(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 102693246 B
(45)授权公告日 2015.03.11

(21)申请号 201110077432.9
(22)申请日 2011.03.22
(73)专利权人 日电(中国)有限公司
地址 100191 北京市海淀区学院路35号世宁大厦20层
(72)发明人 赵彧 李建强 刘博
(74)专利代理机构 北京市金杜律师事务所 11256
代理人 吴立明 庞淑敏
(51)Int.Cl.
G06F 17/30(2006.01)
G06N 5/04(2006.01)
(56)对比文件
CN 1659546 A,2005.08.24, 说明书第16页及附图6.
CN 1987866 A,2007.06.27, 说明书第5-17页及附图1-3.
CN 101571870 A,2009.11.04, 全文.
审查员 李楠

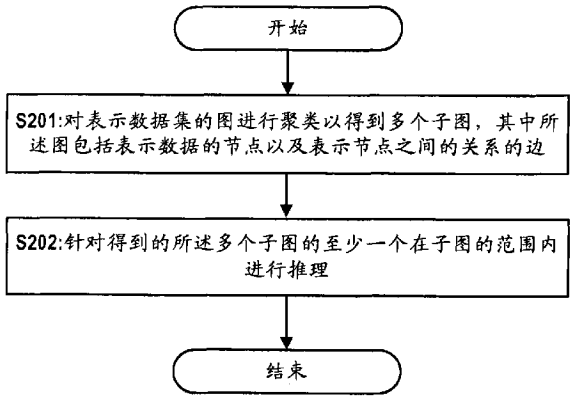
权利要求书2页 说明书7页 附图5页

(54)发明名称

一种用于从数据集获取信息的方法和系统

(57)摘要

本发明公开了一种用于从数据集获取信息的方法和系统。该方法可以包括对表示所述数据集的图进行聚类以得到多个子图,其中所述图包括表示数据的节点,以及表示所述节点之间的关系的边;以及在得到的所述多个子图的至少一个子图内进行推理。根据该方法,对图的聚类是以无监督的方式进行的,其不需要任何预先定义的模型,因此非常灵活且具有很强的适应性。另外,基于聚类而得到的每个子图中的节点数目及其关系均得到限制,因此根据本发明的实施方式,推理规模易于控制从而使得推理效率得以提高。



CN 102693246 B

1. 一种用于从数据集获取信息的方法，包括：

对表示所述数据集的图进行聚类以得到多个子图，其中所述图包括表示数据的节点以及表示所述节点之间的关系的边；以及

根据各个子图的重要性，对得到的所述多个子图进行排序；

针对得到的所述多个子图的至少一个在子图的范围内进行推理；其中，其中所述推理按照所述多个子图的排序进行；

其中所述重要性由子图的独立性或子图的紧密度和独立性来度量；

其中，所述子图的独立性通过连通度来度量；且子图 S 的连通度 Φ(S) 可以通过下式来计算得到：

$$\Phi(S)=\frac{c(S,G\setminus S)}{\min\{\deg(S),\deg(G/S)\}}$$

其中 G 表示整个图；S 表示子图；G\S 是图 G 排除 S 后的剩余部分；c(S，G\S) 是子图 S 与剩余部分 G\S 之间的分割尺寸，即在 S 与 G\S 之间的关系数目；deg(S) 是子图 S 的度，即子图内部的关系数目；

其中，紧密度和独立性为相对紧密度，且子图 S 的相对紧密度 ρ(S) 通过下式来表示：

$$\rho(S)=\frac{\deg(S)}{\deg(S)+\deg(G/S)}$$

其中 S 表示子图；G\S 是图 G 排除 S 后的剩余部分；deg(S) 和 deg(G/S) 分别是子图 S 和剩余部分 G/S 的度。

2. 根据权利要求 1 所述的方法，进一步包括：

响应于在一子图中进行推理未得到结果，通过扩展到其他子图中与该子图相连的节点来进行推理。

3. 根据权利要求 2 所述的方法，其中，按照以下各项其中之一来选择其他子图中与该子图相连的节点：

子图的排序；

该子图中缺少的节点关系类型；以及

节点之间的关系的优先级。

4. 根据权利要求 2 所述的方法，进一步包括：

响应于通过扩展到其他子图中与该子图相连的节点来进行推理得到结果，合并所述子图和所述其他子图以形成新子图；以及

在所述新子图内进行推理。

5. 根据权利要求 4 所述的方法，进一步包括：

保存所述新子图以供随后使用。

6. 一种用于从数据集获取信息的系统，包括：

聚类装置，配置用于对表示所述数据集的图进行聚类以得到多个子图，其中所述图包括表示数据的节点以及表示所述节点之间的关系的边；

排序装置，配置用于根据各个子图的重要性，对得到的所述多个子图进行排序；以及

推理装置，配置用于针对得到的所述多个子图的至少一个在子图的范围内进行推理；

所述推理装置配置用于按照所述多个子图的排序进行推理；

其中所述重要性由子图的独立性或子图的紧密度和独立性来度量；

其中，所述子图的独立性通过连通度来度量；且子图S的连通度Φ(S)可以通过下式来计算得到：

$$\Phi(S)=\frac{c(S,G\setminus S)}{\min\{\deg(S),\deg(G/S)\}}$$

其中G表示整个图；S表示子图；G\S是图G排除S后的剩余部分；c(S，G\S)是子图S与剩余部分G\S之间的分割尺寸，即在S与G\S之间的关系数目；deg(S)是子图S的度，即子图内部的关系数目；

其中，紧密度和独立性为相对紧密度，且子图S的相对紧密度 ρ(S)通过下式来表示：

$$\rho(S)=\frac{\deg(S)}{\deg(S)+\deg(G/S)}$$

其中S表示子图；G\S是图G排除S后的剩余部分；deg(S)和deg(G/S)分别是子图S和剩余部分G/S的度。

7. 根据权利要求6所述的系统，其中所述推理装置进一步配置用于：

响应于在一子图中进行推理未得到结果，通过扩展到其他子图中与该子图相连的节点来进行推理。

8. 根据权利要求7所述的系统，其中所述推理装置配置用于按照以下各项其中之一来选择其他子图中与该子图相连的节点：

子图的排序；

该子图中缺少的节点关系类型；以及

节点之间的关系的优先级。

9. 根据权利要求7所述的系统，进一步包括：

合并装置，配置用于响应于通过扩展到其他子图中与该子图相连的节点来进行推理得到结果，合并所述子图和所述其他子图以形成新子图；以及

其中所述推理装置配置用于在所述新子图内进行推理。

10. 根据权利要求9所述的系统，进一步包括：

保存装置，配置用于保存所述新子图以供随后使用。

一种用于从数据集获取信息的方法和系统

技术领域

[0001] 本发明涉及智能数据分析技术领域，更具体地涉及用于从数据集获取信息的方法和系统。

背景技术

[0002] 随着信息技术和网络技术的发展，各种数据和信息资源越来越为丰富，为了有效地对这些信息进行管理，W3C 提出了用于描述这些资源的方法，即资源描述框架 (RDF)。

