具有高文档频率的项。

[0080] 如以上描述的，可自动地获得用于训练分类器的经标记的示例。为了将一对web表格(T,T')标记为正面示例，即使T和T'不具有共同的记录，第三web表格T"可包括分别与T和T'共同的一些记录（在此称为标记web表格）。例如，考虑图4中的表格442和444。表格4411被发现为是针对它们的标记web 表格；表格441与442以及与表格444重

叠。如果没有找到这样的标记 web 表格，则特征矢量被考虑为负面示例。

[0081] 一旦 SMW 图被构造，就可例如经由线性代数技术（诸如幂迭代或蒙特卡洛 (Monte Carlo)）来计算完全个性化的页面排名矩阵，其中基本思想是通过直接模拟对应的随机走查并接着使用所执行的走查的经验分布以估计静止分布来近似个性化的页面排名。映射减少 (MapReduce) 算法可被用于计算 FPPR，其是基于 Monte Carlo 方式的。一般地，在图中的每个节点处开始的给定长度的单随机走查被高效地计算，并且接着这些随机走查被用于高效地计算每个节点的 PPR 矢量。

[0082] 以上已经描述的通过属性的扩充，并且通过属性的扩充包括标识种子表格，例如，使用 WIK 索引来标识 Q(K,A) 的种子表格使得在 A ≈ B 的情况下 web 表格 T(K,B) 被考虑。可使用等式 (3) 来计算直接映射分数，并且以上描述了计算表格 TSP 分数。聚集和处理值可使用以上描述的“预测值”步骤；如在以下算法中阐述的：

[0083]

算法 1 扩充_ABA(查询表格 $Q(K,A)$)

1: $\forall q \in Q$, $\mathcal{P}_q = \{\}$
2: $\mathcal{R} = WIK(Q)$. { 相关web表格。 }
3: **for all** $T \in \mathcal{R}$ **do**
4: **for all** $q \in Q$ **and** $t \in T$ **s.t.** $q[Q.K] = t[T.K]$ **do**
5: $\mathcal{P}_q = \mathcal{P}_q \cup \{(\ v = t[T.B]\ ,\ S_T(v) = S(T)\ )\}$
6: **for all** $q \in Q$ **do**
7: $\forall\, v \in \mathcal{P}_q,\ \ S(v) = \mathcal{F}_{(x_t, S_{T_t}(x_t)) \in \mathcal{P}_q | x_t = v}\ \ S_{T_t}(x_t)$
8: $\forall\, q \in Q,\ \ q[A] = q[Q.A] = \mathrm{argmax}_v\ \ S(v)$

[0084] 这个在线聚集过程可分布在多个服务器机器上，每个服务器机器服务于在预处理期间生成的索引的一部分。

[0085] 通过示例的扩充 (ABE) 是通过属性的扩充操作的变型。替代于提供扩充属性的名称，用户向查询表格提供一些完整的记录作为示例，即，对于键中的一些，提供对扩充属性的值。以上描述的用于使用示例化的体系结构的步骤可以与通过属性的扩充操作的步骤一致，除了种子表格被标识以及直接映射分数被计算的方式。直接映射在当且仅当记录 Q_c 与 T 中的那些记录重叠的情况下认为 web 表格 T 匹配于查询表格 Q。例如，在图 4 中，表格 441 被认为是在图 3B 中示出的查询表格 440 的种子表格，因为它们在记录 (D3150,ABCDCo) 上重叠。给定查询表格，WIKV 索引被用于直接取得种子表格。一般地，如果对于每个在 T 和 Q_c 之间共享的键，两个表格也在扩充属性上达成一致，则 web 表格 T 被分配高的直接映射分数。因此，直接映射将匹配分数计算为：

[0086] $$S_{DMA}(T) = \frac{|Q^c \cap_{KV} T|}{|Q^c \cap_K T|} \qquad (7)$$

[0087] 其中 $|Q^c \cap_{KV} T|$ 是查询表格 Q 和 web 表格 T 的完整记录之间的重叠记录的数量。输出匹配分数在 0 和 1 之间，其中 0 表示对于 Q^c 和 T 之间共享的每个键，表格没有就扩充属性达成一致，而 1 表示对于每个共享的键，对应的记录在扩充属性上匹配。还注意，这个测量可进一步用 $|Q^c \cap_{KV} T|$ 来缩放以将记录的重叠大小作为直接匹配的附加证据考虑在内。可使用各种技术（诸如语料库中的反记录频率）来对各个记录进行加权以增加共享特定记

录的两个表格的相对重要性。

[0088] 与每个估算的值一起，服务可返回关于所返回的值的附加元数据，包括似然性分数和标识被用于估算该值的源表格的URL的子集。这个元数据可自己在与查询表格相关的值上聚集。对该元数据的一个示例使用是使得能够实现寻找最相关于用户的查询表格的语料库表格的任务。这个操作可被认为是执行表格搜索，其返回对底层原始表格而非虚拟的、被整体聚集的表格的指针。这个操作对于在从外联网以及内联网源两者收集的语料库中定位结构化的数据是有用的。这个操作还使得用户能够查看被用于到达所估算的值的原始数据；这允许用户取得关于特定值被如何生成的更深的理解。分数元数据可被用户界面用于基于预测的置信度水平对所预测的值进行颜色编码(color-code)。用户界面还可选择针对给定实体和属性呈现前k个估算的值。

