

一种对知识图谱进行数据分片的方法及装置

本申请要求于2022年03月28日提交中国国家知识产权局、申请号为202210312004.8、申请名称为“一种对知识图谱进行数据分片的方法及装置”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本说明书一个或多个实施例涉及数据处理技术领域，尤其涉及一种对知识图谱进行数据分片的方法及装置。

背景技术

知识图谱是采用由节点和边构成的多关系图的形式，对知识进行表达的一种知识库。通常，知识图谱采用节点来代表实体，用节点与节点之间的边来表达实体之间的“关系”。实体指的是现实世界中的事物，例如人、地名、概念、药物、公司等，边则用来表达不同实体之间的某种联系。例如，知识图谱中“张三”-“居住在”-“北京”这样的一条边，包含了两个端节点。将知识采用知识图谱的形式进行表达，能够应用在搜索、信息查询等领域，从而大大提高搜索、查询的准确性。

通常，大规模知识图谱中会包含大量的边和节点，其超大的数据量无法在一个设备中存储，需要将知识图谱分别存储在不同的设备中，通过分布式存储，解决数据存储和数据查询等需求。为了将大规模知识图谱进行分布式存储，需要对大规模知识图谱进行数据分片，以便使得多个设备分别得到符合要求的数据分片。

因此，希望能有改进的方案，可以较好地控制对知识图谱的数据分片过程，使得多个设备分得的分片数据更加均衡。

发明内容

本说明书一个或多个实施例描述了一种对知识图谱进行数据分片的方法及装置，以较好地控制对知识图谱的数据分片过程，使得多个设备分得的分片数据更加均衡。具体的技术方案如下。

第一方面，实施例提供了一种对知识图谱进行数据分片的方法，用于将知识图谱拆分成多个分片数据，多个分片数据分别归属于多个设备，所述知识图谱包括代表实体的多个节点，以及体现节点与节点之间关系的边；所述方法通过所述多个设备中任意的第一设备执行，包括：

获取所述知识图谱的第一部分边，所述第一部分边是对所述知识图谱中的多个边进行初始拆分后得到；基于第一扩散速度，从所述第一部分边的端节点中选择扩散节点；获取所述知识图谱中以所述扩散节点为一侧端节点的边，作为待分片边；将所述待分片边中的目标边加入第一分片数据中；其中，所述第一分片数据归属于所述第一设备，所述第一分片数据中包含已分片边；获取其他设备的分片数据中已分片边包含的端节点，作为其他设备的已分片节点；基于所述第一设备的已分片节点与所述其他设备的已分片节点的对比，对所述第一扩散速度进行调整；基于调整后的第一扩散速度继续选择扩散节点，返回执行所述获取所述知识图谱中以所述扩散节点为一侧端节点的边的步骤。

在一种实施方式中，第一扩散速度的取值在(0,1]之间，用于表示选择的数量比例。

在一种实施方式中，所述基于第一扩散速度，从所述第一部分边的端节点中选择扩散

节点的步骤，包括：从所述第一部分边的端节点中选择第一数量个节点作为初始边界点；按照初始边界点关联的边数量从少到多，对多个初始边界点进行排序；基于第一扩散速度，从排序后的多个初始边界点中选择扩散节点。

5 在一种实施方式中，采用以下方法确定所述初始边界点关联的边数量：获取其他设备中以所述初始边界点作为一侧端节点的边；其中，所述其他设备包括所述多个设备中除所述第一设备之外的设备，获取的边是所述其他设备从其自身拥有的部分边中确定；针对任意一个初始边界点，基于所述第一部分边中以该初始边界点作为一侧端节点的边数量，以及所述其他设备中以该初始边界点作为一侧端节点的边数量之和，确定该初始边界点关联的边数量。

10 在一种实施方式中，所述获取所述知识图谱中以所述扩散节点为一侧端节点的边的步骤，包括：从其他设备拥有的部分边中获取以所述扩散节点为一侧端节点的边；将获取的边以及所述第一部分边中以所述扩散节点为一侧端节点的边，确定为所述知识图谱中以所述扩散节点为一侧端节点的边。

15 在一种实施方式中，采用以下方式从所述待分片边中确定目标边：基于所述第一扩散速度，从所述待分片边中选择目标边。

在一种实施方式中，该方法还包括：接收其他设备发送的获取请求，所述获取请求用于获取所述第一设备的已分片节点；将所述第一设备的已分片节点发送至所述其他设备，以使所述其他设备基于所述第一设备的已分片节点对所述其他设备的扩散速度进行调整。

20 在一种实施方式中，所述基于调整后的第一扩散速度继续选择扩散节点的步骤，包括：基于调整后的第一扩散速度，从所述目标边的另一侧端节点中选择扩散节点；或者，基于调整后的第一扩散速度，从所述第一部分边中尚未选择的端节点中选择扩散节点。

25 在一种实施方式中，所述基于所述第一设备的已分片节点与所述其他设备的已分片节点的对比，对所述第一扩散速度进行调整的步骤，包括：当基于所述第一设备的已分片节点数量和所述其他设备的已分片节点数量的对比，确定所述第一设备的节点分片进度大于第一预设进度时，利用第一校正因子减小所述第一扩散速度。

在一种实施方式中，采用以下方式确定第一设备的节点分片进度大于第一预设进度：当第一设备的已分片节点数量大于多个设备的平均已分片节点数量，且第一设备的节点均衡度大于预设节点均衡度时，确定第一设备的节点分片进度大于所述第一预设进度；其中，所述平均已分片节点数量和所述节点均衡度基于所述多个设备的已分片节点数量确定。

30 在一种实施方式中，该方法还包括：当基于所述第一设备的已分片节点数量和所述其他设备的已分片节点数量的对比，确定所述第一设备的节点分片进度小于第二预设进度时，利用所述第一校正因子增大所述第一扩散速度。

35 在一种实施方式中，采用以下方式确定第一设备的节点分片进度小于第二预设进度：当第一设备的已分片节点数量不大于平均已分片节点数量，且所述多个设备中的最大节点均衡度大于预设节点均衡度时，确定所述第一设备的节点分片进度小于所述第二预设进度。

在一种实施方式中，所述利用第一校正因子减小所述第一扩散速度的步骤，包括：

按照第一校正因子的对数规律减小所述第一扩散速度。

在一种实施方式中，所述第一校正因子基于所述第一设备的已分片节点数量和所述多个设备的平均已分片节点数量的对比确定。

在一种实施方式中，在对所述第一扩散速度进行调整之前，还包括：

获取其他设备的分片数据中的已分片边；

所述对所述第一扩散速度进行调整的步骤，包括：

- 5 基于所述第一设备的已分片节点和所述其他设备的已分片节点的对比，以及所述第一设备的已分片边和所述其他设备的已分片边的对比，对所述第一扩散速度进行调整。

在一种实施方式中，所述对所述第一扩散速度进行调整的步骤，包括：基于所述第一设备的已分片节点和所述其他设备的已分片节点的对比，对所述第一扩散速度进行初步调整；基于所述第一设备的已分片边和所述其他设备的已分片边的对比，继续对调整后的第一扩散速度进行调整。

- 10 在一种实施方式中，所述基于所述第一设备的已分片边和所述其他设备的已分片边的对比，继续对调整后的第一扩散速度进行调整的步骤，包括：当基于所述第一设备的已分片边数量和所述其他设备的已分片边数量的对比，确定所述第一设备的边分片进度大于第三预设进度时，利用第二校正因子减小调整后的第一扩散速度。

在一种实施方式中，采用以下方式确定所述第一设备的边分片进度大于第三预设进度：

- 15 当所述第一设备的已分片边数量大于所述多个设备的平均已分片边数量，且所述第一设备的边均衡度大于预设边均衡度时，确定所述第一设备的边分片进度大于第三预设进度；其中，所述平均已分片边数量和所述边均衡度基于所述多个设备的已分片边数量确定。

- 20 在一种实施方式中，该方法还包括：当基于所述第一设备的已分片边数量和所述其他设备的已分片边数量的对比，确定所述第一设备的边分片进度小于第四预设进度时，利用所述第二校正因子增大调整后的第一扩散速度。

在一种实施方式中，采用以下方式确定所述第一设备的边分片进度小于第四预设进度：

当所述第一设备的已分片边数量不大于平均已分片边数量，且所述多个设备的最大边均衡度大于预设边均衡度时，确定所述第一设备的边分片进度小于第四预设进度。

- 25 在一种实施方式中，所述利用第二校正因子减小调整后的第一扩散速度的步骤，包括：按照第二校正因子的对数规律减小调整后的第一扩散速度。

在一种实施方式中，所述第二校正因子基于所述第一设备的已分片边数量和所述多个设备的平均已分片边数量的对比确定。

- 30 第二方面，实施例提供了一种对知识图谱进行数据分片的装置，用于将知识图谱拆分成多个分片数据，多个分片数据分别归属于多个设备，所述知识图谱包括代表实体的多个节点，以及体现节点与节点之间关系的边；所述装置部署在所述多个设备中任意的第一设备中，包括：