[0003] 根据 RDF 的定义，数据或者资源可以使用一个三元组来表示，该三元组包括主体、谓语和客体，其中主体和客体指示数据或者资源，谓语指示主体和客体之间的关系。例如，对于三元组 instanceOf(X,author)，instanceOf(Y,paper)，hasPaper(X,Y) 以及 Topic(Y,D)，其分别表示 X 是作者，Y 是论文，作者 X 是论文 Y 的作者，以及 D 是论文 Y 的主题。

[0004] 基于这样的三元组，可以利用逻辑语言来描述规则（或者公理）以便利用该规则来执行自动推理过程。一个规则的实例为“hasPaper(X,Y)AND Topic(Y,D)- ＞ author(X,D)”，其表示如果作者 X 是论文 Y 的作者且论文 Y 的主题为 D，则作者 X 是主题 D 的作者或者主题 D 方面的专家。这样的描述机制使得自动化推理成为可能，从而可以基于三元组和规则来实现智能语义信息检索和挖掘。

[0005] 在对大规模数据或者网络规模数据进行推理时，效率一直是令人困扰并亟待解决的难题。而且这一问题也成为在实际的信息处理中广泛应用语义推理的障碍。

[0006] 针对该问题，在本领域中已经提出了一些解决方案。例如在美国专利US7689526B2 中提出了一种解决方案，根据该解决方案，首先基于已有的知识规则对数据进行分类，然后针对分类后的数据利用压缩模型来表示知识规则。该方案旨在通过使用由压缩模型表示的规则来提高推理效率。

[0007] 另外，在由 E.Amir 和 S.McIlraith 发表于 Representation and Reasoning(2000) 题为“Partition-based Logical Reasoning”一文中，公开了另一种解决方案，该解决方案首先对规则集进行分析，然后对规则进行划分，以由此来改善推理效率。

[0008] 前述两种解决方案都是通过对规则进行预处理来改善推理效率，因此属于基于规则的技术。然而，由于规则通常是仅仅适用于特定情形（例如，依赖于查询），因此这两种解决方案具有适应性不好、灵活性较差的缺陷。此外，对于数据量巨大的情况，即便是利用一条规则来进行推理也需花费大量时间，因此在这种情况下，这两种解决方案对效率改善具有有限的作用。

[0009] 此外，在由 Y.Zeng，Y.Wang，Z.Huang 和 N.Zhong 发表于 Lecture Notes in Computer Science(2009，Vol.5820，第 418-429 页）题为“Unifying Web-Scale Search and Reasoning from the Viewpoint of Granularity”一文中，公开了一种基于模型的节点分组技术的解决方案。出于说明的目的，在图 1A 至图 1C 中示意性地示出了根据该技术方案的原理的图示。

[0010] 如图 1A 所示，根据该解决方案，数据集通过包括节点和边的图来表示，其中节点

表示数据或者资源，例如 RDF 的主体和客体，边表示数据或者资源之间的关系，例如 RDF 的谓语。为了清晰起见，在图 1A 中分别采用圆形、方形和三角形图案的节点来表示前面给出的三元组的示例中的作者、论文和主题，圆形节点与方形节点之间的边（链接）表示“hasPaper”这一关系，以及方形节点与三角形节点之间边（链接）表示“hasTopic”这一关系。

[0011] 接着，如图 1B 所示，可以基于预先建立的节点分组模型，对该图中的节点执行分组操作，从而得到一个排序的节点组列表。节点组列表的排序是基于例如作者论文的数目来进行，论文数目较多的作者排序较为靠前。然后，如图 1C 所示，按照各个节点组的排序，分别在第一、第二和第三推理事务中，对各个节点组逐个地执行推理，推理的范围为对应节点组及与该节点组相连通的所有其他节点，以及这些节点之间的边。

[0012] 由于这一技术方案是通过预先建立的节点分组模型来针对数据进行预处理，所以这种方案对于预先建立的节点分组模型具有很大的依赖性，这使得该解决方案的灵活性较差，不能适用于动态的查询需求。另外，该解决方案是通过节点分组限制推理规模，但其仅仅限制了触发推理的节点的数量，由于节点之间还存在大量错综复杂的关系，所以推理规模实际上难以得到有效的控制。此外，根据该技术方案，在每个推理事务中还涉及大量重复的节点，这也进一步恶化了推理模块的控制有效性。

[0013] 为此，本领域存在一种对于在数据分析过程中采用的推理技术进行改进的迫切需要。

发明内容

[0014] 有鉴于此，本发明提供了一种用于从数据集获取信息的方法和系统，以克服或者至少部分消除现有技术中存在的缺陷。

[0015] 根据本发明的一个方面，提供了一种用于从数据集获取信息的方法。该方法可以包括对表示所述数据集的图进行聚类以得到多个子图，其中所述图包括表示数据的节点以及表示所述节点之间的关系的边；以及针对得到的所述多个子图的至少一个在子图的范围内进行推理。

[0016] 在根据本发明的一个优选实施方式中，该方法可以进一步包括：根据各个子图的重要性，对得到的所述多个子图进行排序。在该实施方式中，推理可以按照所述多个子图的排序依次地进行。

[0017] 在根据本发明的一个实施方式中，各个子图的重要性可以由以下其中一项或者多项来度量：子图的紧密度；子图的独立性；以及子图的层级。

[0018] 在根据本发明的另一实施方式中，该方法可以进一步包括：响应于在一子图中进行推理未得到结果，通过扩展到其他子图中与该子图相连的节点来进行推理。

[0019] 在根据本发明的再一实施方式中，按照以下各项其中之一来选择其他子图中与该子图相连的节点：子图的排序；该子图中缺少的节点关系类型；以及节点之间的关系的优先级。