[0089] 示例性框架的另一操作是用显著(“重要”)属性的同义词来发现查询表格的显著(“重要”)属性(例如，图3C)。为此，如果给定每个web表格的同义词，群集可基于同义词之中重叠的概念来执行并可被用于在属性和它们的同义词之间作出区分。为了发现每个web表格的同义词，可使用针对查询表格遵循的相同步骤。对于web表格T(K,B)，过程可标识匹配表格并对每个web表格T’(K’,B’)分配相对于T的匹配分数S(T’)。接着，每个web表格T’使用对应的分数S(T’)来预测作为B的同义词的B’并返回(B’;S(T’))。这提供了预测集合$P_T = \{(B_1, S(T_1)), (B_2, S(T_2)), \ldots\}$。分数可被聚集(类似于等式(1))并且前d个(例如，20个)插入到同义词索引中的同义词作为T的同义词，即T2Syn(T)。

[0090] 直接映射方式在第一个步骤中标识匹配表格。在这种情况下，对于web表格T，只有对T直接连接的web表格被用在SMW图中来提供同义词。注意对于表格441，直接映射不能将“Make(制作)”和“Vendor(供应商)”报告为来自表格445和446的同义词，因为它们没有直接连接到表格441。还可能从直接映射中产生不准确的群集。

[0091] 如果不是整体方式被用于取得匹配表格，则表格445和446是可从表格441到达的并且由此445和446相对于表格441的PPR分数(即$S_{Hol}(445)$，$S_{Hol}(446)$)将是非零，并且表格445和446两者均对表格441的同义词集合作出贡献。对整体同义词进行群集一般提供两个群集(或同义词集合)，例如表示\Brand″(品牌)属性的一个以及表示\Resolution″(分辨率)属性的另一个，这对于相机领域而言是准确且有意义的。

[0092] 对于属性发现操作的处理包括标识种子表格，计算TSP分数(如以上一般描述的)以及对值进行聚集并处理。标识种子表格包括使用查询表格的键，使得DMA将种子表格标识为那些在键属性上与查询表格重叠的表格。因此，直接映射将Q和T之间的匹配分数计算为：

[0093] $$S_{DMA}(T) = \frac{|Q \cap_K T|}{\min\{|Q|, |T|\}} \qquad (8)$$

[0094] 例如在图4中，由于表格441与查询表格440重叠，所以表格441是种子表格。给定查询表格，WIK索引直接提供种子表格。

[0095] 对值进行聚集和处理在以下算法中被一般地描述：

[0096]

算法 2 属性_发现 (查询表格 Q, Web 表格)

1: $\mathcal{R}$ =WIK(Q) { 使用倒排索引的相关表格 }
2: AllSynSet = {}
3: for all $T \in \mathcal{R}$ do
4: AllSynSet = AllSynSet ∪ $(Syn(T), S(T))$
5: return Cluster(AllSynSet)

[0097] 对于属性发现操作，每个 web 表格 T 用 $S_{Hol}(T)$ 分数预测同义词集合 T2Syn(T) 并返回 (T2Syn(T), $S_{Hol}(T)$)。在处理了匹配的表格后，与一组同义词集合的对应的预测分数一起提供该组同义词集合。同义词集合接着被群集并且为每个群集聚集预测分数。返回群集的经排序的列表，其中每个群集表示具有其同义词的经发现的属性。

[0098] 对于群集而言，可使用标准的凝聚群集步骤，例如初始地每个同义词集合 T2Syn(T) 是一群集。web 表格的同义词集合包括每个项根据聚集的分数的权重，并且由此，余弦相似性可被用于计算两个群集之间的相似性。而且，当两个群集被合并时，它们的对应的 S_{Hol}（或预测）分数被聚集以返回群集的经排名的列表。

[0099] 在以上的示例中，使用了实体 - 属性二元关系。这提供了具有节点中单个分数的较简单的图，其能够将问题建模成 TSP 问题。使用 n 元 web 表格和节点中的单个分数，查询表格和 web 表格之间的匹配分数将不能表明 web 表格的哪个列是所需的扩充属性。然而，web 上的关系表格一般旨在用于人类解释并通常具有主题列，并且存在有效的试探法来标识 web 表格的主题，例如使用其中主题列名称具有高的搜索查询命中的 web 搜索日志，并且通常主题列出现在最左列。如果主题列能被标识，则这样的表格可被划分到若干个实体 - 属性而言关系中，即，主题列与其他列的每一个构成了实体 - 属性二元关系集合。一般而言，主题出现在单个列中并且多个列并不认为是主题。

[0100] 图 6 示出其上可实现例如图 1-6 的各示例和实现的合适的计算和联网环境 600 的示例。计算系统环境 600 只是合适计算环境的一个示例，而非意在暗示对本发明使用范围或功能有任何限制。也不应该将计算环境 600 解释为对示例操作环境 600 中示出的任一组件或其组合有任何依赖性或要求。

[0101] 本发明可用各种其他通用或专用计算系统环境或配置来操作。适用于本发明的公知计算系统、环境、和 / 或配置的示例包括但不限于 :个人计算机、服务器计算机、手持式或膝上型设备、平板设备、多处理器系统、基于微处理器的系统、机顶盒、可编程消费电子产品、网络 PC、微型计算机、大型计算机、包括任何以上系统或设备的分布式计算环境等等。

[0102] 本发明可在诸如程序模块等由计算机执行的计算机可执行指令的通用上下文中描述。一般而言，程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等等。本发明也可在其中由通过通信网络链接的远程处理设备执行任务的分布式计算环境中实现。在分布式计算环境中，程序模块可以位于包括存储器存储设备在内的本地和 / 或远程计算机存储介质中。

[0103] 参考图 6，用于实现本发明的各方面的示例系统可包括计算机 610 形式的通用计算设备。计算机 610 的组件可以包括但不限于 :处理单元 620、系统存储器 630 和将包括系统存储器在内的各种系统组件耦合至处理单元 620 的系统总线 621。系统总线 621 可以是若