- 35 第一获取模块，配置为，获取所述知识图谱的第一部分边，所述第一部分边是对所述知识图谱中的多个边进行初始拆分后得到；第一选择模块，配置为，基于第一扩散速度，从第一部分边的端节点中选择扩散节点；第二获取模块，配置为，获取所述知识图谱中所述扩散节点为一侧端节点的边，作为待分片边；第一分片模块，配置为，将所述待分片边中的目标边加入第一分片数据中；其中，所述第一分片数据归属于所述第一设备，所述第一分片数据中包含已分片边；第三获取模块，配置为，获取其他设备的分片数据中已分片边包含的端节点，作为其他设备的已分片节点；第一调整模块，配置为，基于所述第一设备的已分片节点与所述其他设备的已分片节点的对比，对所述第一扩散速度进行调整；

第二选择模块，配置为，基于调整后的第一扩散速度继续选择扩散节点，返回执行所述第二获取模块。

第三方面，实施例提供了一种计算机可读存储介质，其上存储有计算机程序，当所述计算机程序在计算机中执行时，令计算机执行第一方面中任一项所述的方法。

5 第四方面，实施例提供了一种计算设备，包括存储器和处理器，所述存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现第一方面中任一项所述的方法。

本说明书实施例提供的方法及装置中，在对知识图谱进行数据分片时，以扩散节点开始沿着其邻居节点的方向进行扩散分片。设备可以基于自身的已分片节点和其他设备的已分片节点的对比，对扩散速度进行调整，通过对扩散速度的控制，实现对已分片节点数量的控制。也就是说，本说明书实施例基于边的数量对知识图谱进行拆分，并通过对多个设备中已分片节点的对比，自适应地修改扩散速度，使得多个设备分得的节点数量达到要求的均衡度，进而使得多个设备分得的片数据中的节点数量和边数量更加均衡。

附图说明

15 为了更清楚地说明本发明实施例的技术方案，下面将对实施例描述中所需要使用的附图作简单的介绍。显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他附图。

图1为本说明书披露的一个实施例的实施场景示意图；

图2为实施例提供的一种对知识图谱进行数据分片的方法的流程示意图；

图3为实施例提供的另一种对知识图谱进行数据分片的方法的流程示意图；

20 图4为实施例提供的一种对知识图谱进行数据分片的装置的示意性框图；

图5为实施例提供的另一装置实施例的示意性框图。

具体实施方式

下面结合附图，对本说明书提供的方案进行描述。

25 图1为本说明书披露的一个实施例的实施场景示意图。其中，知识图谱包含代表实体的节点和表示节点与节点之间关系的边，且包含多个节点和多个边。图1中用圆圈代表节点，其中的数字代表节点编号，例如节点1~29，两个节点之间的连线代表边，例如“1-2”“1-3”等代表两条边。首先，对知识图谱的全量数据进行初始拆分，将知识图谱的所有边随机平均，并存储在设备1~设备3中，例如设备1分得的部分边包括1-2、1-8、3-7、8-11、8-15、12-14、15-16、18-19和20-21，设备2和设备3也分别分得部分边。接着，
30 每一个设备，基于一定的扩散速度（表示选择的数量比例），从自身拥有的部分边包含的端节点中选择一定数量比例的扩散节点，从这些扩散节点开始扩散，并将扩散的边确定为自身的分片数据。例如，设备1从节点1和8开始，沿着边的方向进行扩散。在每次扩散中，设备之间交互已分片节点的数量，并基于设备之间已分片节点的数量差异，对扩散速度进行调整，最终使得多个设备的分片数据中的节点数量比较均衡，并且，有邻居关系的
35 边尽可能地分到同一设备中。图1中的知识图谱和多个设备等信息仅仅是一种举例，并不构成对本申请的限定。

下面结合图1对本申请的相关概念和实施场景进行详细说明。

知识图谱是一种以图形式表达的知识库，能够将庞大且繁杂的知识以更加有序的方式表达出来。知识图谱可以应用多个领域，例如可以应用在基于语义的搜索领域，应用在

推荐领域，或应用在生成用户画像等领域。当应用在搜索领域时，可以从知识图谱中搜索待搜索的实体，根据实体节点之间的关系获得与待搜索实体相关的数据。当应用在推荐领域时，可以从知识图谱中确定待推荐实体，根据实体节点之间的关系获得与待推荐实体相关的数据，并基于该数据对待推荐实体进行推荐。在生成用户画像时，可以利用实体节点之间的关系获得实体节点的相关数据，利用该相关数据生成用户画像。

知识图谱包含多个节点以及节点之间的连接边，节点代表实体（entity），因此也可以将节点称为实体节点，节点之间的连接边用于表示实体之间的关系（relation）。实体指的是现实世界中的事物，例如人、地名、概念、药物、公司、组织、机构、设备、数字、日期、货币和地址，等等，不胜枚举。实体可以采用实体词来表示，实体词具有名词性质。例如，用户的昵称张三、地址北京等都是实体。关系用来表达不同实体之间的某种联系，例如“张三”-“居住在”-“北京”中，关系是“居住在”，体现了张三居住在北京这样的关系数据。

知识图谱可以利用业务数据进行构建，例如可以利用与以下对象相关的业务数据进行构建：店铺、用户、商品和事件等等。在大规模知识图谱中，节点和边的数量非常多，通常无法由一个设备对其进行存储。为了适应大规模知识图谱的存储和查询需求，可以将知识图谱分别存储在不同的设备中，通过分布式存储，解决数据存储和数据查询的需求。

通常，可以将大规模知识图谱的数据分别存储在多个设备中，这些设备的存储空间和计算能力等配置基本相同。在处理针对大规模知识图谱的查询等请求时，需要在这多个设备中进行，这就存在负载均衡的要求。也就是说，存储在多个设备中的知识图谱的分片数据的边数量和节点数量应大致均衡，这是对知识图谱进行拆分时的一个要求，即节点和边的均衡原则。

另一方面，为了提高查询等处理的效率，应使得相邻的节点和相邻的边尽可能地拆分在同一设备的分片数据中。这是对知识图谱进行拆分时的另一个要求，即邻居扩散原则。

为了较好地控制对知识图谱的数据分片过程，使得多个设备分得的分片数据更加均衡，并且尽量满足邻居扩散原则，本实施例提供了一种对知识图谱进行数据分片的方法。该方法通过多个设备中任意的第一设备执行，包括：步骤 S210，获取知识图谱的第一部分边，所述第一部分边是对所述知识图谱中的多个边进行初始拆分后得到；步骤 S220，基于第一扩散速度，从第一部分边的端节点中选择扩散节点；步骤 S230，获取知识图谱中以扩散节点为一侧端节点的边，作为待分片边；步骤 S240，将待分片边中的目标边加入第一分片数据中，其中，第一分片数据归属于第一设备，第一分片数据中包含已分片边；步骤 S250，获取其他设备的分片数据中已分片边包含的端节点，作为其他设备的已分片节点；步骤 S260，基于第一设备的已分片节点与其他设备的已分片节点的对比，对第一扩散速度进行调整；步骤 S270，基于调整后的第一扩散速度继续选择扩散节点，返回执行步骤 S230 获取知识图谱中以扩散节点为一侧端节点的边。

在本实施例中，在对知识图谱的边进行扩散分片的过程中，第一设备会基于自身的已分片节点与其他设备的已分片节点之间的差异，对自身的扩散速度进行调整，动态地调整扩散速度，使得自身的已分片节点与其他设备的已分片节点达到相对均衡的状态。

在知识图谱中，节点代表实体，每一个节点可以包括实体类型和实体属性等数据；节点与节点之间的边可以包括关系类型和关系属性等数据。将知识图谱拆分成多个分片数据，

包括对知识图谱包含的所有数据的划分，但是对知识图谱中所有数据的划分依托于对节点和边的划分，而节点的划分依托于对边的划分。因此，本实施例可以基于对知识图谱中的边划分，实现对知识图谱整体数据的拆分。并且，在对知识图谱进行数据分片的过程中，可以先对以编号形式表示的节点和边进行划分，当此过程完成时，再将知识图谱中相应编号的节点和边的数据存储在对应的设备中。

下面结合图 2 对上述实施例进行详细说明。

图 2 为实施例提供的一种对知识图谱进行数据分片的方法的流程示意图。该方法用于将知识图谱拆分成多个分片数据，多个分片数据分别归属于多个设备。多个设备的存储空间、计算能力等配置可以相同，也可以不同。多个设备也可以理解为逻辑意义上的多个设备。任意一个设备可以通过任何具有计算、处理能力的装置、设备、平台、设备集群等来实现。

知识图谱包括代表实体的多个节点，以及体现节点与节点之间关系的边。该知识图谱是待拆分的知识图谱，初始时，知识图谱的全量数据可以存储在一个超级计算机中，也可以分别不等数据量地存储在多个设备中。本方法实施例通过 N 个设备中任意的第一设备 A 执行， N 个设备均可以执行与第一设备 A 相同的过程，进而实现 N 个设备得到的分片数据是比较均衡的。 N 可以是 2 或大于 2 的整数。本方法包括以下步骤。

在步骤 S210 中，第一设备 A 获取知识图谱的第一部分边 `edges1`。第一部分边 `edges1` 是对知识图谱中的多个边进行初始拆分后得到。同样的， N 个设备中除第一设备 A 之外的其他设备，也分别获取到部分边。多个设备获取的知识图谱的部分边，可以是对知识图谱的全量边进行随机平均分而得到的。