[0020] 在根据本发明的又一实施方式中，该方法可以进一步包括：响应于通过扩展到其他子图中与该子图相连的节点来进行推理得到结果，合并所述子图和所述其他子图以形成新子图；以及在所述新子图内进行推理。

[0021] 在根据本发明的另一优选实施方式中，该方法可以进一步包括保存新子图以供随后使用。

[0022] 此外，根据本发明的另一方面，还提供了一种用于从数据集获取信息的系统。该系统可以包括：聚类装置，配置用于对表示所述数据集的图进行聚类以得到多个子图，其中所述图包括表示数据的节点以及表示所述节点之间的关系的边；以及推理装置，配置用于针对得到的所述多个子图的至少一个在子图的范围内进行推理。

[0023] 根据本发明的实施方式，对图的聚类是以无监督的方式进行的，其不需要任何预先定义的模型，因此非常灵活且具有很强的适应性。另外，推理在子图范围内进行，基于聚类而得到的每个子图中的节点数目及其关系均得到限制，且基于聚类而得到的每个子图中没有重复的节点和关系。因此根据本发明的实施方式，推理规模易于控制，从而使得推理效率得以提高。

附图说明

[0024] 通过对结合附图所示出的实施方式进行详细说明，本发明的上述以及其他特征将更加明显，本发明附图中相同的标号表示相同或相似的部件。在附图中：

[0025] 图 1A 至图 1C 示出了根据现有技术的一种从数据集获取信息的技术方案。

[0026] 图 2 示出了根据本发明的一个实施方式用于从数据集获取信息的方法的流程图。

[0027] 图 3 是示出了本发明的原理的示意图。

[0028] 图 4 示出了根据本发明的另一实施方式用于从数据集获取信息的方法的流程图。

[0029] 图 5 示出了根据本发明的优选实施方式用于对子图进行调整的原理示意图。

[0030] 图 6 示出了根据本发明的一个实施方式用于从数据集获取信息的系统的方框图。

具体实施方式

[0031] 在下文中，将参考附图通过实施方式对本发明提供的用于从数据集获取信息的方法和系统进行详细的描述。

[0032] 首先将参考图 2 至图 5 来描述本发明所提供的方法。参考图 2，该图 2 示出根据本发明的一个实施方式用于从数据集获取信息的方法的流程图。

[0033] 如图 2 所示，首先在步骤 201，对表示数据集的图进行聚类以得到多个子图。该图包括表示数据的节点以及表示所述节点之间的关系的边。

[0034] 聚类是图论中的一项重要技术，其目标是将图中的节点和关系划分成类簇。图聚类的总体思路是基于图中的边（关系）结构来进行聚类，以使得每个类簇内部的关系比两个类簇之间的关系更加密切。为此，本发明人将图形聚类技术应用于智能数据分析的领域，利用图聚类技术将表示数据集的图分割成若干子图。

[0035] 图 3 是示出了本发明的原理的示意图。如图 3 所示，在本发明中，采用图来表示数据集，其中图的节点表示数据，而节点之间的链路或者边表示节点之间的关系。针对该表示数据集的图，基于图形聚类技术来进行聚类，从而将该图聚类成如图 3 中所示的以圆形虚线示出的若干子图（类簇）。

[0036] 表示数据集的图可以存储在存储单元中。例如，各个节点可以存储在节点存储单元中，并例如以［节点 ID，节点名］的形式存储；节点之间的关系可以存储在节点关系存储

单元中，且例如以［关系 ID，关系名，主体节点 ID，客体节点 ID］的形式存储；聚类得到的子图或者类簇可以存储在例如子图存储单元中，例如以［子图 ID，节点列表，关系列表］的形式来存储。

[0037] 对表示数据集的图进行聚类可以采用已知的或者将来开发出的任何图聚类方法来实现。例如，可以采用基于连通性的图聚类算法，依据该算法可以将每对节点之间存在的路径的数目作为进行聚类时的一种度量，对于属于相同类簇的节点，它们之间应当具有高度的连通性。

[0038] 在根据本发明的一种具体实现中，可以采用高连通子图 (HCS) 算法，其中设置了边连通性阈值 k。然后，可以针对图 G 执行对该图的最小割算法 (minimum-cut) 以将该图分割成两个子图 H，H'。如果子图 G 的边连通性高于连通性阈值 k，则返回图 G 作为分类后的类簇，否则将子图 H 和 H'作为新的输入以便进行下一次迭代处理。该过程一直重复直至得到的子图的连通性均高于阈值 k。这样，就可以得到若干个具有高度内部关联性的子图。得到的子图，如前所述可以存储在子图存储单元中。

[0039] 此外，也可以在全局层次上利用分层聚类将子图形成为分层结构。例如，在采用 HCS 算法的情况下，可以设置多个边连通性阈值，其中可以将较低的阈值应用于更高层次的聚类，而将较高的阈值应用较低层次的聚类。通过这样的聚类，就可以获得具有分层结构的多个子图。

[0040] 然后，可以在步骤 S202 中，针对得到的所述多个子图的至少一个在子图的范围内进行推理。推理使用的推理规则，例如可以存储在规则（公理）存储单元中，且例如以［规则 ID，规则语句］的形式存储。在各个子图中执行推理可以采用现有技术中的方法来进行，此处出于简化的目的，不再对推理的具体细节进行赘述。

[0041] 根据本发明的实施方式，对表示数据集的图进行分组是基于图聚类技术，其是以自动地、无监督的方式进行的，而无需依赖任何预定义的分组模型，因此，本发明具有很高的灵活性和很强的适应性。此外，本发明是基于图聚类技术实现的分组，每个组中的节点和关系都得到了限制，因此可以提高推理效率，同时可以很好地控制推理的规模。

[0042] 此外，图 4 中还给出了根据本发明的另一实施方式的方法的流程图。在图 4 中，步骤 S401 和 S402 基本对应于图 2 中的步骤 S201 和 S202，因此不再对此进行详细描述。与图 2 中不同的是，在图 4 所示的实施方式中，在步骤 S402 之前还进一步包括步骤 S403。在该步骤 S403 中，进一步根据各个子图的重要性，对得到的所述多个子图进行排序，以便确定对子图执行推理的顺序。

[0043] 根据本发明，在获得了多个子图之后，逐个地对子图进行推理。然而，对于诸如搜索等在线应用场景，通常设置有系统的响应时间，如果能在有限的响应时间内对最重要的子图进行推理，这将是有益的。