干类型的总线结构中的任一种，包括使用各种总线体系结构中的任一种的存储器总线或存储器控制器、外围总线、以及局域总线。作为示例而非限制，这样的体系结构包括工业标准体系结构 (ISA) 总线、微通道体系结构 (MCA) 总线、增强型 ISA(EISA) 总线、视频电子技术标准协会 (VESA) 局部总线和外围部件互连 (PCI) 总线 (也称为夹层 (Mezzanine) 总线)。

[0104] 计算机 610 通常包括各种计算机可读介质。计算机可读介质可以是能由计算机 610 访问的任何可用介质，并包含易失性和非易失性介质以及可移动、不可移动介质。作为示例而非限制，计算机可读介质可包括计算机存储介质和通信介质。计算机存储介质包括以存储诸如计算机可读的指令、数据结构、程序模块或其他数据之类的信息的任何方法或技术实现的易失性和非易失性、可移动和不可移动介质。计算机存储介质包括，但不仅限于，RAM、ROM、EEPROM、闪存或其他存储器技术、CD-ROM、数字多功能盘 (DVD) 或其他光盘存储、磁带盒、磁带、磁盘存储或其他磁存储设备，或可以用来存储所需信息并可以被计算机 610 访问的任何其他介质。通信介质通常以诸如载波或其他传输机制之类的已调制数据信号来体现计算机可读指令、数据结构、程序模块或其他数据，并且包括任何信息传送介质。术语“已调制数据信号”是指使得以在信号中编码信息的方式来设置或改变其一个或多个特征的信号。作为示例而非限制，通信介质包括诸如有线网络或直接线连接之类的有线介质，以及诸如声学、RF、红外及其他无线介质之类的无线介质。上面各项中的任何项的组合也包括在计算机可读介质的范围内。

[0105] 系统存储器 630 包括易失性和 / 或非易失性存储器形式的计算机存储介质，如只读存储器 (ROM)631 和随机存取存储器 (RAM)632。包含诸如在启动期间帮助在计算机 610 内的元件之间传输信息的基本例程的基本输入 / 输出系统 633(BIOS) 通常存储在 ROM 631 中。RAM 632 通常包含处理单元 620 可立即访问和 / 或当前正在操作的数据和 / 或程序模块。作为示例而非限制，图 6 示出了操作系统 634、应用程序 635、其他程序模块 636 和程序数据 637。

[0106] 计算机 610 也可以包括其它可移动 / 不可移动、易失性 / 非易失性计算机存储介质。仅作为示例，图 6 示出了从不可移动、非易失性磁介质中读取或向其写入的硬盘驱动器 641，从可移动、非易失性磁盘 652 中读取或向其写入的磁盘驱动器 651，以及从诸如 CD ROM 或其他光学介质等可移动、非易失性光盘 656 中读取或向其写入的光盘驱动器 655。可以在该示例操作环境中使用的其它可移动 / 不可移动、易失性 / 非易失性计算机存储介质包括但不限于，磁带盒、闪存卡、数字多功能盘、数字录像带、固态 RAM、固态 ROM 等等。硬盘驱动器 641 通常通过诸如接口 640 之类的不可移动存储器接口连接到系统总线 621，并且磁盘驱动器 651 和光盘驱动器 655 通常通过诸如接口 650 之类的可移动存储器接口连接到系统总线 621。

[0107] 以上描述并在图 6 中示出的驱动器及其相关联的计算机存储介质为计算机 610 提供了对计算机可读指令、数据结构、程序模块和其他数据的存储。例如，在图 6 中，硬盘驱动器 641 被示为存储操作系统 644、应用程序 645、其他程序模块 646 和程序数据 647。注意，这些组件可与操作系统 634、应用程序 635、其它程序模块 636 和程序数据 637 相同，也可与它们不同。操作系统 644、应用程序 645、其他程序模块 646 和程序数据 647 在这里被标注了不同的附图标记是为了说明至少它们是不同的副本。用户可通过诸如平板或者电子数字化仪 664、话筒 663、键盘 662 和定点设备 661(通常指的是鼠标、跟踪球或触摸垫) 等输入

设备向计算机 610 输入命令和信息。图 6 中未示出的其他输入设备可以包括操纵杆、游戏手柄、圆盘式卫星天线、扫描仪等。这些以及其它输入设备通常通过耦合到系统总线的用户输入接口 660 连接到处理单元 620，但也可通过诸如并行端口、游戏端口或通用串行总线 (USB) 之类的其它接口和总线结构来连接。监视器 691 或其它类型的显示设备也经由诸如视频接口 690 之类的接口连接至系统总线 621。监视器 691 也可以与触摸屏面板等集成。注意到监视器和 / 或触摸屏面板可以在物理上耦合至其中包括计算设备 610 的外壳，诸如在平板型个人计算机中。此外，诸如计算设备 610 等计算机还可以包括其他外围输出设备，诸如扬声器 695 和打印机 696，它们可以通过输出外围接口 694 等连接。

[0108] 计算机 610 可使用到一个或多个远程计算机（诸如，远程计算机 680）的逻辑连接而在联网环境中操作。远程计算机 680 可以是个人计算机、服务器、路由器、网络 PC、对等设备或其他常见网络节点，并且通常包括许多或所有以上相对计算机 610 所描述的元件，但在图 6 中仅示出了存储器存储设备 681。图 6 中所示的逻辑连接包括一个或多个局域网 (LAN)671 和一个或多个广域网 (WAN)673，但也可以包括其他网络。此类联网环境在办公室、企业范围的计算机网络、内联网和因特网中是常见的。