第一部分边 `edges1` 并不是最终分给第一设备 A 的分片数据，因为第一部分边 `edges1` 虽然可以与其他设备分得的部分边之间达到了边数量上的均衡，但是尚未达到节点数量上的均衡，也不符合邻居扩散原则。在后续的处理步骤中，基于 N 个设备分别分得的部分边，对知识图谱的边进行重新分配，使得每个设备最终得到的分片数据中的节点和边满足均衡度原则和邻居扩散原则。

知识图谱中的边可以采用两个节点编号来表示，例如节点 1 和节点 2 之间的边可以表示为 1-2。大规模知识图谱可以包含大量的边，有时候是百亿甚至千亿级别的边数量。在本步骤中，对知识图谱中的多个边进行初始拆分，将拆分后的 N 部分边存储在 N 个设备中。例如，知识图谱包含 100 亿个边，有 10 个设备，则每个设备大致存储 10 亿个边，每个设备得到部分边。对知识图谱的多个边进行的初始拆分可以是随机进行的，并不符合邻居扩散原则。

在一种实施方式中，初始时主设备存储有知识图谱的全量边，主设备基于 N 个设备的数量 N 对知识图谱的全量边进行拆分，例如随机拆分或者按照先后顺序拆分，得到 N 部分边，将这 N 部分边分别发送至 N 个设备中。为了使得描述更方便，第一设备 A 获取的部分边作为第一部分边 `edges1`。

在一种实施方式中，初始时， N 个设备中可以分别存储有知识图谱的部分边。 N 个设备中的部分边可能并不是平均分的。在这种情况下， N 个设备可以通过相互的通信，使得 N 个设备分别得到大致平均分的部分边。

例如，在图 1 所示的场景示意图中，假设知识图谱包含 27 个边，将这 27 个边随机平

均分配至设备 1、设备 2 和设备 3，每个设备分别得到 9 个边。

在步骤 S220 中，第一设备 A 基于第一扩散速度 V_e ，从第一部分边 edges1 的端节点中选择扩散节点。第一扩散速度 V_e 也可以称为第一扩散比例，用于表示选择的数量比例，例如可以用于表示选择的节点数量比例，或者用于表示选择的边数量比例。

5 例如，在图 1 中，设备 1 的部分边中包含端节点 1、2、3、7、8、11、12、14、15、16、18、19、20 和 21。基于第一扩散速度 V_e ，从这些端节点中选择扩散节点，也就是按照第一扩散速度的比例，从这些端节点中选择一定数量的端节点，作为扩散节点。第一扩散速度的取值可以在 (0, 1] 之间，即第一扩散速度可以是大于 0、小于等于 1 的值。

本实施例中，N 个设备中的其他设备也分别基于其扩散速度，从其部分边的端节点中
10 选择扩散节点。初始时，N 个设备的扩散速度可以是相同的，例如都可以设置为 0.1。

在一种实施方式中，步骤 S220 在执行时，可以包括以下步骤 1a~3a。

步骤 1a，从第一部分边 edges1 的端节点中选择第一数量个节点作为初始边界点，得到多个初始边界点。该选择操作可以是随机选择。第一数量可以是预设值，也可以在每次扩散迭代执行时进行修改。N 个设备的第一数量可以设置为相同值。

15 步骤 2a，按照初始边界点关联的边数量从少到多，对多个初始边界点进行排序。

初始边界点关联的边数量，可以理解为以初始边界点为一侧端节点的所有边的数量，也可以称为度 (degree)。按照度从小到大，对多个初始边界点进行排序。

步骤 3a，基于第一扩散速度 V_e ，从排序后的多个初始边界点中选择扩散节点。在选择时，可以从排序后的序列的起始位置开始选择，按照第一扩散速度 V_e 的值，选择度较小的
20 初始边界点作为扩散节点。

本实施方式中，在选择扩散节点时，从度小的节点开始进行扩散，能够避免在初始时将超级热点选择为扩散节点。超级热点，是存在大量关联边的节点。

在上述步骤 2a 中，可以采用以下操作确定初始边界点关联的边数量：

第一设备 A 获取其他设备中以初始边界点作为一侧端节点的边，针对任意一个初始边界点，基于第一部分边 edges1 中以该初始边界点作为一侧端节点的边数量，以及其他设备中以该初始边界点作为一侧端节点的边数量之和，确定该初始边界点关联的边数量。
25

其中，上述其他设备包括 N 个设备中除第一设备 A 之外的设备，获取的边是其他设备从其自身拥有的部分边中确定。

第一设备 A 可以针对自身的初始边界点，生成获取请求，并将该获取请求发送至其他设备。该获取请求用于获取其他设备中以该初始边界点作为一侧端节点的边，其中可以携带该初始边界点的编号。其他设备在接收到该获取请求时，将自身拥有的部分边中，以获取请求携带的初始边界点作为一侧端节点的边发送至第一设备 A。其他设备发送至第一设备 A 的边，是从其他设备的未分片中确定的，已分片边不会再发送至第一设备 A。
30

同样的，其他设备也可以向第一设备 A 发送获取请求，以获取以其他设备中的初始边界点作为一侧端节点的边，第一设备 A 也响应其他设备发送的获取请求。
35

例如，在图 1 中，设备 1 将节点 1 和节点 8 作为扩散节点，从自身的部分边中确定以扩散节点作为一侧端节点的边包括：1-2、1-8、8-11 和 8-15。设备 1 可以从设备 2 的部分边中获取以节点 1 和节点 8 作为一侧端节点的边，包括：1-4、8-9、8-10 和 8-12，从设备 3 的部分边中获取以节点 1 和节点 8 作为一侧端节点的边，包括：1-3。假设，设备 1

发送的针对节点 1 和节点 8 的获取请求先于设备 2 和设备 3 发送的具有相同作用的获取请求。也就是说，谁先对某个节点进行扩散，那么就先得到该节点关联的边。

在上述步骤 2a 中，也可以采用其他方式确定初始边界点关联的边数量。例如，第一设备 A 可以从包含知识图谱的全量边的主设备中，获取以初始边界点为一侧端节点的所有边。主设备在将以初始边界点为一侧端节点的所有边发送至第一设备 A 之后，还可以向其他设备发送通知消息，以使得其他设备基于该通知消息，将自身拥有的对应的边进行删除。

在上述针对步骤 2a 的实施方式中，第一设备 A 也可以仅获取以初始边界点为一侧端节点的边的数量值。

当第一设备 A 获取到以初始边界点为一侧端节点的所有边时，可以将这些边加入第一部分边 edges1 中。

在步骤 S230 中，第一设备 A 获取知识图谱中以扩散节点为一侧端节点的边，作为待分片边。知识图谱中以扩散节点为一侧端节点的边，包含了以扩散节点为一侧端节点的所有边。例如，在图 1 所示示意图中，假设节点 1 为第一设备 A 的一个扩散节点，则边 1-2、1-3、1-4 和 1-8 即是该以扩散节点 1 为一侧端节点的所有边。

第一设备 A 可以从其他设备拥有的部分边中获取以扩散节点为一侧端节点的边，将获取的边以及第一部分边 edges1 中以扩散节点为一侧端节点的边，确定为知识图谱中以扩散节点为一侧端节点的边。

如果在执行步骤 S220，确定初始边界点关联的边数量时，已经从其他设备中获取到以初始边界点作为一侧端节点的边，由于扩散节点是从初始边界点中选择的，因此可以直接从步骤 S220 获取的以初始边界点作为一侧端节点的边中，获取以扩散节点为一侧端节点的边。

第一设备 A 也可以采用其他方式获取知识图谱中以扩散节点为一侧端节点的边，例如，可以从包含知识图谱的全量边的主设备中，获取以扩散节点为一侧端节点的所有边。主设备在将以扩散节点为一侧端节点的所有边发送至第一设备 A 之后，还可以向其他设备发送通知消息，以使得其他设备基于该通知消息，将自身拥有的对应的边进行删除。

在步骤 S240 中，第一设备 A 将待分片边中的目标边加入第一分片数据 data1 中。

其中，第一分片数据 data1 归属于第一设备 A，第一分片数据 data1 中包含已分片边，已分片边的端节点可以称为已分片节点，第一分片数据 data1 中也可以包含已分片节点。通常，待分片边的数量是非常多的。第一设备 A 可以将多个待分片边都作为目标边，加入第一分片数据 data1 中，也可以从多个待分片边中选择一定数量的待分片边，作为目标边加入第一分片数据 data1 中。例如，可以基于第一扩散速度 V_e ，从多个待分片边中选择目标边，也就是将多个分片边中一定比例的边加入第一分片数据 data1 中。在实施时，可以基于第一扩散速度 V_e ，从多个待分片边中随机选择目标边，也可以顺序选择目标边。

将目标边加入第一分片数据 data1 中，可以通过将目标边的状态修改为已分片实现。第一设备 A 中未加入第一分片数据 data1 中的边，其状态可以是未分片。

第一分片数据 data1 是归属于第一设备 A 的数据，是已经过分片之后的数据，其他设备一般不能再从第一分片数据 data1 中获取数据。当第一设备 A 接收到其他设备发送的用于获取以某个节点为一侧端节点的边的获取请求时，第一设备 A 从处于未分片状态的边中确定。

同样的，N 个设备中的各个其他设备，分别拥有归属于其自身的分片数据。随着扩散操作的不断进行，逐渐地向分片数据中添加已分片边，直至将知识图谱的所有边分别加入 N 个设备的分片数据中。