[0044] 为此，根据本发明的优选实施例，对聚类得到的子图进行排序以使得包括重要信息的子图排序更为靠前。这样对子图执行聚类时，可以按照子图的排序依次来进行，以便使得在响应时间结束后向用户返回最为有效的结果。

[0045] 例如，可以使用子图的内部特征作为重要性的度量。通常，紧密度更高、独立性更高的子图更可能得到更有效的结果，因此，这样的子图也更为重要。

[0046] 在根据本发明的一个实施方式中，选择子图与其他子图之间的独立性作为对子图

进行排序的依据。该独立性例如通过连通度 (conductance) 来度量。子图 S 的连通度 Φ (S) 可以通过下式来计算得到 :

[0047] $$\Phi(S)=\frac{c(S,G\setminus S)}{\min\{\deg(S),\deg(G\setminus S)\}}$$ 式 (1)

[0048] 其中 G 表示整个图 ;S 表示子图 ;G\S 是图 G 排除 S 后的剩余部分 ;c(S, G\S) 是子图 S 与剩余部分 G\S 之间的分割尺寸,即在 S 与 G\S 之间的边数目 ;deg(S) 是子图 S 的度,即子图内部的边数目。对子图的排序可以基于该连通度 Φ (S) 进行,连通度值较低 (即独立性高) 的子图可以排序较为靠前,连通度值较高 (即独立性低) 的子图可以排序较为靠后。

[0049] 另外,也可以将紧密度和独立性两者 (即相对紧密度) 作为排序的一种度量。子图 S 的相对紧密度 ρ (S),例如可以通过下式来表示 :

[0050] $$\rho(S)=\frac{\deg(S)}{\deg(S)+c(S,G\setminus S)}$$ 式 (2)

[0051] 其中类似地,S 表示子图 ;G\S 是图 G 排除 S 后的剩余部分 ;deg(S) 和 c(S,G\S) 分别是子图 S 的度以及子图 S 和剩余部分 (G\S) 之间的分割尺寸。在采用相对紧密度的情况下,可以将具有较大相对紧密度值的子图排在较为靠前的位置,而将将具有较小相对紧密度值的子图排在较为靠后的位置。

[0052] 此外,在采用分层聚类的情况下,还可以进一步基于各个子图的层级来进行排序。例如,可以将位于在层级中较低层的那些子图排在位于较高层中的那些子图之前。

[0053] 这样,就可以在步骤 S402,基于多个子图的排序,逐一地对子图执行推理,直至总的推理时间已经达到限制或者已经完成对所有子图的推理。这样就可以尽可能在推理时间结束时向用户返回最重要的推理结果。

[0054] 此外,还优选的是,可以在步骤 S402 之后,在步骤 S404 中响应于在一子图中进行推理未得到结果,通过扩展到其他子图中与该子图相连的节点来进行推理。

[0055] 如前所述,基于图聚类的推理有效地限制了推理规模,但发明人也注意到,这种方式同时也断开了一些节点之间的关系。而在一些特定情况下,聚类很可能断开了将用于推理的重要关系,从而导致针对特定的推理规则在一些子图内无法得出推理结果。

[0056] 考虑到这一情况,特别是对于重要性较高的子图,本发明优选地,通过考虑其他子图中与该子图相连的节点来进行推理,以便能够得到有效的推理结果。此外,如果通过考虑这些节点能够得到有效的推理结果,则可以将这些节点合并到该子图中,以便在随后推理时使用。备选地,也可以在步骤 S405 将该子图与这些节点所在的子图合并,从而形成新的子图并在步骤 S406 针对新子图执行推理,以便得到有效的推理结果。此外还优选的是,可以保存合并得到的新子图,以便例如随后在利用相应的推理规则进行推理时使用。

[0057] 根据本发明的一个实施方式,如果在一个子图 C1 的范围内进行推理得到的推理结果为空,即该推理没有得到结果,则调查其他子图,即考虑其他子图中与该子图相连的节点。鉴于与该子图相连的节点可能存在于多个子图中,因此可以设置选择这些节点的优先次序。例如,可以按照备选节点所在子图的排序,来选择其他子图中与该子图相连的节点。对于排序较为靠前的子图的节点,可以优先考虑。此外,也可以考虑推理中缺少的节点关系

类型，并优先考虑涉及到缺少的节点关系类型的子图。另外，也可以考虑节点之间的关系的优先级。这些优先级可以针对各个推理规则预先设定。可以优先考虑涉及到优先级较高的关系的子图。此外，也可以将上述选择依据结合使用。

[0058] 在根据本发明的另一实施方式中，可以通过评估外部链路的必要性来并入重要的子图，以便用于进一步推理。例如可以将对于一个子图而言重要的其他子图定义如下：假设子图 C1 和子图 C2 之间的边集为 E，与该边集 E 中的边相连且位于 C2 中的边节点集合为 V，如果在 C1+E+V 的范围内进行推理能够得有意义的结果，则对于该特定的推理规则而言，C2 是该 C1 的重要子图。当然，这只是用于确定对于 C1 重要的子图的一个示例，本发明并局限于此，而是可以采用任何适当的方法来确定。

[0059] 这样，在找到重要的子图 C2 的情况下，则可以将子图 C1 和 C2 合并，以得到以新子图，然后在新子图内执行推理。

[0060] 图 5 示出了根据本发明的优选实施方式用于对子图进行调整的原理示意图。如图所示，在排序第一的子图中，没有得到推理结果。因此，可以将推理范围扩展到相邻的子图（具有第二排序）中与该子图相连的节点，例如扩展至子图 2 的 α 和 d。如果在子图 1 扩展了外部边 (b，α)，(c，α)，(1，d) 和 (2，d) 及外部节点 α 和 d 的情况下能够得到推理结果，则将该子图 2 识别为重要子图。然后，例如可以将子图 1 与被识别为是重要子图的子图 2 合并，从而得到新子图，如在图 5 中以点划线圆圈所示。接着，可以在该合并后的新子图的范围内执行推理。

[0061] 通过这样的调整操作，就可以避免基于聚类的这种分组方法可能带来的重要关系被切断的情况，从而使得本发明的技术方案在考虑推理效率的同时，也能更充分地考虑到推理的有效性。

[0062] 此外，本发明还提供了一种用于从数据集获取信息的系统。在下文中将参考图 6 对其进行描述，该图 6 示意性地示出了根据本发明的一个实施方式的用于从数据集获取信息的系统。