[0109] 当在 LAN 联网环境中使用时，计算机 610 通过网络接口或适配器 670 连接到 LAN 671。当在 WAN 联网环境中使用时，计算机 610 通常包括调制解调器 672 或用于通过诸如因特网等 WAN 673 建立通信的其它手段。可为内置或可为外置的调制解调器 672 可以经由用户输入接口 660 或其他合适的机构连接至系统总线 621。诸如包括接口和天线的无线联网组件 674 等无线网络可以通过诸如接入点或对等计算机等合适的设备耦合到 WAN 或 LAN。在联网环境中，相关于计算机 610 所示的程序模块或其部分可被存储在远程存储器存储设备中。作为示例而非限制，图 6 示出了远程应用程序 685 驻留在存储器设备 681 上。可以理解，所示的网络连接是示例，也可以使用在计算机之间建立通信链路的其他手段。

[0110] 辅助子系统 699（例如，用于内容的辅助显示）可经由用户接口 660 连接，从而即使计算机系统的主要部分处于低功率状态中，也允许诸如程序内容、系统状态和事件通知等数据被提供给用户。辅助子系统 699 可连接至调制解调器 672 和 / 或网络接口 670，从而在主处理单元 620 处于低功率状态中时，也允许在这些系统之间进行通信。

[0111] 结语

[0112] 尽管本发明易于作出各种修改和替换构造，但其某些说明性实施例在附图中示出并在上面被详细地描述。然而应当了解，这不旨在将本发明限于所公开的具体形式，而是相反地，旨在覆盖落入本发明的精神和范围之内的所有修改、替换构造和等效方案。

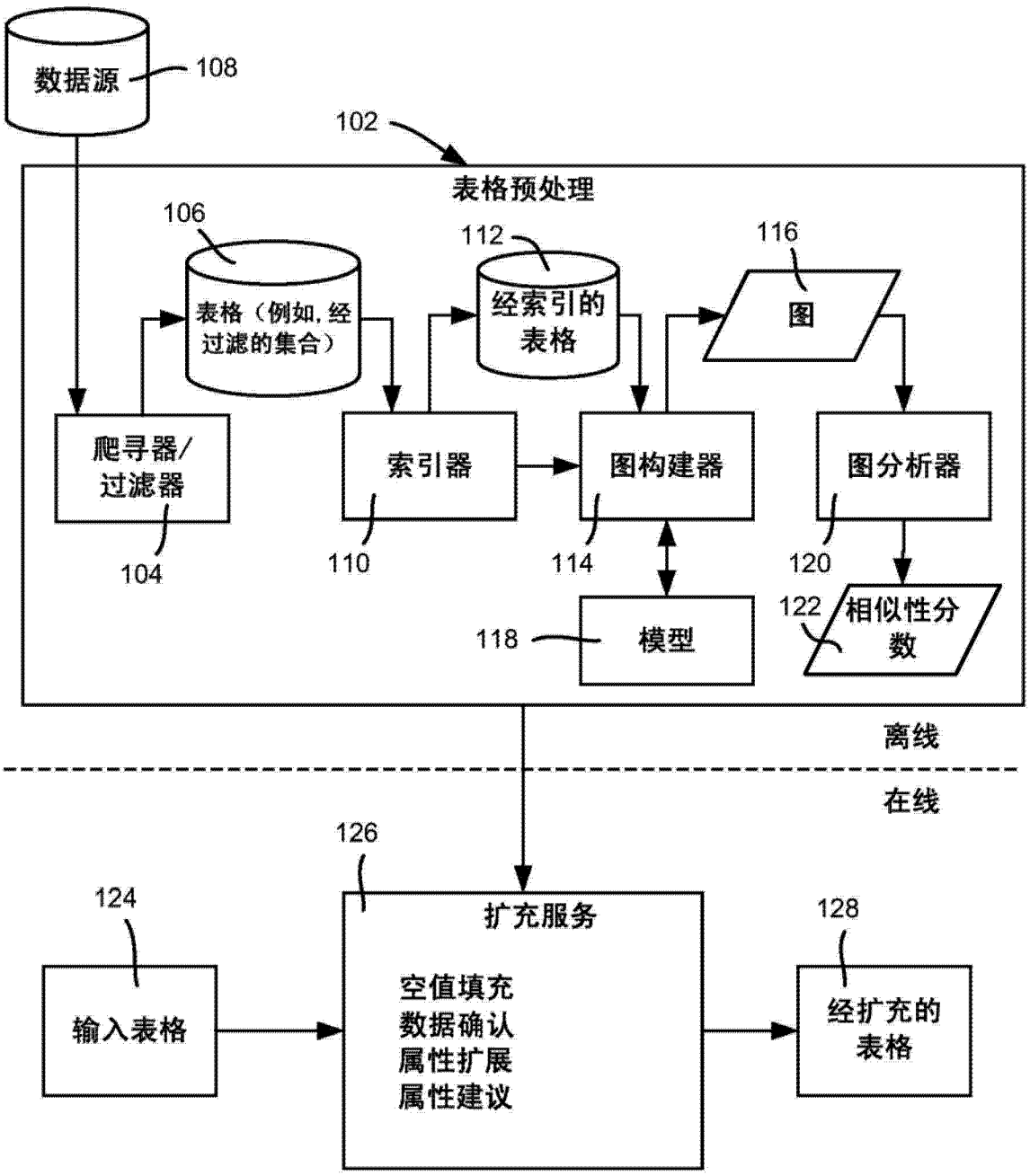

数据源
108
102
表格预处理
106
表格（例如，经过滤的集合）
112
经索引的表格
116
图
爬寻器/过滤器
104
索引器
110
图构建器
114
图分析器
120
118
模型
122
相似性分数
离线
在线
126
扩充服务
空值填充
数据确认
属性扩展
属性建议
124
输入表格
128
经扩充的表格

