(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 113196277 A
(43)申请公布日 2021.07.30

(21)申请号 201980082810.7

(22)申请日 2019.10.13

(30)优先权数据
20185863 2018.10.13 FI

(85)PCT国际申请进入国家阶段日
2021.06.11

(86)PCT国际申请的申请数据
PCT/FI2019/050731 2019.10.13

(87)PCT国际申请的公布数据
WO2020/074786 EN 2020.04.16

(71)申请人 伊普拉利技术有限公司
地址 芬兰赫尔辛基

(72)发明人 S·阿维拉 J·卡利奥
S·比约克维斯特

(74)专利代理机构 北京汇知杰知识产权代理有限公司 11587
代理人 李洁 董江虹

(51)Int.Cl.
G06F 40/205(2006.01)
G06F 40/279(2006.01)
G06N 20/00(2006.01)
G06N 3/08(2006.01)

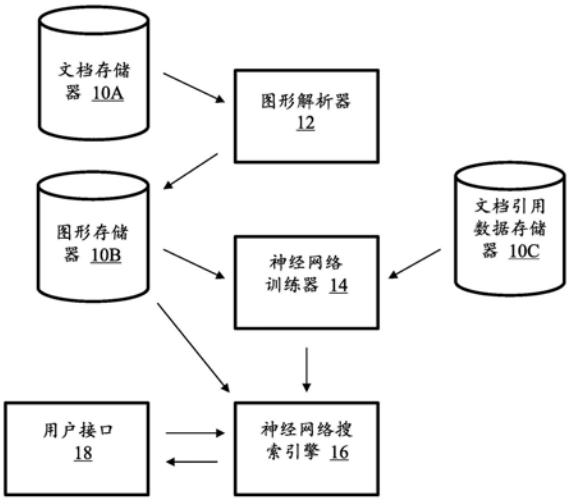
权利要求书2页 说明书13页 附图8页

(54)发明名称

用于检索自然语言文档的系统

(57)摘要

本发明提供了一种自然语言检索系统和方法。所述系统包括数字数据存储装置，用于存储多个自然语言块和对应于所述块的数据图形。第一数据处理装置适于将所述块转换为所述图形，所述图形被存储在所述存储装置中。所述图形包含多个节点，每个节点包含从所述块提取的自然语言单元作为节点值。还提供了：第二数据处理装置，用于执行机器学习算法，所述机器学习算法能够遍历所述图形并且读取所述节点值，以用于基于所述图形的节点结构和所述图形的节点值形成经训练的机器学习模型；以及第三数据处理装置，适于读取新鲜的图形并且利用所述模型以用于基于所述新鲜的图形确定所述自然语言块的子集。



CN 113196277 A

1.一种自然语言检索系统,包括

-数字数据存储装置(10A,10B),用于存储

o多个自然语言块,

o对应于所述块的数据图形,

-第一数据处理装置(12),适于将所述块转换为所述图形,所述图形被存储在所述存储装置中,由此所述图形包含多个节点,每个节点包含从所述块提取的自然语言单元作为节点值,

其特征在于,所述系统还包括

-第二数据处理装置(14),用于执行机器学习算法,所述机器学习算法能够遍历所述图形以用于基于所述图形的节点结构和所述图形的节点值形成经训练的机器学习模型,

-第三数据处理装置(16),适于读取新鲜的图形或被转换为新鲜的图形的新鲜的自然语言块,并且利用所述机器学习模型以用于基于所述新鲜的图形确定所述自然语言块的子集。

2.根据权利要求1所述的系统,其中至少一些图形中的包含特定自然语言单元值的至少一些节点的数目被配置为小于所述特定自然语言单元值在对应的自然语言块中出现的数目。

3.根据权利要求1或2所述的系统,其中所述第一数据处理装置(12)适于通过以下方式将所述块转换为所述图形:

-从所述块识别第一组自然语言符号以及与所述第一组自然语言符号不同的第二组自然语言符号,

-利用所述第一组符号和所述第二组符号执行匹配器,以用于形成第一组符号的匹配对,

-利用所述匹配对将所述第一组符号的至少一部分布置为所述图形的连续节点。

4.根据前述权利要求中任一项所述的系统,其中所述第一数据处理ns(12)适于形成包含多个边缘的图形,所述图形的各个节点包含相互之间具有部分词关系的自然语言单元,如从所述块导出的。

5.根据前述权利要求中任一项的任一项所述的系统,其中所述第一数据处理装置(12)适于形成包含多个边缘的图形,所述图形的各个节点包含相互之间具有下位词关系的自然语言单元,如由所述块导出的。