[0063] 如图 6 所示，系统 600 可以包括聚类装置 601 和推理装置 602。该聚类装置 60 1 配置用于对表示数据集的图进行聚类以得到多个子图。该图包括表示数据的节点以及表示所述节点之间的关系的边。该推理装置 602，配置用于针对得到的所述多个子图的至少一个在子图的范围内进行推理。

[0064] 根据本发明的一个优选实施方式，该系统 600 可以进一步包括：排序装置 603，配置用于根据各个子图的重要性，对得到的多个子图进行排序。在该实施方式中，所述推理装置 602 配置用于按照所述多个子图的排序进行推理。

[0065] 根据本发明的另一优选实施方式，子图的重要性可以由以下其中一项或者多项来度量：子图的紧密度；子图的独立性；以及子图的层级。

[0066] 根据本发明的再一优选实施方式，推理装置 602 可以进一步配置用于响应于在一子图中进行推理未得到结果，通过扩展到其他子图中与该子图相连的节点来进行推理。

[0067] 根据本发明的又一优选实施方式，推理装置 602 可以配置用于按照以下各项其中之一来选择其他子图中与该子图相连的节点：子图的排序；该子图中缺少的节点关系类型；以及节点之间的关系的优先级。

[0068] 根据本发明的另一优选实施方式，该系统 600 可以进一步包括合并装置 604，配置

用于响应于通过扩展到其他子图中与该子图相连的节点来进行推理得到结果，合并所述子图和所述其他子图以形成新子图。在该实施方式中，推理装置 602 可以配置用于在所述新子图内进行推理。

[0069] 根据本发明的再一优选实施方式，该系统 600 可以进一步包括保存装置 605，配置用于保存所述新子图以供随后使用。

[0070] 需要指出的是，该系统 600 中所包括的各个装置的操作与前面描述的各个方法步骤基本上是对应的，因此，关于该系统 600 中的各个装置的具体操作，可以参考前文结合图 2 至图 5 对本发明的方法的描述。

[0071] 在上文中主要参考搜索和查询对本发明进行了描述。然而本发明并不仅限于此，而是可以应用于其他任何适当的情形，例如数据挖掘。

[0072] 此外，在上文中在描述对聚类得到的子图进行排序时主要以紧密度、独立性、层级等标准为示例。然而需要说明的是，还可以替代地或者附加地采用任何其他适当的标准来进行排序。

[0073] 另外，在将推理扩展至其他子组中的节点时，也可以以按照本文所给的依据之外的其他依据来选择其他子图中与该子图相连的节点。

[0074] 在上文中主要结合 RDF 进行了描述，然而需要说明的是，并不局限于此，而是也可以与其他任何适当的数据或者资源表述方式结合使用。

[0075] 此外，在上文中结合简单的特定三元组示例对本发明进行了描述，然而需要说明的是，这只是出于示例的目的。本发明实际可以应用于大规模或者网络规模的数据，而且实际推理过程可能也会复杂得多。

[0076] 此外，本发明的实施方式可以以软件、硬件或者软件和硬件的结合来实现。硬件部分可以利用专用逻辑来实现；软件部分可以存储在存储器中，由适当的指令执行系统，例如微处理器或者专用设计硬件来执行。本领域的普通技术人员可以理解上述的方法和系统可以使用计算机可执行指令和 / 或包含在处理器控制代码中来实现，例如在诸如磁盘、CD 或 DVD-ROM 的载体介质、诸如只读存储器（固件）的可编程的存储器或者诸如光学或电子信号载体的数据载体上提供了这样的代码。本实施例的系统及其组件可以由诸如超大规模集成电路或门阵列、诸如逻辑芯片、晶体管等的半导体、或者诸如现场可编程门阵列、可编程逻辑设备等的可编程硬件设备的硬件电路实现，也可以用由各种类型的处理器执行的软件实现，也可以由上述硬件电路和软件的结合例如固件来实现。

[0077] 虽然已经参考目前考虑到的实施方式描述了本发明，但是应该理解本发明不限于所公开的实施方式。相反，本发明旨在涵盖所附权利要求的精神和范围内所包括的各种修改和等同布置。以下权利要求的范围符合最广泛解释，以便包含所有这样的修改及等同结构和功能。