图 1

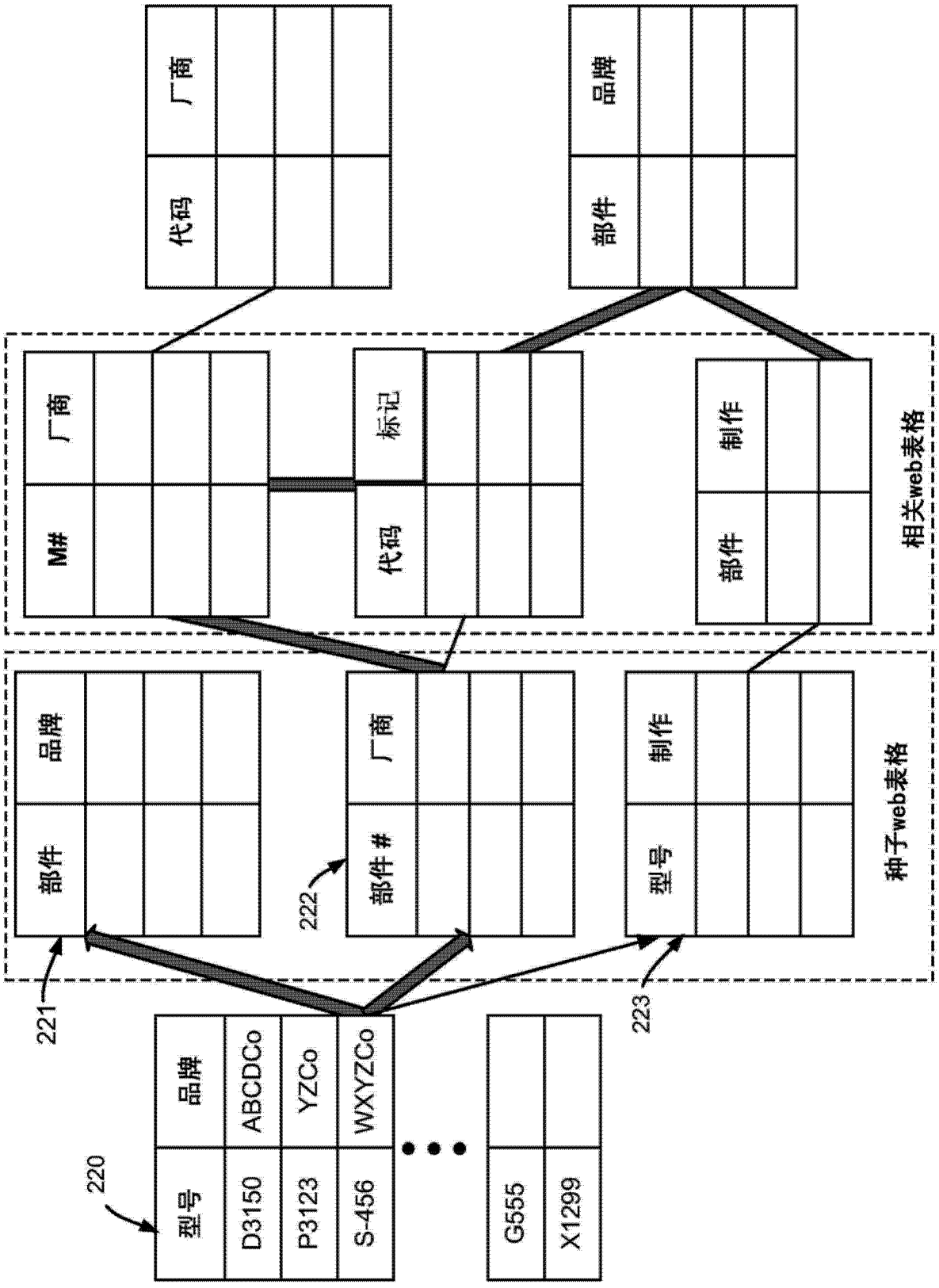

厂商
代码
品牌
部件
M#
厂商
标记
代码
制作
部件
相关web表格
品牌
部件
厂商
部件 #
制作
型号
种子web表格
221
222
223
220
品牌
型号
ABCDCo
D3150
YZCo
P3123
WXYZCo
S-456
G555
X1299

图 2

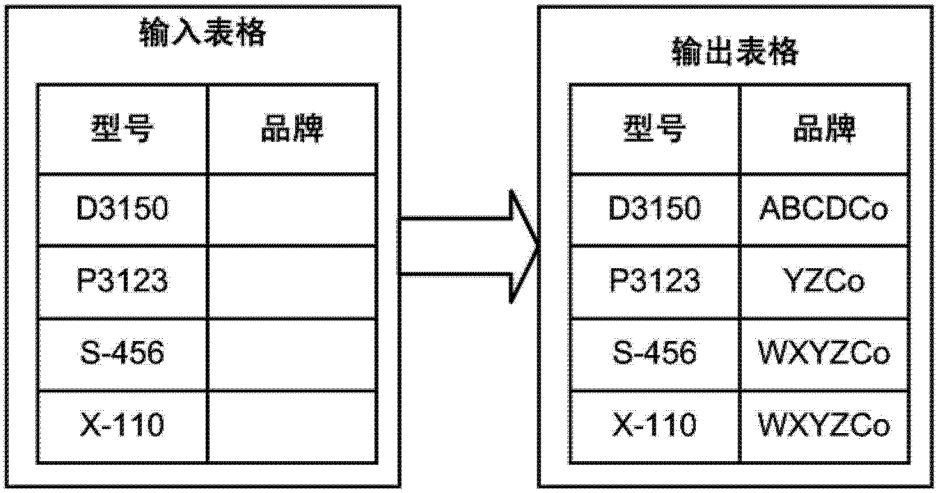

输入表格
型号
品牌
D3150
P3123
S-456
X-110
输出表格
型号
品牌
D3150
ABCDCo
P3123
YZCo
S-456
WXYZCo
X-110
WXYZCo