6.根据前述权利要求中任一项所述的系统,其中所述第一数据处理装置(12)适于形成包含多个边缘的图形,所述图形的至少一个节点能够包含对同一图形中的一个或多个节点的引用以及附加地从相应的自然语言块导出的至少一个自然语言单元。

7.根据前述权利要求中任一项所述的系统,其中所述图形是树形图形,所述树形图形的节点值包含词或多词组块,所述词或多词组块是通过所述第一处理装置使用词的词性和句法依赖性从所述自然语言块导出的或从其向量化形式导出的。

8.根据前述权利要求中任一项所述的系统,其中所述第一数据处理装置(12)适于使用概率图形模型(PGM)以用于确定所述图形的边缘概率,以及使用所述边缘概率来形成所述图形。

9.根据前述权利要求中任一项所述的系统,其中所述第二数据处理装置(14)适于执行

基于图形的神经网络算法,诸如循环神经网络(RNN)图形算法,特别是长短期记忆(LSTM)算法,诸如Tree-LSTM算法。

10.根据前述权利要求中任一项所述的系统,其中所述经训练的机器学习模型适于将图形映射成多维向量,所述多维向量的相对角度由所述图形的节点结构和所述图形的节点值限定。

11.根据前述权利要求中任一项所述的系统,其中所述机器学习模型适于根据图形的节点结构和图形的节点值将所述图形或图形对分类为两个或更多个分类。

12.根据前述权利要求中任一项所述的系统,其中

-所述存储装置还被配置为存储将所述块中的至少一些相互链接的引用数据,并且

-所述机器学习算法具有依赖于用于训练所述机器学习模型的所述引用数据的学习目标。

13.根据前述权利要求中任一项所述的系统,其中所述存储装置被配置为存储自然语言文档,每个所述自然语言文档包含第一自然语言块和第二自然语言块。

14.根据权利要求12和13所述的系统,其中所述第二数据处理装置(14)在所述训练中被配置为使用对应于第一文档的第一块的多个第一图形,并且对于每个第一图形,使用至少部分地基于与所述第一文档不同的第二文档的第二块的一个或多个第二图形,如由所述引用数据所定义的。

15.根据权利要求12-14中任一项所述的系统,其中所述第二数据处理装置(14)在所述训练中被配置为使用对应于第一文档的第一块的多个第一图形,并且对于每个第一图形,使用至少部分地基于所述第一文档的第二块的第二图形。

16.根据前述权利要求中任一项所述的系统,其中所述第三数据处理装置(16)适于读取以新鲜的图形的形式或以被转换为对应的图形的新鲜的自然语言块的形式输入的所述新鲜的自然语言。

17.根据前述权利要求中任一项所述的系统,是利用权利要求和全说明书作为所述自然语言块的专利检索系统。

18.一种计算机实施的检索自然语言文档的方法,所述方法包括

-将多个自然语言块存储到数字数据存储器中,

-将所述块转换为对应的图形,所述图形包含多个节点,每个节点包含从所述块提取的自然语言单元作为节点值,

-将所述图形存储在所述数字数据存储器中,

其特征在于,所述方法还包括

-执行机器学习算法,所述机器学习算法能够遍历所述图形,以用于基于所述图形的节点结构和所述图形的节点值形成经训练的机器学习模型,

-读取新鲜的图形或被转换为新鲜的图形的新鲜的自然语言块,以及

-利用所述机器学习模型以用于基于所述新鲜的图形确定所述自然语言块的子集。

用于检索自然语言文档的系统

技术领域

[0001] 本发明涉及自然语言处理。特别地，本发明涉及基于机器学习的——诸如基于神经网络的——用于检索、比较或分析包含自然语言的文档(document)的系统和方法。所述文档可以是技术文档或科学文档。特别地，所述文档可以是专利文档。

背景技术

[0002] 在商业、工业、经济和文化的许多领域都需要书面技术概念的比较。一个具体示例是对专利申请的审查，其中一个目的是确定在专利申请的权利要求中限定的技术概念是否在语义上涵盖在另一个文档中限定的另一个技术概念。

[0003] 当前，存在越来越多的可用于查找单独文档的检索工具，但是对由文档公开的概念的分析和比较仍然主要是手工工作，涉及对词、句子和更大的语言实体的含义的人类推断。

[0004] 围绕自然语言处理的科学研究已经产生用于通过计算机自动解析语言的工具。这些工具可以被使用，例如，以符号化(tokenize)文本、词性标注(part-of-speech tagging)、实体识别以及识别词或实体之间的相关性。