在步骤 S250 中，第一设备 A 获取其他设备的分片数据中已分片边包含的端节点，作为其他设备的已分片节点。其他设备是指 N 个设备中除第一设备 A 之外的设备。

第一设备 A 可以分别向其他设备发送获取请求，用于获取其他设备的已分片节点。其他设备在接收到第一设备 A 的该获取请求时，从自身的分片数据中确定已分片节点，并将其发送至第一设备 A。第一设备 A 获取的其他设备的已分片节点，可以理解为获取的是其他设备的已分片节点的数量。

第一设备 A 还可以接收其他设备发送的获取请求，该获取请求用于获取第一设备 A 的已分片节点。第一设备 A 可以将第一设备 A 的已分片节点发送至其他设备，以使其他设备基于第一设备 A 的已分片节点对其他设备的扩散速度进行调整。设备之间交互的已分片节点均可以是已分片节点数量。

在步骤 S260 中，第一设备 A 基于第一设备 A 的已分片节点与其他设备的已分片节点的对比，对第一扩散速度 V_e 进行调整。

当基于第一设备 A 的已分片节点数量 N_A^V 和其他设备的已分片节点数量的对比，确定第一设备 A 的节点分片进度大于第一预设进度时，说明第一设备 A 的节点分片进度过快，可以利用第一校正因子 D_e 减小第一扩散速度 V_e 。当确定第一设备 A 的节点分片进度不大于第一预设进度时，可以保持第一扩散速度 V_e 不变。

具体的，第一设备 A 可以采用以下方式确定第一设备 A 的节点分片进度大于第一预设进度：

当第一设备 A 的已分片节点数量 N_A^V 大于多个设备的平均已分片节点数量 $N_{均}^V$ ，且所述第一设备 A 的节点均衡度 M_A^V 大于预设节点均衡度 B_e 时，确定第一设备 A 的节点分片进度大于第一预设进度。

其中，预设节点均衡度 B_e 可以是预先根据经验设置的阈值。平均已分片节点数量 $N_{均}^V$ 和第一设备 A 的节点均衡度 M_A^V 都是基于多个设备的已分片节点数量确定的。平均已分片节点数量 $N_{均}^V$ 是多个设备的已分片节点数量的平均值。第一设备 A 的节点均衡度 M_A^V 可以按照以下式 (1) 计算：

$$M_A^V = \frac{N_A^V}{\min(N_i^V)} \quad (1)$$

其中， N_i^V 是第 i 个设备的已分片节点数量， $\min(N_i^V)$ 是第 i 个设备中的已分片节点数量的最小值，参量的上角标 V 代表该参量与节点相关。第一设备 A 的节点均衡度，表示的是该第一设备 A 的已分片节点在多个设备的已分片节点中的均衡程度值。

节点均衡度还可以采用其他公式确定，例如利用已分片节点数量 N_A^V 与平均已分片节点数量 $N_{均}^V$ 的差值确定，例如可以采用该差值与平均已分片节点数量 $N_{均}^V$ 的比值确定等。

同时比较已分片节点数量与平均值、节点均衡度与阈值，当已分片节点数量大于平均值，且节点均衡度大于阈值时，确定第一设备 A 的节点分片进度大于第一预设进度。采用上述两个比较，能够排除已分片节点数量比平均值大一点点，但是节点均衡度不大于阈值的设备，这样的设备的扩散速度不必减小。本实施例中，第一预设进度并不是一个具体的

数值,而是一种复杂的特殊状态,是第一设备A相对于其他设备的超前扩散状态。

当基于第一设备A的已分片节点数量 N_A^v 和其他设备的已分片节点数量的对比,确定第一设备A的节点分片进度小于第二预设进度时,说明第一设备A的节点分片进度过慢,可以利用第一校正因子 D_v 增大第一扩散速度 V_e 。当确定第一设备A的节点分片进度不小于第二预设进度时,可以保持第一扩散速度 V_e 不变。

具体的,第一设备A可以采用以下方式确定第一设备A的节点分片进度小于第二预设进度:

当第一设备A的已分片节点数量 N_A^v 不大于平均已分片节点数量 $N_{均}^v$,且多个设备中的最大节点均衡度 $\max(M_i^v)$ 大于预设节点均衡度 B_v 时,确定第一设备A的节点分片进度小于第二预设进度。当已分片节点数量大于平均值,且最大的节点均衡度大于阈值时,说明不仅仅第一设备A的已分片节点数量低于平均值,连最大的节点均衡度也已经大于阈值了,已经有设备的节点均衡度超前了。此时无需比较第一设备A的节点均衡度与阈值的大小,因为其节点均衡度一般来说是比较小的。

本实施例中,第二预设进度并不是一个具体的值,而是一种复杂的特殊状态,是第一设备A相对于其他设备的落后扩散状态。第二预设进度小于第一预设进度。

下面再详细说明当确定第一设备A的节点分片进度大于第一预设进度时,如何对第一扩散速度 V_e 进行调整。在一种实施方式中,可以将第一扩散速度暂时设置为0,也就是第一设备A进入等待状态。当判断出第一设备A的节点分片进度不大于第一预设进度时,

为了提高设备的分片速度,可以在确定第一设备A的节点分片进度大于第一预设进度时,利用第一校正因子 D_v 来减小第一扩散速度 V_e 。当确定第一设备A的节点分片进度小于第二预设进度时,利用第一校正因子 D_v 来增大第一扩散速度 V_e 。

在一种实施方式中,可以通过将第一扩散速度 V_e 减去 D_v ,来减小第一扩散速度 V_e ;通过将第一扩散速度 V_e 加上 D_v ,来增大第一扩散速度 V_e 。在这种实施方式中,第一扩散速度 V_e 的改变依赖于对 D_v 的设置。当 D_v 设置得较小,则节点均衡度的收敛速度会相对较慢,当 D_v 设置得较大,则节点均衡度的收敛速度会相对较快。其中, D_v 可以是预设的常数,也可以随着迭代过程的进行而调整,例如随着迭代次数的增加逐渐减小。

在一些知识图谱中,其节点的度(degree)的分布大致呈幂律分布,即度小的节点通常占比非常大,而度大的节点占比非常小。经过申请人的研究发现,不同度的节点的数量与其度呈指数关系。在 V_e 上增加0.1,即 D_v 取0.1,所带来的扩散的边数量并不是上涨10%,而是上涨 $(e^{0.1}-1)*100\%=10.52\%$,而在 V_e 上增加0.3时,上涨幅度为34.99%。反过来,将 V_e 减去0.1时,扩散的边数量大约下降9.52%,而将 V_e 减去0.3时,下降幅度大约为25.92%。因此,对节点的度较大的设备,将其扩散速度降低0.1(或0.3),和对节点的度较小的设备,将其扩散速度增大0.1(或0.3),对扩散时边数量的影响是不一样的。

为了使得对扩散速度的调整速度更加合理,使得数据分片过程的进行更加可控,可以按照第一校正因子 D_v 的对数规律减小第一扩散速度 V_e ;或者,按照第一校正因子 D_v 的对数规律增大第一扩散速度 V_e 。

在一种实施方式中,第一设备A可以按照以下式(2)减小第一扩散速度 V_e :

$$V_e^2 = V_e^1 * [1 - \log_a(1 + D_v)] \quad (2)$$

第一设备A可以按照以下式(3)增大第一扩散速度 V_e :

$$V_e^2 = V_e^1 * [1 + \log_a(1 + D_v)] \quad (3)$$

其中, V_e^2 是调整后的第一扩散速度, V_e^1 是调整之前的第一扩散速度, $\log_a$ 是以 a 为底的对数, a 可以是预设值, 例如可以是自然常数 e 。上述式 (2) 和式 (3) 仅仅是按照第一校正因子 D_v 的对数规律对第一扩散速度 V_e 进行调整时的一种实施方式, 基于这些公式,

5 很容易得出其他形式的实施方式, 例如乘以一个系数或除以一个系数等。

本实施例中, 按照第一校正因子 D_v 的对数规律对第一扩散速度进行调整。对于其中节点的度数量偏大或偏小的设备, 其扩散速度的下降幅度和上升幅度都能达到基本相同, 并且对 D_v 的设置范围也可以更加宽松, 整体的收敛速度也较快。

10 在一种实施场景中, 知识图谱中可能会包含超级热点。超级热点关联的边数量 (也就是度) 巨大, 远远大于其他节点关联的边数量。当超级热点存在于某个设备的部分边中, 并且该超级热点被选为初始分界点或前几次扩散时的扩散节点时, 在超级热点附近确定的目标边很快就能达到较大值, 从而使得已分片节点的数量达到很大值。并且, 随着数据分片过程的进行, 不同设备中的数据分布在不同迭代步数上也可能是不同的。

15 为了使得对第一扩散速度 V_e 的调整更加合理, 避免多个设备的数据分片过程出现较大的偏离, 本实施例可以对第一校正因子 D_v 进行自适应调整。例如, 第一校正因子 D_v 可以基于第一设备 A 的已分片节点数量 N_A^v 和多个设备的平均已分片节点数量 $N_{均}^v$ 的对比确定。