图 3A

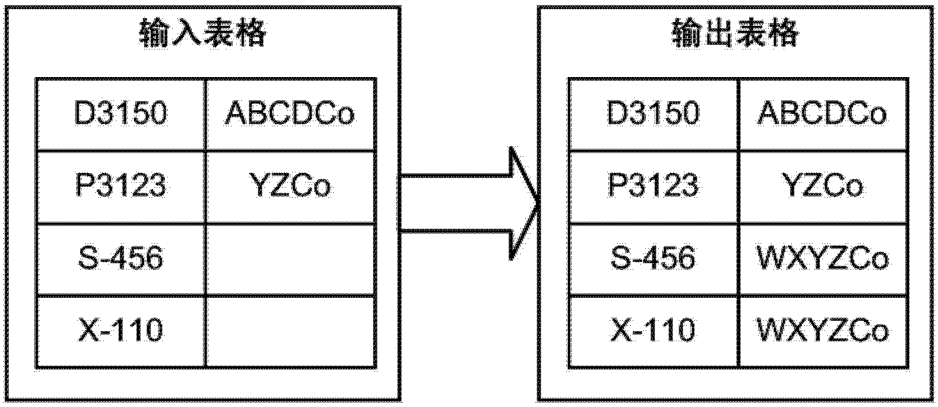

输入表格
D3150
ABCDCo
P3123
YZCo
S-456
X-110
输出表格
D3150
ABCDCo
P3123
YZCo
S-456
WXYZCo
X-110
WXYZCo