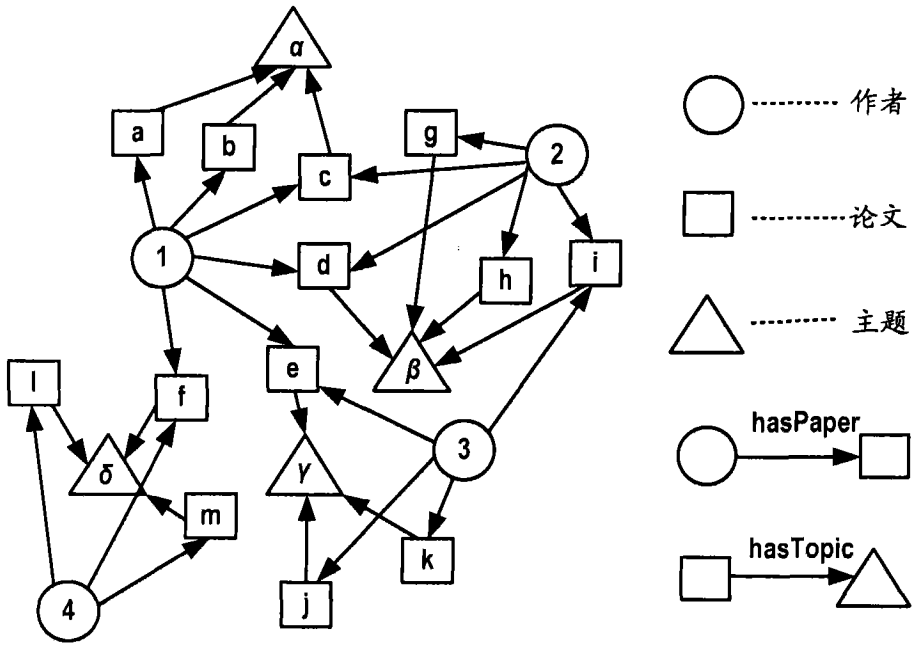


图 1A

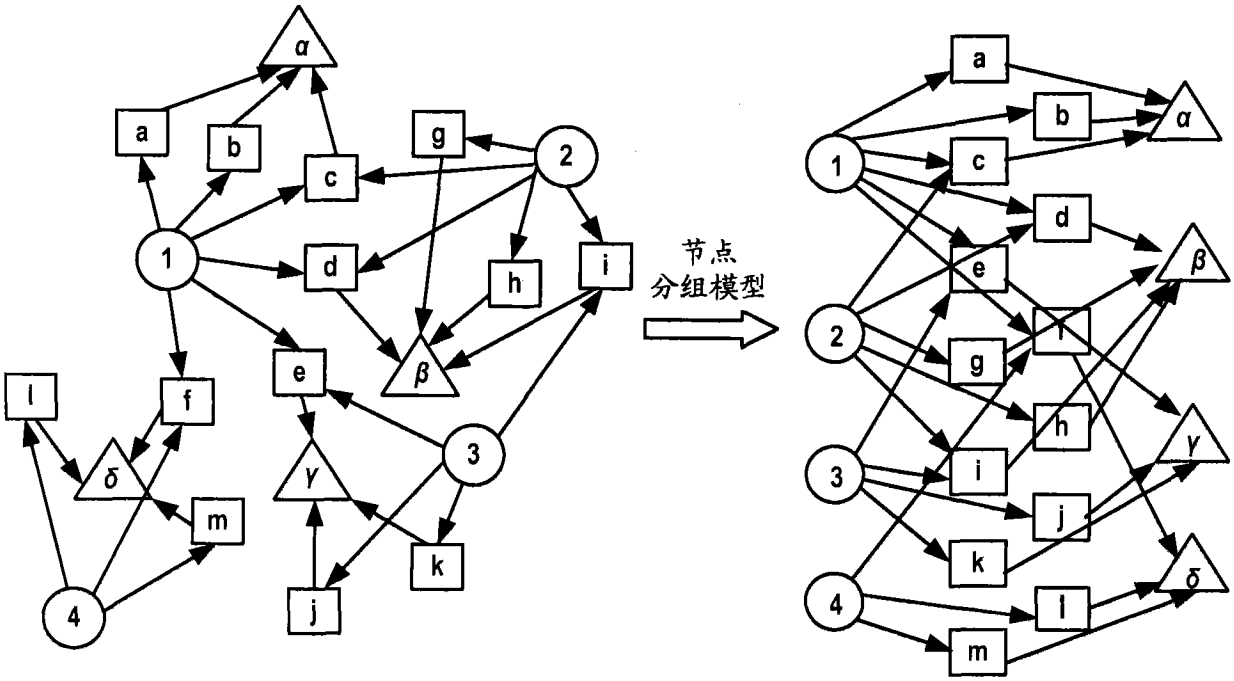


图 1B

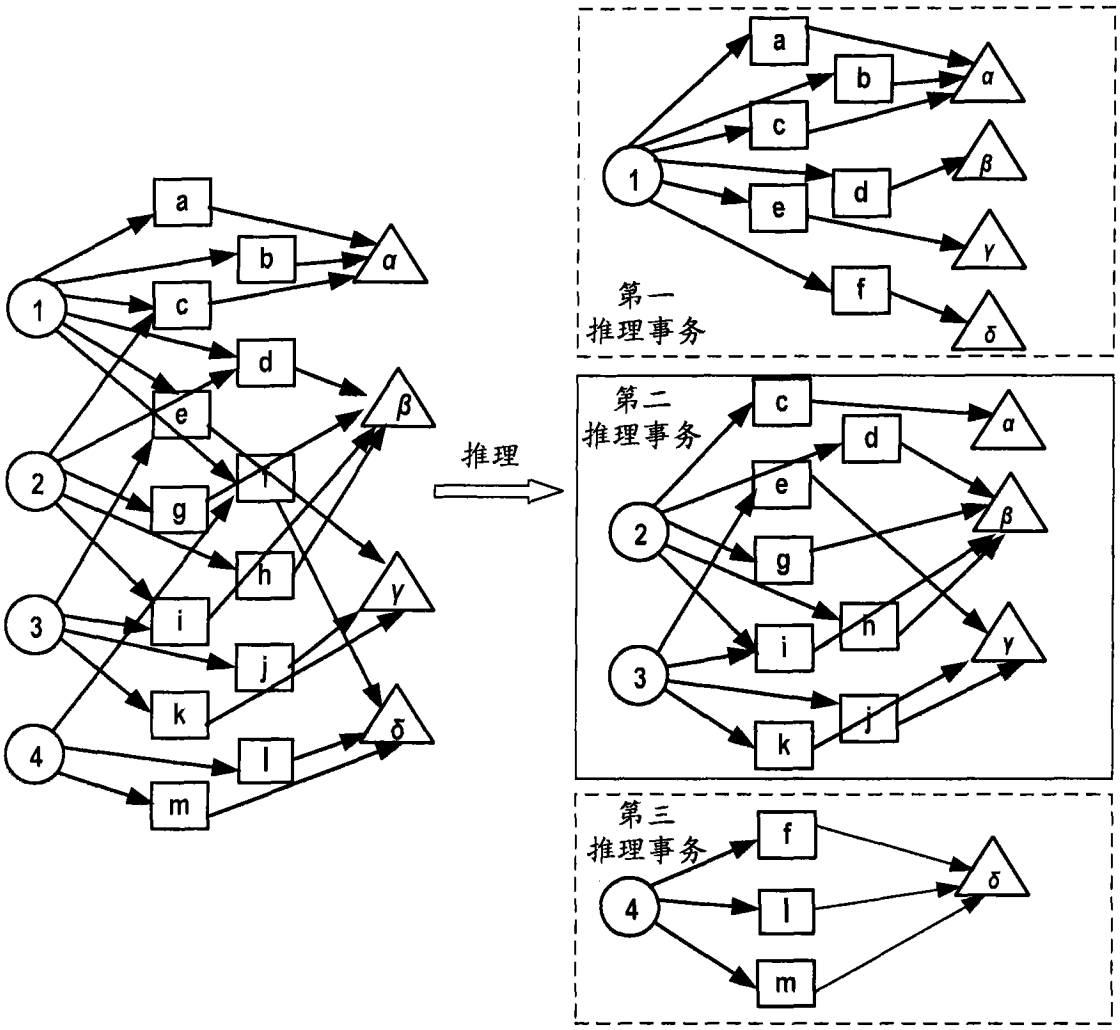


图 1C

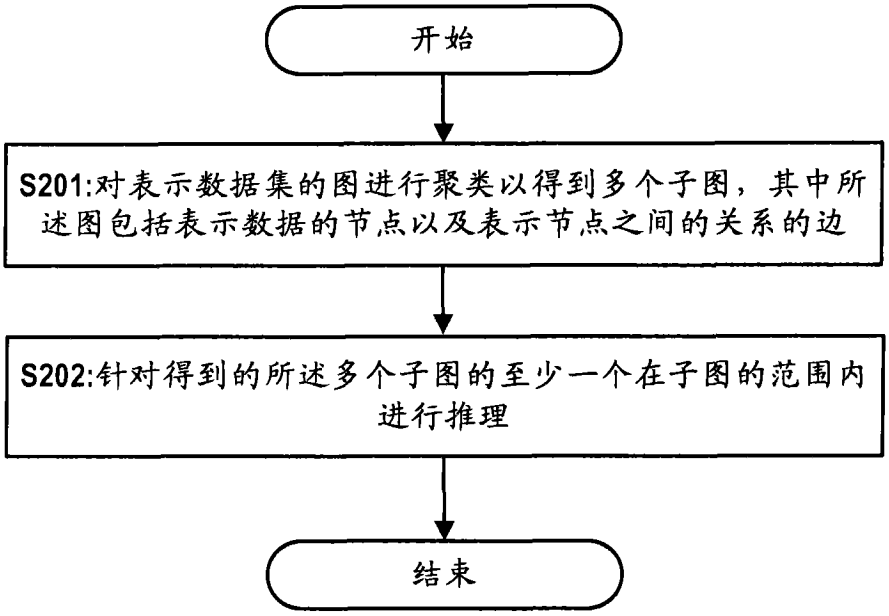

开始
S201:对表示数据集的图进行聚类以得到多个子图，其中所述图包括表示数据的节点以及表示节点之间的关系的边
S202:针对得到的所述多个子图的至少一个在子图的范围内进行推理
结束