在一种实施方式中, 第一设备 A 可以按照以下式 (4) 减小第一扩散速度 V_e :

$$V_e^2 = V_e^1 * \left\{ 1 - \log_a \left[1 + D_v * \log_b \left(b + \frac{N_A^v - N_{均}^v}{N_{均}^v} \right) \right] \right\} \quad (4)$$

第一设备 A 可以按照以下式 (5) 增大第一扩散速度 V_e :

$$V_e^2 = V_e^1 * \left\{ 1 + \log_a \left[1 + D_v * \log_b \left(b + \frac{N_{均}^v - N_A^v}{N_A^v} \right) \right] \right\} \quad (5)$$

20 其中, V_e^2 是调整后的第一扩散速度, V_e^1 是调整之前的第一扩散速度, $\log_b$ 是以 b 为底的对数, b 可以是预设值, 例如可以是自然常数 e , a 和 b 的取值可以相同, 也可以不同, N_A^v 是第一设备 A 的已分片节点数量, $N_{均}^v$ 是平均已分片节点数量。从式 (4) 和 (5) 可以看出, 第一校正因子 D_v 被乘以了中括号中的对数校正项, 该对数校正项与第一设备 A 的已分片节点数量 N_A^v 和多个设备的平均已分片节点数量 $N_{均}^v$ 的对比相关。该对数校正项以及校正方式仅仅是一种实施方式, 基于这些式子, 可以容易地得出其他形式的实施方式, 例如

25 直接将第一校正因子 D_v 修改为以下形式也是可行的实施方式:

$$D_v * \log_b \left[b + \gamma * \frac{N_A^v - N_{均}^v}{N_{均}^v} \right] \text{ 或 } D_v * \log_b \left[b + \gamma * \frac{N_{均}^v - N_A^v}{N_A^v} \right]$$

其中, γ 为预设系数。

在步骤 S270 中, 第一设备 A 基于调整后的第一扩散速度 V_e 继续选择扩散节点, 返回执行步骤 S230, 即获取知识图谱中以扩散节点为一侧端节点的边。

30 第一设备 A 在继续选择扩散节点时, 可以基于调整后的第一扩散速度 V_e , 从步骤 S240 中目标边的另一侧端节点中选择扩散节点。在步骤 S230 中的扩散边是目标边的一侧端节点, 目标边的另一侧端节点是与步骤 S230 中提到的一侧端节点不同的端节点。从目标边

的另一侧端节点中选择扩散节点，是在按照目标边所指的方向向相邻的邻居边方向进行扩散。

第一设备 A 在基于调整后的第一扩散速度 V_e 从目标边的另一侧端节点中选择扩散节点时，可以包括 1b~3b。

- 5 步骤 1b，从目标边的另一侧端节点中选择第一数量个节点作为边界点，得到多个边界点。对本步骤的说明可以参见步骤 1a，此处不再赘述。

步骤 2b，按照边界点关联的边数量从少到多，对多个边界点进行排序。

在确定边界点关联的边数量，也就是确定度时，可以按照步骤 2a 中确定初始边界点的方式进行，此处不再赘述。

- 10 步骤 3b，基于调整后的第一扩散速度 V_e ，从排序后的多个边界点中选择扩散节点。在选择时，可以从排序后的序列的起始位置开始选择，按照调整后的第一扩散速度 V_e 的值，选择度较小的边界点作为扩散节点。

- 上述步骤 S220~S240 可以理解为一次分片迭代过程（或者称为扩散迭代过程）。步骤 S250~S260 是一次对第一扩散速度的调整过程。在实际应用中，可以每一次分片迭代过程之后都执行一次对第一扩散速度的调整过程，也可以执行多次分片迭代过程之后，执行一次对第一扩散速度的调整过程。在多次执行分片迭代过程时，在步骤 S240 中将目标边加入第一分片数据中之后，可以继续基于第一扩散速度 V_e 从目标边的另一侧端节点中选择扩散节点，返回执行步骤 S230。
- 15

- 当在步骤 S230 中，第一设备 A 获取了知识图谱中以扩散节点为一侧端节点的边，作为待分片边时，还可以判断待分片边的数量是否大于预设阈值，如果大于，则说明待分片边的数量足够多；如果不大于，则第一设备 A 可以在下一个分片迭代过程中，基于第一扩散速度 V_e ，从第一部分边中尚未选择的端节点中选择扩散节点。当知识图谱中的所有边都被分别加入到多个设备的分片数据中时，认为数据分片过程已完成，图 2 所示的迭代过程结束。
- 20

- 图 2 所述实施例仅仅示出一种可行的实施方式的核心思路，在实际应用中可以选择多种具体的操作方式进行。在图 2 所示实施例中，初始时，针对边对知识图谱进行随机的平均划分，将部分边分别存储在多个设备中。在数据分片的过程中，基于第一扩散速度选择扩散节点，体现出了对边的均衡处理。而利用多个设备的已分片节点数量的对比，对第一扩散速度进行调整，则在对边的均衡处理的基础上实现了对节点的均衡处理。
- 25

- 在本说明书的另一实施例中，还可以实现对边更合理的均衡处理。对图 2 所示实施例加以改进，可以得到图 3 所示实施例。图 3 为实施例提供的另一种对知识图谱进行数据分片的方法的流程示意图。其中，图 3 实施例包括步骤 S310~S370，其中，步骤 S310~S350 以及步骤 S380 分别与图 2 所示实施例中的步骤 S210~S250 以及步骤 S270 完全相同，本实施例不再对其赘述。在以下针对图 3 所示实施例的描述中，着重对与图 2 所示实施例的不同之处进行说明，而相同之处可以参见图 2 所示实施例的说明，本实施例不再赘述。
- 30
- 35

步骤 S360，在对第一扩散速度 V_e 进行调整之前，第一设备 A 获取其他设备的分片数据中的已分片边。步骤 S360 可以在步骤 S350 之前或之后或同时执行。

第一设备 A 可以分别向其他设备发送获取请求，用于获取其他设备的已分片边。其他设备在接收到第一设备 A 的该获取请求时，从自身的分片数据中确定已分片边，并将其发

送至第一设备 A。第一设备 A 获取的其他设备的已分片边，可以理解为获取的是其他设备的已分片边的数量。

第一设备 A 还可以接收其他设备发送的获取请求，该获取请求用于获取第一设备 A 的已分片边。第一设备 A 可以将第一设备 A 的已分片边发送至其他设备，以使其他设备基于
5 第一设备 A 的已分片边对其他设备的扩散速度进行调整。设备之间交互的已分片边均可以是已分片边数量。

步骤 S370，第一设备 A，基于第一设备 A 的已分片节点和其他设备的已分片节点的对比，以及第一设备 A 的已分片边和其他设备的已分片边的对比，对第一扩散速度 V_e 进行调整。本步骤在执行时可以包括以下步骤 1c 和 2c。

10 步骤 1c，第一设备 A，基于第一设备 A 的已分片节点和其他设备的已分片节点的对比，对第一扩散速度 V_e 进行初步调整。

步骤 2c，第一设备 A，基于第一设备 A 的已分片边和其他设备的已分片边的对比，继续对调整后的第一扩散速度 V_e 进行调整。

对第一扩散速度 V_e 的调整，也可以先基于已分片边之间的对比进行调整，再基于已分片节点之间的对比进行调整。下面仅以步骤 1c 和 2c 所示的顺序为例，对本实施例中
15 对第一扩散速度 V_e 进行调整的过程进行详细说明。

其中，步骤 1c 的执行过程可以与步骤 S260 的执行过程完全相同，具体说明可以参见图 2 所示实施例的说明，此处不再赘述。下面着重说明步骤 2c。

当基于第一设备 A 的已分片边数量 N_A^e 和其他设备的已分片边数量的对比，确定第一设备 A 的边分片进度大于第三预设进度时，说明第一设备 A 的边分片进度过快，可以利用第二校正因子 D_e 减小调整后的第一扩散速度 V_e 。当确定第一设备 A 的边分片进度不大于第三预设进度时，可以保持调整后的第一扩散速度 V_e 不变，继续执行步骤 S380。

具体的第一设备 A 可以采用以下方式确定第一设备 A 的边分片进度大于第三预设进度：

20 当第一设备 A 的已分片边数量 N_A^e 大于多个设备的平均已分片边数量 $N_{均}^e$ ，且第一设备 A 的边均衡度 M_A^e 大于预设边均衡度 B_e 时，确定第一设备 A 的边分片进度大于第三预设进度。

其中，预设边均衡度 B_e 可以是预先根据经验设置的阈值。平均已分片边数量 $N_{均}^e$ 和边均衡度 M_A^e 都是基于多个设备的已分片边数量确定的。平均已分片边数量 $N_{均}^e$ 是多个设备的已分片边数量的平均值。第一设备 A 的边均衡度 M_A^e 可以按照式 (6) 计算：

$$M_A^e = \frac{N_A^e}{\min(N_i^e)} \quad (6)$$

30 其中， N_i^e 是第 i 个设备的已分片边数量， $\min(N_i^e)$ 是第 i 个设备中的已分片边数量的最小值，参量的上角标 e 代表该参量与边相关。第一设备 A 的边均衡度，表示的是该第一设备 A 的已分片边在多个设备的已分片边中的均衡程度值。

上述边均衡度还可以采用其他公式确定，例如利用已分片边数量 N_A^e 与平均已分片边数量 $N_{均}^e$ 的差值确定，例如可以采用该差值与平均已分片边数量 $N_{均}^e$ 的比值确定等。