图 3B

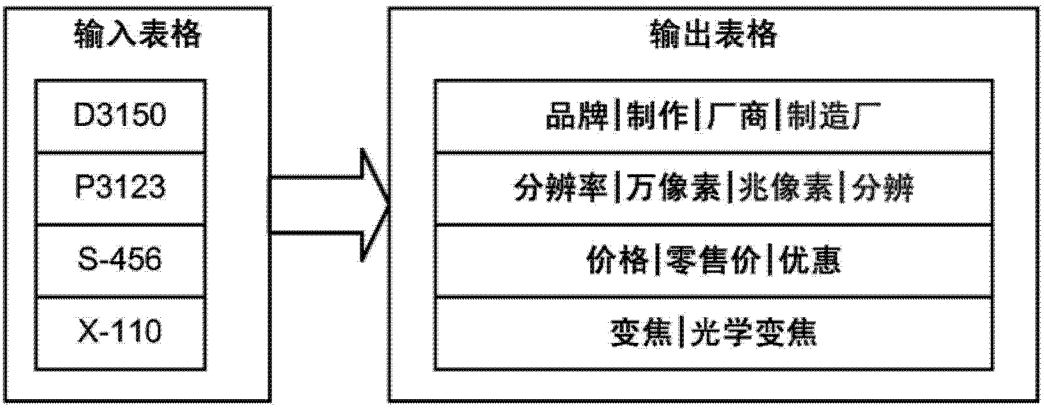

输入表格
D3150
P3123
S-456
X-110
输出表格
品牌|制作|厂商|制造厂
分辨率|万像素|兆像素|分辨
价格|零售价|优惠
变焦|光学变焦

图 3C

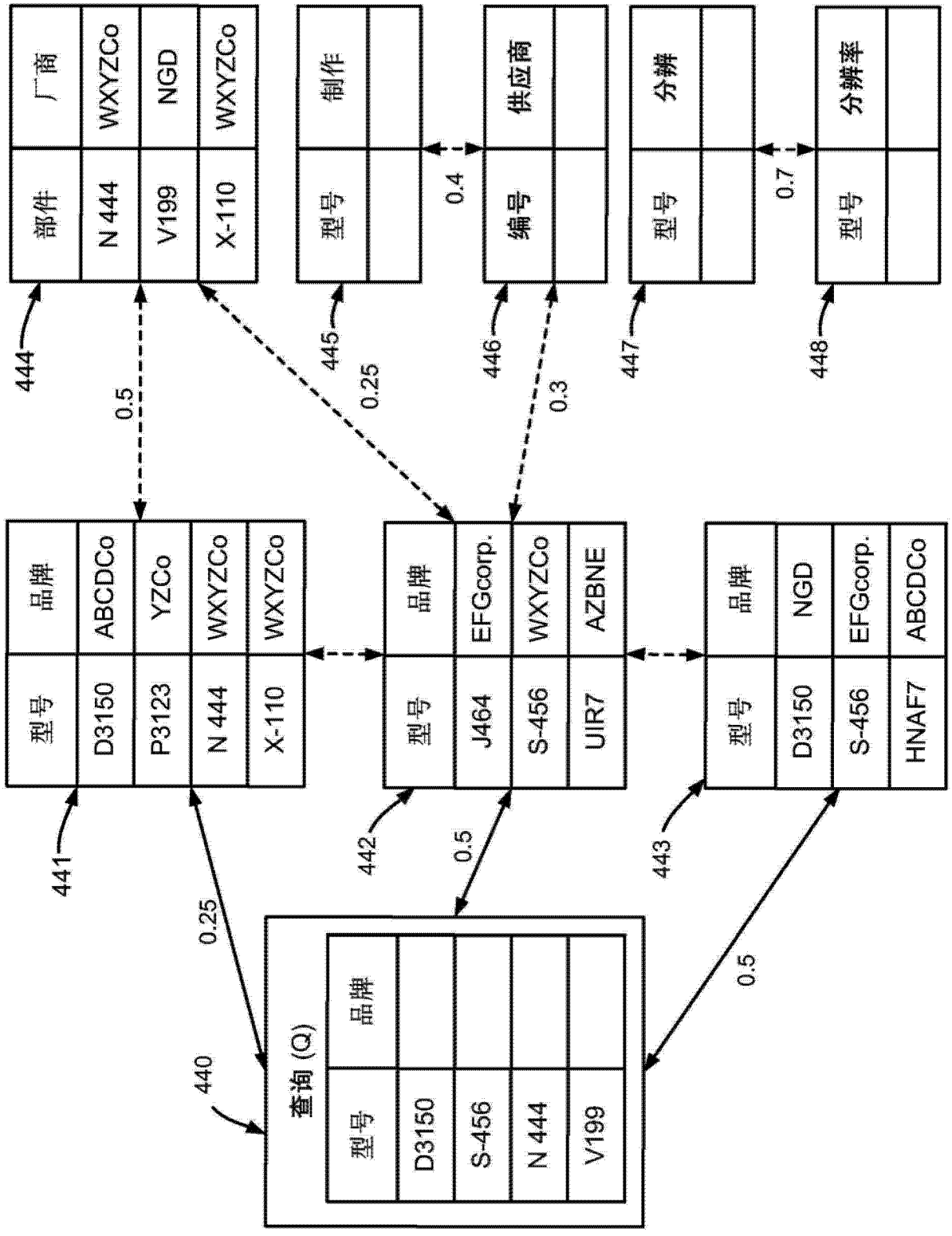

查询 (Q)
型号
品牌
D3150
S-456
N 444
V199
440
441
442
443
444
445
446
447
448
0.25
0.5
0.4
0.7
0.3
部件
厂商
N 444
V199
X-110
WXYZCo
NGD
制作
型号
供应商
编号
分辨
分辨率
品牌
型号
ABCDCo
YZCo
P3123
EFGcorp.
J464
S-456
UIR7
AZBNE
D3150
HNAF7