图 2

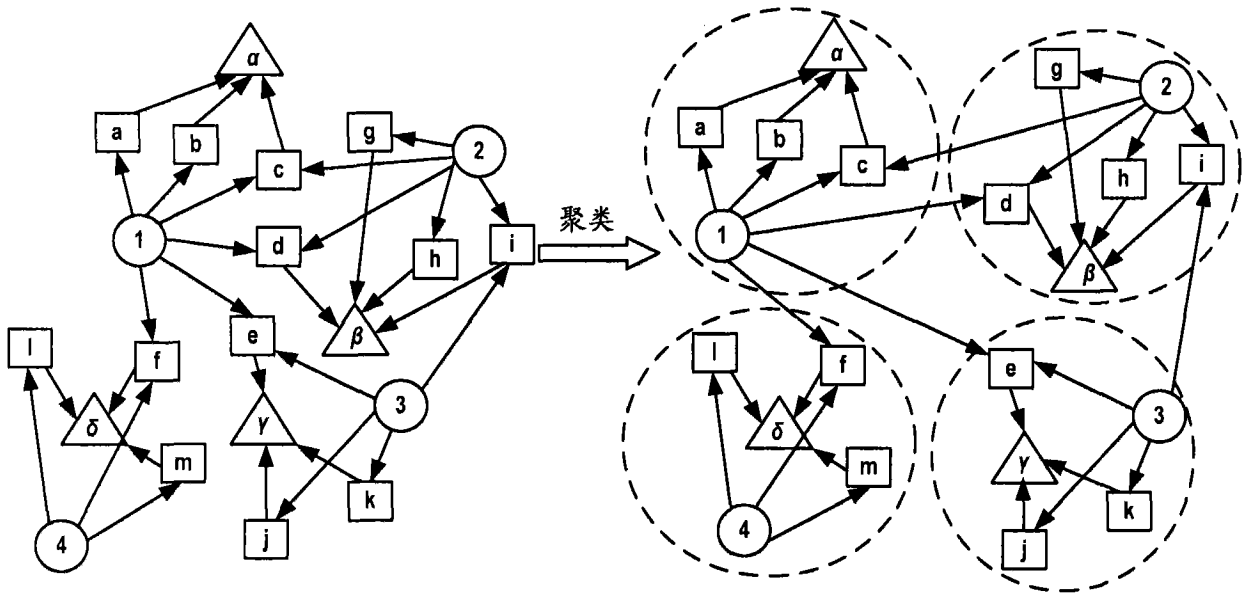

聚类

图 3

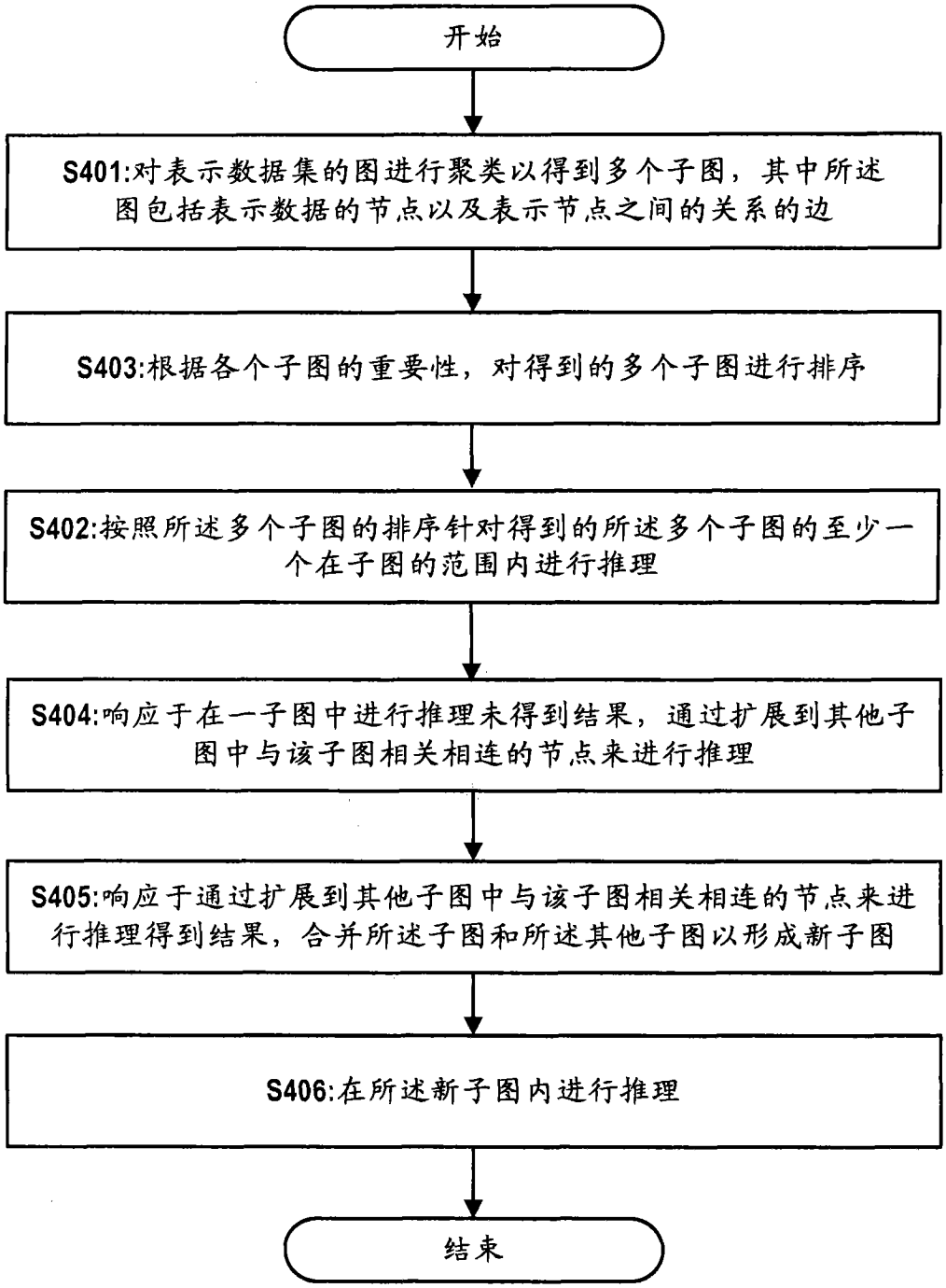

开始
S401:对表示数据集的图进行聚类以得到多个子图，其中所述图包括表示数据的节点以及表示节点之间的关系的边
S403:根据各个子图的重要性，对得到的多个子图进行排序
S402:按照所述多个子图的排序针对得到的所述多个子图的至少一个在子图的范围内进行推理
S404:响应于在一子图中进行推理未得到结果，通过扩展到其他子图中与该子图相关相连的节点来进行推理