35 同时比较已分片边数量与平均值、边均衡度与阈值，当已分片边数量大于平均值，且边均衡度大于阈值时，确定第一设备 A 的边分片进度大于第三预设进度。采用上述两个比较，能够排除已分片边数量比平均值大一点点，但是边均衡度不大于阈值的设备，这样的设备的扩散速度不必因为边而减小。本实施例中，第三预设进度并不是一个具体的数值，

而是一种复杂的特殊状态，是第一设备 A 相对于其他设备的超前扩散状态。

当基于第一设备 A 的已分片边数量 N_A^* 和其他设备的已分片边数量的对比，确定第一设备 A 的边分片进度小于第四预设进度时，说明第一设备 A 的边分片进度过慢，可以利用第二校正因子 D_e 增大调整后的第一扩散速度 V_e 。当确定第一设备 A 的边分片进度不小于第四预设进度时，可以保持调整后的第一扩散速度 V_e 不变。

具体的，第一设备 A 可采用以下方式确定第一设备 A 的边分片进度小于第四预设进度：

当第一设备 A 的已分片边数量 N_A^* 不大于平均已分片边数量 $N_{均}^*$ ，且多个设备的最大边均衡度 $\max(M_i^*)$ 大于预设边均衡度 B_e 时，确定第一设备 A 的边分片进度小于第四预设进度。当已分片边数量大于平均值，且最大的边均衡度大于阈值时，说明不仅仅第一设备 A 的已分片边数量低于平均值，连最大的边均衡度也已经大于阈值了，已经有设备的边均衡度超前了。此时无需比较第一设备 A 的边均衡度与阈值的大小，因为其边均衡度一般来说是比较小的。

本实施例中，第四预设进度并不是一个具体的数值，而是一种复杂的特殊状态，是第一设备 A 相对于其他设备的落后扩散状态。第四预设进度小于第三预设进度。

下面再详细说明当确定第一设备 A 的边分片进度大于第三预设进度时，如何对调整后的第一扩散速度 V_e 进行调整。为了提高设备的分片速度，可以在确定第一设备 A 的边分片进度大于第三预设进度时，利用第二校正因子 D_e 来减小调整后的第一扩散速度 V_e 。当确定第一设备 A 的边分片进度小于第四预设进度时，利用第二校正因子 D_e 来增大调整后的第一扩散速度 V_e 。

在一种实施方式中，可以通过将调整后的第一扩散速度 V_e 减去 D_e ，来减小调整后的第一扩散速度 V_e ；通过将第一扩散速度 V_e 加上 D_e ，来增大调整后的第一扩散速度 V_e 。在这种实施方式中，第一扩散速度 V_e 的改变依赖于对 D_e 的设置。当 D_e 设置得较小，则边均衡度的收敛速度会相对较慢，当 D_e 设置得较大，则边均衡度的收敛速度会相对较快。其中， D_e 可以是预设的常数，也可以随着迭代过程的进行而调整，例如随着迭代次数的增加逐渐减小。

为了使得对扩散速度的调整速度更加合理，使得数据分片过程的进行更加可控，可以按照第二校正因子 D_e 的对数规律减小调整后的第一扩散速度 V_e 。

在一种实施方式中，第一设备 A 可以按照以下式 (7) 减小调整后的第一扩散速度 V_e ：

$$V_e^3 = V_e^2 * [1 - \log_a(1 + D_e)] \quad (7)$$

第一设备 A 可以按照以下式 (8) 增大调整后的第一扩散速度 V_e ：

$$V_e^3 = V_e^2 * [1 + \log_a(1 + D_e)] \quad (8)$$

其中， V_e^2 是调整后的第一扩散速度， V_e^3 是利用第二校正因子 D_e 对调整后的第一扩散速度继续进行调整得到的扩散速度， $\log_a$ 是以 a 为底的对数， a 可以是预设值，例如可以是自然常数 e 。上述式 (7) 和式 (8) 仅仅是按照第二校正因子 D_e 的对数规律对调整后的第一扩散速度 V_e^2 继续进行调整时的一种实施方式，基于这些公式，很容易得出其他形式的实施方式，例如乘以一个系数或除以一个系数等。

本实施例中，按照第二校正因子 D_e 的对数规律对调整后的第一扩散速度进行调整。对于其中节点的度数量偏大或偏小的设备，其扩散速度的下降幅度和上升幅度都能达到基本相同，并且对 D_e 的设置范围也可以更加宽松，整体的收敛速度也较快。

为了使得对调整后的第一扩散速度 V_e 的调整更加合理, 避免多个设备的数据分片过程出现较大的偏离, 本实施例可以对第二校正因子 D_e 进行自适应调整。例如, 第二校正因子 D_e 可以基于第一设备 A 的已分片边数量 N_A^e 和多个设备的平均已分片边数量 $N_{均}^e$ 的对比确定。

在一种实施方式中, 第一设备 A 可以按照以下式 (9) 减小调整后的第一扩散速度:

$$V_e^3 = V_e^2 * \left\{ 1 - \log_a \left[1 + D_e * \log_b \left(b + \frac{N_A^e - N_{均}^e}{N_{均}^e} \right) \right] \right\} \quad (9)$$

5 第一设备 A 可以按照以下式 (10) 增大调整后的第一扩散速度:

$$V_e^3 = V_e^2 * \left\{ 1 + \log_a \left[1 + D_e * \log_b \left(b + \frac{N_{均}^e - N_A^e}{N_A^e} \right) \right] \right\} \quad (10)$$

其中, V_e^2 是调整后的第一扩散速度, V_e^3 利用第二校正因子 D_e 对调整后的第一扩散速度继续进行调整得到的扩散速度 $\log_b$ 是以 b 为底的对数, b 可以是预设值, 例如可以是自然常数 e , a 和 b 的取值可以相同, 也可以不同, N_A^e 是第一设备 A 的已分片边数量, $N_{均}^e$ 是平均已分片边数量。从式 (9) 和式 (10) 可以看出, 第二校正因子 D_e 被乘以了对数校正项, 10 该对数校正项与第一设备 A 的已分片边数量 N_A^e 和平均已分片节点数量 $N_{均}^e$ 的对比相关。该对数校正项以及校正方式仅仅是一种实施方式, 基于这些式子, 可以容易地得出其他形式的实施方式, 例如直接将第二校正因子 D_e 修改为以下形式也是可行的实施方式:

$$D_e * \log_b \left(b + \gamma * \frac{N_A^e - N_{均}^e}{N_{均}^e} \right) \text{ 或 } D_e * \log_b \left(b + \gamma * \frac{N_{均}^e - N_A^e}{N_A^e} \right)$$

其中, γ 为预设系数。

15 在一种实施方式中, 对第一扩散速度 V_e 的调整还可以与迭代次数 S 结合起来。例如, 可以设置初始的 V_e 为 0.1, 当迭代次数 S 在 $0 \sim S_f$ 之间时, 为精细迭代分片阶段, 对第一扩散速度 V_e 的调整, 可以以初始的 V_e 取 0.1 开始。其中 S_f 例如可以取 500。当迭代次数大于 S_f 小于 S_c 时, 为粗迭代分片阶段, 其中 S_c 可以取大于 500 的值, 例如可以取 1000。对第一扩散速度 V_e 的调整可以包括式 (11):

$$20 \quad V_e^1 = V_e^0 + (1 - V_e^0) * \frac{S - S_f}{S_c - S_f} \quad (11)$$

其中, V_e^0 是调整之前的第一扩散速度, V_e^1 是利用迭代次数调整后的第一扩散速度, S 为迭代次数, 其随着迭代过程逐渐增加。 S_f 和 S_c 是预先设置的值。式 (11) 对第一扩散速度的调整, 可以与基于已分片节点数量的对比对第一扩散速度的调整, 以及基于已分片边数量的对比对第一扩散速度的调整结合起来, 例如, 将式 (11) 与“式 (4) 和式 (5) 中的一个”, 以及“式 (9) 和式 (10) 中的一个”结合起来使用。

25 在上述各个实施例中, 由于扩散速度仅仅是一个数量比例, 当不同的设备的边和节点等数据具有不同的特点时, 利用该扩散速度选择的扩散节点以及得到的目标边的数量也会不同, 从而导致不同设备之间的分片数据中节点数量不均衡, 或者边数量不均衡。上述各个实施例通过调整扩散速度来使得多个设备之间的分片数据的节点数量达到均衡, 边数量 30 达到均衡, 从而使得多个设备中的分片数据达到均衡。当针对多个设备中的分片数据进行查询时, 能够达到负载均衡。

本说明书中, 第一部分边、第一设备、第一扩散速度、第一分片数据、第一预设进度、

第一数量等词语中的“第一”，以及文中相应的“第二”，仅仅是为了区分和描述方便，而不具有任何限定意义。

上述内容对本说明书的特定实施例进行了描述，其他实施例在所附发明内容的范围内。在一些情况下，在发明内容中记载的动作或步骤可以按照不同于实施例中的顺序来执行，并且仍然可以实现期望的结果。另外，在附图中描绘的过程不一定要按照示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中，多任务处理和并行处理也是可以的，或者可能是有利的。