图 4

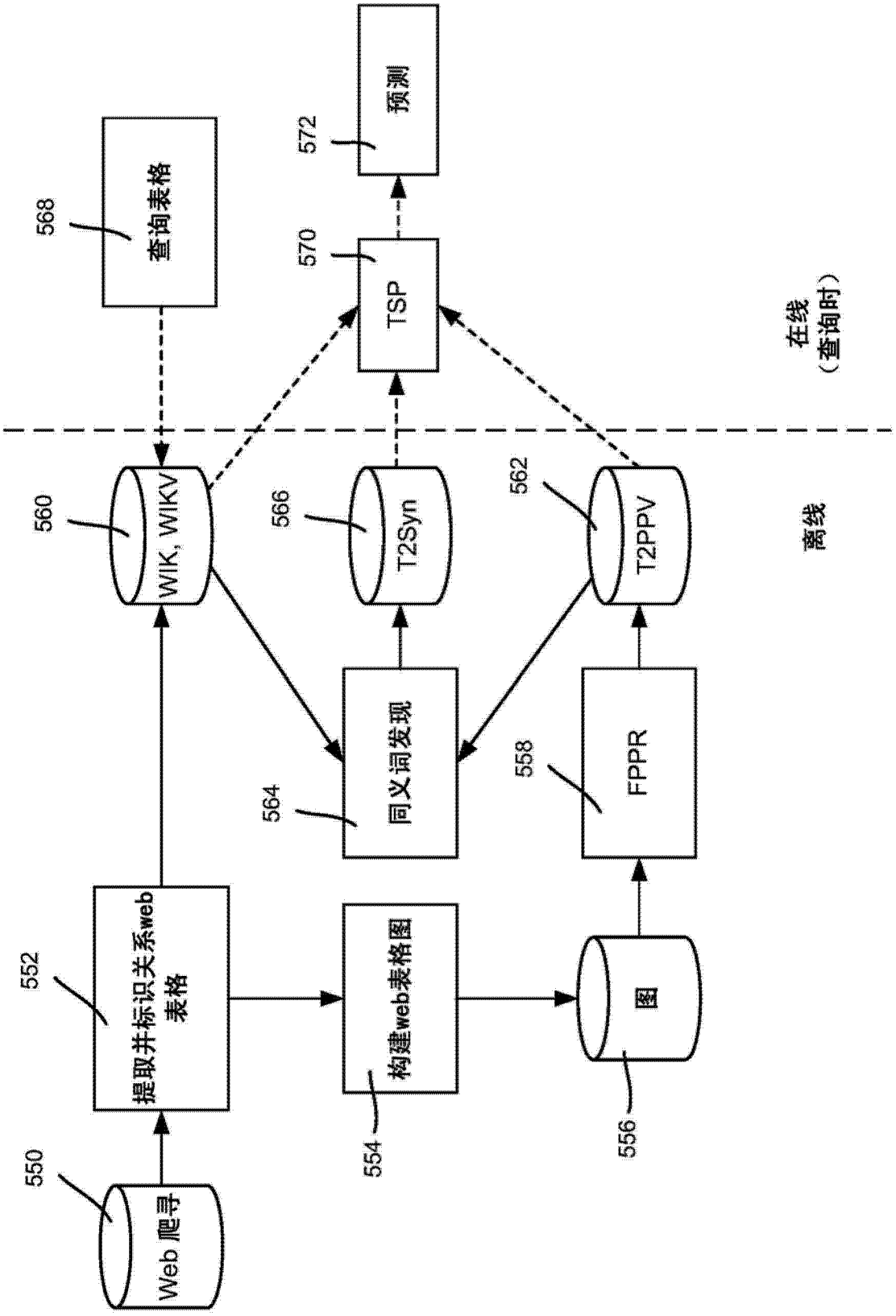

Web 爬寻
550
提取并标识关系web表格
552
构建web表格图
554
图
556
FPPR
558
T2PPV
562
同义词发现
564
T2Syn
566
WIK, WIKV
560
查询表格
568
TSP
570
预测
572
离线
在线（查询时）

图 5

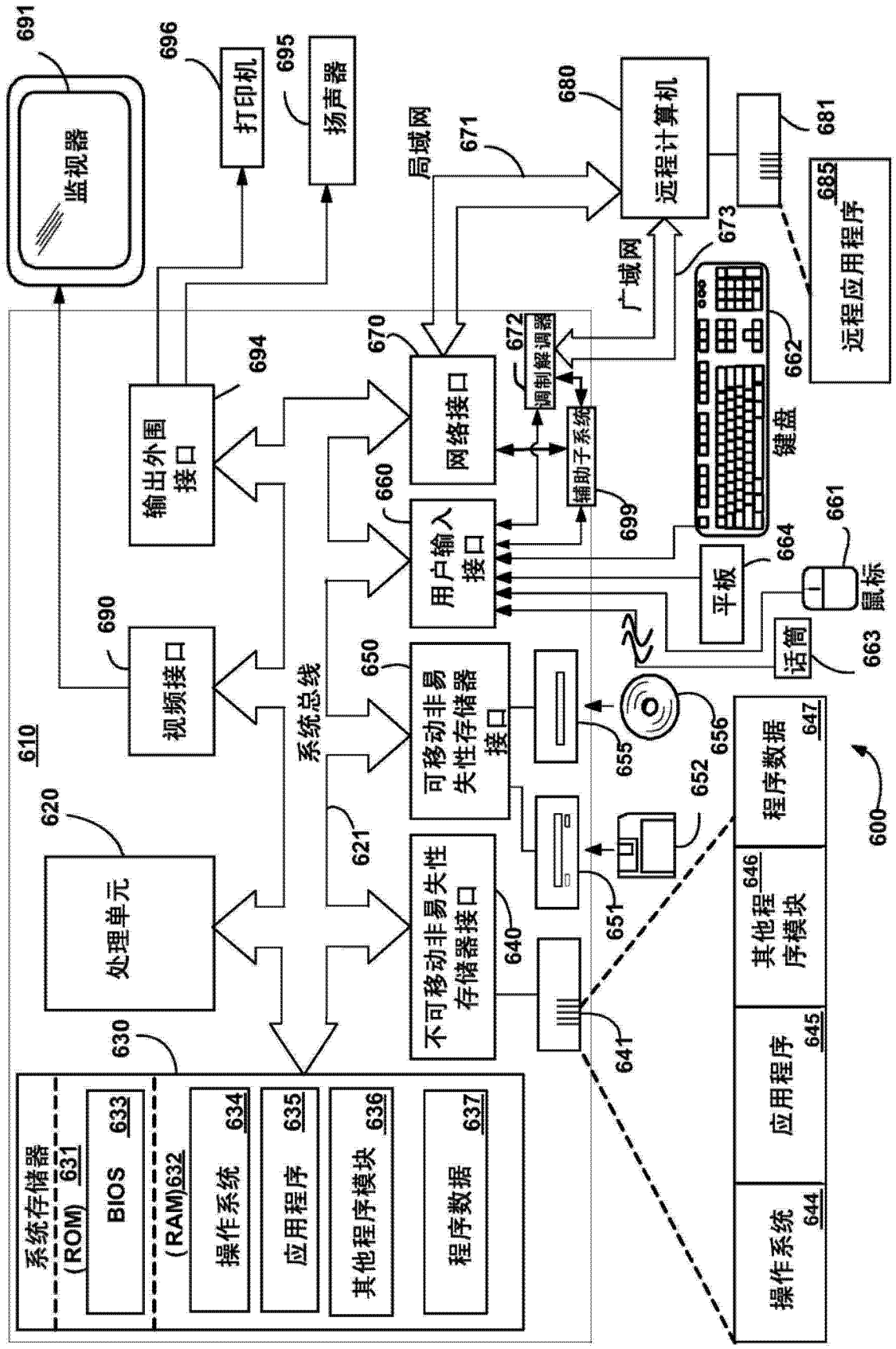

691
监视器
696
打印机
695
扬声器
局域网
671
680
远程计算机
681
685
远程应用程序
673
广域网
672
调制解调器
670
网络接口
694
输出外围接口
662
键盘
699
辅助子系统
660
用户输入接口
661
鼠标
664
平板
690
视频接口
663
话筒
650
可移动非易失性存储器接口
656
655
652
610
系统总线
620
处理单元
621
640
651
647
程序数据
646
其他程序模块
645
应用程序
644
操作系统
630
641
600
系统存储器
(ROM) 631
BIOS 633
(RAM)632
操作系统 634
应用程序 635
其他程序模块 636
程序数据 637
不可移动非易失性存储器接口

图 6