S405:响应于通过扩展到其他子图中与该子图相关相连的节点来进行推理得到结果，合并所述子图和所述其他子图以形成新子图
S406:在所述新子图内进行推理
结束

图 4

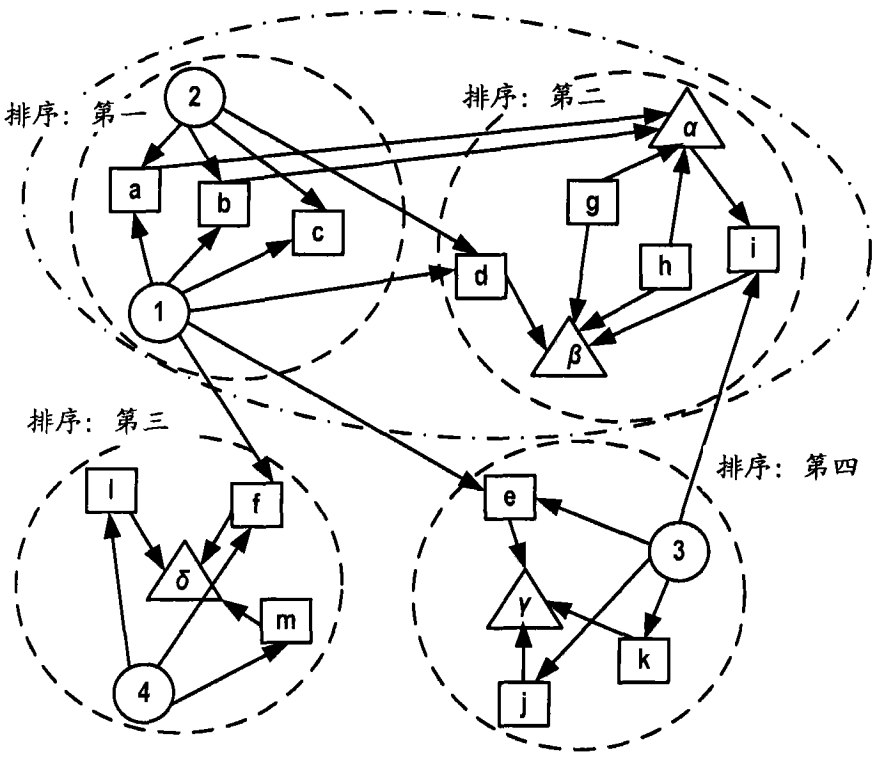

排序：第一
排序：第二
排序：第三
排序：第四
1
2
3
4
a
b
c
d
e
f
g
h
i
j
k
l
m
α
β
γ
δ

图 5

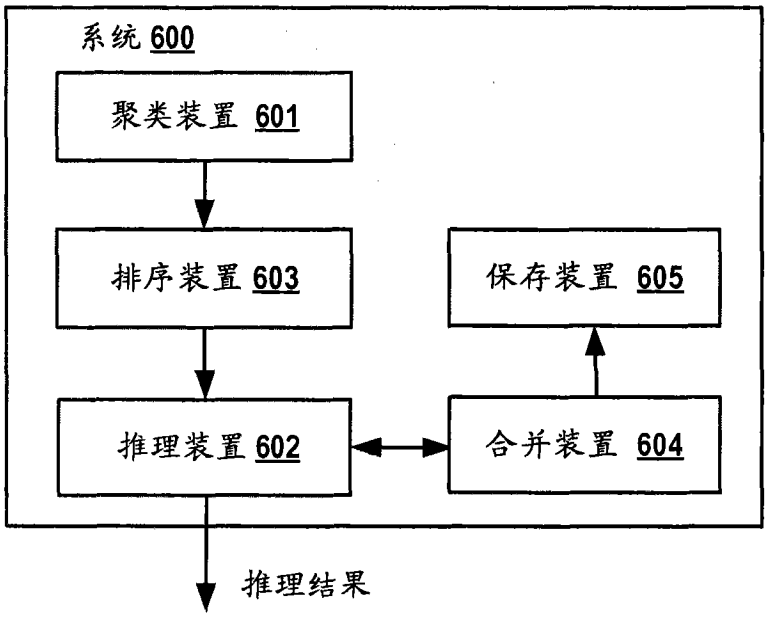

系统 600
聚类装置 601
排序装置 603
推理装置 602
合并装置 604
保存装置 605
推理结果

图 6