图4为实施例提供的一种对知识图谱进行数据分片的装置的示意性框图。该装置400用于将知识图谱拆分成多个分片数据，多个分片数据分别归属于多个设备。任意一个设备可以通过任何具有计算、处理能力的装置、设备、平台、设备集群等来实现。知识图谱包括代表实体的多个节点，以及体现节点与节点之间关系的边。该装置实施例与图2所示方法实施例相对应。该装置400部署在多个设备中任意的第一设备中，包括：

第一获取模块410，配置为，获取所述知识图谱的第一部分边，第一部分边是对所述知识图谱中的多个边进行初始拆分后得到；第一选择模块420，配置为，基于第一扩散速度，从第一部分边的端节点中选择扩散节点；第二获取模块430，配置为，获取所述知识图谱中以所述扩散节点为一侧端节点的边，作为待分片边；第一分片模块440，配置为，将所述待分片边中的目标边加入第一分片数据中；其中，第一分片数据归属于第一设备，所述第一分片数据中包含已分片边；第三获取模块450，配置为，获取其他设备的分片数据中已分片边包含的端节点，作为其他设备的已分片节点；第一调整模块460，配置为，基于所述第一设备的已分片节点与所述其他设备的已分片节点的对比，对所述第一扩散速度进行调整；第二选择模块470，配置为，基于调整后的第一扩散速度继续选择扩散节点，返回执行第二获取模块430，即获取所述知识图谱中以所述扩散节点为一侧端节点的边。

在一种实施方式中，第一扩散速度的取值在 $(0, 1]$ 之间，用于表示选择的数量比例。

在一种实施方式中，第一选择模块420具体配置为：从所述第一部分边的端节点中选择第一数量个节点作为初始边界点；按照初始边界点关联的边数量从少到多，对多个初始边界点进行排序；基于第一扩散速度，从排序后的多个初始边界点中选择扩散节点。

在一种实施方式中，所述装置400还包括：第一确定模块（图中未示出），配置为采用以下操作确定所述初始边界点关联的边数量：获取其他设备中以所述初始边界点作为一侧端节点的边；其中，所述其他设备包括所述多个设备中除所述第一设备之外的设备，获取的边是所述其他设备从其自身拥有的部分边中确定；针对任意一个初始边界点，基于所述第一部分边中以该初始边界点作为一侧端节点的边数量，以及所述其他设备中以该初始边界点作为一侧端节点的边数量之和，确定该初始边界点关联的边数量。

在一种实施方式中，第二获取模块430具体配置为：从其他设备拥有的部分边中获取以所述扩散节点为一侧端节点的边；将获取的边以及所述第一部分边中以所述扩散节点为一侧端节点的边，确定为所述知识图谱中以所述扩散节点为一侧端节点的边。

在一种实施方式中，装置400还包括：第三选择模块（图中未示出），配置为在将所述待分片边中的目标边加入第一分片数据中之前，基于所述第一扩散速度，从所述待分片边中选择目标边，以从所述待分片边中确定目标边。

在一种实施方式中，装置400还包括：第一接收模块（图中未示出），配置为，接收

其他设备发送的获取请求，所述获取请求用于获取所述第一设备的已分片节点；第一发送模块（图中未示出），配置为，将所述第一设备的已分片节点发送至所述其他设备，以使所述其他设备基于所述第一设备的已分片节点对所述其他设备的扩散速度进行调整。

5 在一种实施方式中，第二选择模块 470 具体配置为：基于调整后的第一扩散速度，从所述目标边的另一侧端节点中选择扩散节点；

或者，第二选择模块 470 具体配置为：基于调整后的第一扩散速度，从所述第一部分边中尚未选择的端节点中选择扩散节点。

10 在一种实施方式中，第一调整模块 460，具体配置为：当基于所述第一设备的已分片节点数量和所述其他设备的已分片节点数量的对比，确定所述第一设备的节点分片进度大于第一预设进度时，利用第一校正因子减小所述第一扩散速度。

在一种实施方式中，第一调整模块 460，利用第一校正因子减小第一扩散速度时，包括，按照第一校正因子的对数规律减小所述第一扩散速度。

15 在一种实施方式中，装置 400 还包括：第二确定模块（图中未示出），配置为，当所述第一设备的已分片节点数量大于所述多个设备的平均已分片节点数量，且所述第一设备的节点均衡度大于预设节点均衡度时，确定所述第一设备的节点分片进度大于所述第一预设进度；其中，平均已分片节点数量和节点均衡度基于所述多个设备的已分片节点数量确定。

在一种实施方式中，第一调整模块 460，还配置为：当基于所述第一设备的已分片节点数量和所述其他设备的已分片节点数量的对比，确定所述第一设备的节点分片进度小于第二预设进度时，利用所述第一校正因子增大所述第一扩散速度。

20 在一种实施方式中，所述装置 400 还包括：第三确定模块（图中未示出），配置为，当所述第一设备的已分片节点数量不大于平均已分片节点数量，且所述多个设备中的最大节点均衡度大于预设节点均衡度时，确定所述第一设备的节点分片进度小于所述第二预设进度。

在一种实施方式中，所述第一校正因子基于所述第一设备的已分片节点数量和所述多个设备的平均已分片节点数量的对比确定。

25 在本说明书的另一实施例中，对图 4 所示实施例加以改进可以得到图 5 所示实施例，图 5 为实施例提供的另一装置实施例的示意性框图。该装置实施例与图 3 所示方法实施例相对应。其中，装置 500 包括：第一获取模块 510、第一选择模块 520、第二获取模块 530、第一分片模块 540、第三获取模块 550、第四获取模块 580、第一调整模块 560 和第二选择模块 570。

30 其中，第一获取模块 510、第一选择模块 520、第二获取模块 530、第一分片模块 540、第三获取模块 550 和第二选择模块 570 分别与图 4 实施例中的第一获取模块 410、第一选择模块 420、第二获取模块 430、第一分片模块 440、第三获取模块 450 和第二选择模块 470 完全相同，本实施例不再对这部分模块进行赘述。下面着重说明与图 4 实施例的不同之处。

35 第四获取模块 580，配置为，在对所述第一扩散速度进行调整之前，获取其他设备的分片数据中的已分片边；第一调整模块 560，配置为，基于所述第一设备的已分片节点和所述其他设备的已分片节点的对比，以及所述第一设备的已分片边和所述其他设备的已分片边的对比，对所述第一扩散速度进行调整。

在一种实施方式中，第一调整模块 560 可以包括：第一调整子模块 561，配置为，基

于所述第一设备的已分片节点和所述其他设备的已分片节点的对比，对所述第一扩散速度进行初步调整；第二调整子模块 562，配置为，基于所述第一设备的已分片边和所述其他设备的已分片边的对比，继续对调整后的第一扩散速度进行调整。

5 在一种实施方式中，第二调整子模块 562 具体配置为：当基于所述第一设备的已分片边数量和所述其他设备的已分片边数量的对比，确定所述第一设备的边分片进度大于第三预设进度时，利用第二校正因子减小调整后的第一扩散速度。

在一种实施方式中，第二调整子模块 562 具体配置为：按照第二校正因子的对数规律减小调整后的第一扩散速度。

10 在一种实施方式中，第一调整模块 560 还包括：第一确定子模块（图中未示出），配置为，当所述第一设备的已分片边数量大于所述多个设备的平均已分片边数量，且所述第一设备的边均衡度大于预设边均衡度时，确定所述第一设备的边分片进度大于第三预设进度；其中，所述平均已分片边数量和所述边均衡度基于所述多个设备的已分片边数量确定。

15 在一种实施方式中，第二调整子模块 562 还配置为：当基于所述第一设备的已分片边数量和所述其他设备的已分片边数量的对比，确定所述第一设备的边分片进度小于第四预设进度时，利用所述第二校正因子增大调整后的第一扩散速度。

在一种实施方式中，第一调整模块 560 还包括：第二确定子模块（图中未示出），配置为，当所述第一设备的已分片边数量不大于平均已分片边数量，且所述多个设备的最大边均衡度大于预设边均衡度时，确定所述第一设备的边分片进度小于第四预设进度。

20 在一种实施方式中，所述第二校正因子基于所述第一设备的已分片边数量和所述多个设备的平均已分片边数量的对比确定。

上述多个装置实施例与方法实施例相对应，具体说明可以参见方法实施例部分的描述，此处不再赘述。装置实施例是基于对应的方法实施例得到，与对应的方法实施例具有同样的技术效果，具体说明可参见对应的方法实施例。

25 本说明书实施例还提供了一种计算机可读存储介质，其上存储有计算机程序，当所述计算机程序在计算机中执行时，令计算机执行图 1 至图 3 任一项所述的方法。

本说明书实施例还提供了一种计算设备，包括存储器和处理器，所述存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现图 1 至图 3 任一项所述的方法。

30 本说明书中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于存储介质和计算设备实施例而言，由于其基本相似于方法实施例，所以描述得比较简单，相关之处参见方法实施例的部分说明即可。

35 本领域技术人员应该可以意识到，在上述一个或多个示例中，本发明实施例所描述的功能可以用硬件、软件、固件或它们的任意组合来实现。当使用软件实现时，可以将这些功能存储在计算机可读介质中或者作为计算机可读介质上的一个或多个指令或代码进行传输。

以上所述的具体实施方式，对本发明实施例的目的、技术方案和有益效果进行了进一步的详细说明。所应理解的是，以上所述仅为本发明实施例的具体实施方式而已，并不用于限定本发明的保护范围，凡在本发明的技术方案的基础之上所做的任何修改、等同替换、改进等，均应包括在本发明的保护范围之内。

权 利 要 求 书

1. 一种对知识图谱进行数据分片的方法，用于将知识图谱拆分成多个分片数据，多个分片数据分别归属于多个设备，所述知识图谱包括代表实体的多个节点，以及体现节点与节点之间关系的边；所述方法通过所述多个设备中任意的第一设备执行，包括：

获取所述知识图谱的第一部分边，所述第一部分边是对所述知识图谱中的多个边进行初始拆分后得到；

基于第一扩散速度，从所述第一部分边的端节点中选择扩散节点；

获取所述知识图谱中以所述扩散节点为一侧端节点的边，作为待分片边；

将所述待分片边中的目标边加入第一分片数据中；其中，所述第一分片数据归属于所述第一设备，所述第一分片数据中包含已分片边；

获取其他设备的分片数据中已分片边包含的端节点，作为其他设备的已分片节点；

基于所述第一设备的已分片节点与所述其他设备的已分片节点的对比，对所述第一扩散速度进行调整；

基于调整后的第一扩散速度继续选择扩散节点，返回执行所述获取所述知识图谱中以所述扩散节点为一侧端节点的边的步骤。

2. 根据权利要求1所述的方法，所述第一扩散速度的取值在(0,1]之间，用于表示选择的数量比例。

3. 根据权利要求1所述的方法，所述基于第一扩散速度，从所述第一部分边的端节点中选择扩散节点的步骤，包括：

从所述第一部分边的端节点中选择第一数量个节点作为初始边界点；

按照初始边界点关联的边数量从少到多，对多个初始边界点进行排序；

基于第一扩散速度，从排序后的多个初始边界点中选择扩散节点。

4. 根据权利要求3所述的方法，采用以下方法确定所述初始边界点关联的边数量：

获取其他设备中以所述初始边界点作为一侧端节点的边；其中，所述其他设备包括所述多个设备中除所述第一设备之外的设备，获取的边是所述其他设备从其自身拥有的部分边中确定；

针对任意一个初始边界点，基于所述第一部分边中以该初始边界点作为一侧端节点的边数量，以及所述其他设备中以该初始边界点作为一侧端节点的边数量之和，确定该初始边界点关联的边数量。

5. 根据权利要求1所述的方法，所述获取所述知识图谱中以所述扩散节点为一侧端节点的边的步骤，包括：

从其他设备拥有的部分边中获取以所述扩散节点为一侧端节点的边；

将获取的边以及所述第一部分边中以所述扩散节点为一侧端节点的边，确定为所述知识图谱中以所述扩散节点为一侧端节点的边。

6. 根据权利要求1所述的方法，采用以下方式从所述待分片边中确定目标边：

基于所述第一扩散速度，从所述待分片边中选择目标边。

7. 根据权利要求1所述的方法，还包括：

接收其他设备发送的获取请求，所述获取请求用于获取所述第一设备的已分片节

点；

将所述第一设备的已分片节点发送至所述其他设备，以使所述其他设备基于所述第一设备的已分片节点对所述其他设备的扩散速度进行调整。

8. 根据权利要求1所述的方法，所述基于调整后的第一扩散速度继续选择扩散节点的步骤，包括：

基于调整后的第一扩散速度，从所述目标边的另一侧端节点中选择扩散节点；

或者，基于调整后的第一扩散速度，从所述第一部分边中尚未选择的端节点中选择扩散节点。

9. 根据权利要求1所述的方法，所述基于所述第一设备的已分片节点与所述其他设备的已分片节点的对比，对所述第一扩散速度进行调整的步骤，包括：

当基于所述第一设备的已分片节点数量和所述其他设备的已分片节点数量的对比，确定所述第一设备的节点分片进度大于第一预设进度时，利用第一校正因子减小所述第一扩散速度。

10. 根据权利要求9所述的方法，采用以下方式确定所述第一设备的节点分片进度大于第一预设进度：

当所述第一设备的已分片节点数量大于所述多个设备的平均已分片节点数量，且所述第一设备的节点均衡度大于预设节点均衡度时，确定所述第一设备的节点分片进度大于所述第一预设进度；其中，所述平均已分片节点数量和所述节点均衡度基于所述多个设备的已分片节点数量确定。

11. 根据权利要求9所述的方法，还包括：

当基于所述第一设备的已分片节点数量和所述其他设备的已分片节点数量的对比，确定所述第一设备的节点分片进度小于第二预设进度时，利用所述第一校正因子增大所述第一扩散速度。

12. 根据权利要求11所述的方法，采用以下方式确定所述第一设备的节点分片进度小于第二预设进度：

当所述第一设备的已分片节点数量不大于平均已分片节点数量，且所述多个设备中的最大节点均衡度大于预设节点均衡度时，确定所述第一设备的节点分片进度小于所述第二预设进度。

13. 根据权利要求9所述的方法，所述利用第一校正因子减小所述第一扩散速度的步骤，包括：

按照第一校正因子的对数规律减小所述第一扩散速度。

14. 根据权利要求9所述的方法，所述第一校正因子基于所述第一设备的已分片节点数量和所述多个设备的平均已分片节点数量的对比确定。

15. 根据权利要求1所述的方法，在对所述第一扩散速度进行调整之前，还包括：

获取其他设备的分片数据中的已分片边；

所述对所述第一扩散速度进行调整的步骤，包括：

基于所述第一设备的已分片节点和所述其他设备的已分片节点的对比，以及所述第一设备的已分片边和所述其他设备的已分片边的对比，对所述第一扩散速度进行调整。

16. 根据权利要求 15 所述的方法, 所述对所述第一扩散速度进行调整的步骤, 包括:

基于所述第一设备的已分片节点和所述其他设备的已分片节点的对比, 对所述第一扩散速度进行初步调整;

基于所述第一设备的已分片边和所述其他设备的已分片边的对比, 继续对调整后的第一扩散速度进行调整。

17. 根据权利要求 16 所述的方法, 所述基于所述第一设备的已分片边和所述其他设备的已分片边的对比, 继续对调整后的第一扩散速度进行调整的步骤, 包括:

当基于所述第一设备的已分片边数量和所述其他设备的已分片边数量的对比, 确定所述第一设备的边分片进度大于第三预设进度时, 利用第二校正因子减小调整后的第一扩散速度。

18. 根据权利要求 17 所述的方法, 采用以下方式确定所述第一设备的边分片进度大于第三预设进度:

当所述第一设备的已分片边数量大于所述多个设备的平均已分片边数量, 且所述第一设备的边均衡度大于预设边均衡度时, 确定所述第一设备的边分片进度大于第三预设进度; 其中, 所述平均已分片边数量和所述边均衡度基于所述多个设备的已分片边数量确定。

19. 根据权利要求 17 所述的方法, 还包括:

当基于所述第一设备的已分片边数量和所述其他设备的已分片边数量的对比, 确定所述第一设备的边分片进度小于第四预设进度时, 利用所述第二校正因子增大调整后的第一扩散速度。

20. 根据权利要求 19 所述的方法, 采用以下方式确定所述第一设备的边分片进度小于第四预设进度:

当所述第一设备的已分片边数量不大于平均已分片边数量, 且所述多个设备的最大边均衡度大于预设边均衡度时, 确定所述第一设备的边分片进度小于第四预设进度。

21. 根据权利要求 17 所述的方法, 所述利用第二校正因子减小调整后的第一扩散速度的步骤, 包括:

按照第二校正因子的对数规律减小调整后的第一扩散速度。

22. 根据权利要求 17 所述的方法, 所述第二校正因子基于所述第一设备的已分片边数量和所述多个设备的平均已分片边数量的对比确定。

23. 一种对知识图谱进行数据分片的装置, 用于将知识图谱拆分成多个分片数据, 多个分片数据分别归属于多个设备, 所述知识图谱包括代表实体的多个节点, 以及体现节点与节点之间关系的边; 所述装置部署在所述多个设备中任意的第一设备中, 包括:

第一获取模块, 配置为, 获取所述知识图谱的第一部分边, 所述第一部分边是对所述知识图谱中的多个边进行初始拆分后得到;

第一选择模块, 配置为, 基于第一扩散速度, 从所述第一部分边的端节点中选择扩散节点;

第二获取模块, 配置为, 获取所述知识图谱中以所述扩散节点为一侧端节点的边,

作为待分片边；

第一分片模块，配置为，将所述待分片边中的目标边加入第一分片数据中；其中，所述第一分片数据归属于所述第一设备，所述第一分片数据中包含已分片边；

第三获取模块，配置为，获取其他设备的分片数据中已分片边包含的端节点，作为其他设备的已分片节点；

第一调整模块，配置为，基于所述第一设备的已分片节点与所述其他设备的已分片节点的对比，对所述第一扩散速度进行调整；

第二选择模块，配置为，基于调整后的第一扩散速度继续选择扩散节点，返回执行所述第二获取模块。

24. 一种计算机可读存储介质，其上存储有计算机程序，当所述计算机程序在计算机中执行时，令计算机执行权利要求 1-22 中任一项所述的方法。

25. 一种计算设备，包括存储器和处理器，所述存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现权利要求 1-22 中任一项所述的方法。