[0005] 也已经进行科学工作以通过从文档提取关键概念例如出于文本概要和技术趋势分析目的自动分析专利。

[0006] 最近，使用多维词向量(word vector)的词嵌入(word embedding)已经成为用于将词的含义映射成数字计算机可处理的形式的重要工具。此方法可以由神经网络——诸如循环神经网络(recurrent neural network)——使用，用于为计算机提供对文档的内容的更深入理解。

[0007] 传统上，使用关键字检索进行专利检索，关键字检索涉及限定正确的关键字以及其同义词、词形变化形式等，以及布尔检索策略的创建。这是耗时的并且需要专门知识。最近，语义检索也已经得到发展，语义检索是模糊的并且可能涉及人工智能技术的使用。它们有助于快速查找到以某种方式与在另一个文档中讨论的概念相关的大量文档。然而，它们在例如专利新颖性检索方面是相对有限的，因为在实践中它们的评价新颖性——即查找公开了落在专利权利要求中限定的一般概念下的具体内容的文档——的能力是有限的。

[0008] 总之，存在很好地适合于一般检索以及例如从文本和文本概要提取核心概念的可用技术。然而，它们并不很好地适合在大量数据中进行在不同文档中公开的概念之间的详细比较，所述详细比较例如对于专利新颖性检索目的或其他技术比较目的来说是至关重要的。

[0009] 特别是为了实现更有效率的检索和新颖性评价工具，需要改进的用于文本分析和比较的技术。

发明内容

[0010] 本发明的一个目的是解决上述问题中的至少一些以及提供一种提高技术检索的

准确度的新颖的系统和方法。一个具体的目的是提供一种能够更好地考虑文档的概念之间的技术关系以进行针对性检索的解决方案。

[0011] 一个具体目的是提供一种用于改进的专利检索和自动新颖性评价的系统和方法。

[0012] 根据一个方面，本发明提供一种自然语言检索系统，所述自然语言检索系统包括数字数据存储装置，用于存储多个自然语言块(block)和对应于所述块的数据图形(graph)。还提供了适于将所述块转换为所述图形的第一数据处理装置，所述图形被存储在所述存储装置中。所述图形包含多个节点，优选地连续节点，每个节点包含从所述块提取的自然语言单元作为节点值或其一部分。还提供了：第二数据处理装置，用于执行机器学习算法，所述机器学习算法能够遍历所述图形并且读取所述节点值，以用于基于所述图形的节点结构和所述图形的节点值形成经训练的机器学习模型；以及第三数据处理装置，适于读取新鲜的(fresh)图形或被转换为新鲜的图形的新鲜的自然语言块，并且利用所述机器学习模型以用于基于所述新鲜的图形确定所述自然语言块的子集。

[0013] 本发明还涉及一种适于读取自然语言块并且实施所述第一数据处理装置、第二数据处理装置和第三数据处理装置的功能的方法。

[0014] 根据一个方面，本发明提供一种检索专利文档的系统和方法，所述方法包括读取多个专利文档，每个专利文档包括全说明书(specification)和权利要求，以及分别将所述全说明书和权利要求转换为全说明书图形和权利要求图形。所述图形包含多个节点，每个节点具有从所述全说明书或权利要求提取的第一自然语言单元作为节点值，并且包含所述节点之间的多个边缘(edge)，所述边缘是基于从所述全说明书或权利要求提取的至少一个第二自然语言单元确定的。所述方法包括使用机器学习算法来训练机器学习模型，所述机器学习算法能够根据所述边缘遍历所述图形，并且利用所述节点值以用于使用所述全说明书图形和权利要求图形的多个不同的对作为训练数据形成经训练的机器学习模型。所述方法还包括读取新鲜的图形或被转换为新鲜的图形的文本块，并且利用所述经训练的机器学习模型以用于基于所述新鲜的图形确定所述专利文档的子集。

[0015] 所述图形可以特别是在至少一些连续节点的节点值之间具有部分词关系的树形递归(recursive)图形。

[0016] 所述方法和系统优选地是基于神经网络的，由此所述机器学习模型是神经网络模型。

[0017] 更具体地，本发明的特征在于独立权利要求中所陈述的内容。

[0018] 本发明提供了显著的益处。与基于关键字的检索相比，本基于图形的并且利用神经网络的方法具有的优点是：检索不仅基于词的文本内容以及可选地其他传统标准例如词的亲密度，而且还考虑了文档中的概念的实际技术关系。这使本方法特别适合于例如专利检索，在专利检索中技术内容——而不是确切表达或书写文档的风格——很重要。因此，可以实施更准确的技术检索。

[0019] 与利用例如基于文本的线性神经网络模型的所谓的语义检索相比，基于图形的方法能够更好地考虑文档的实际技术内容。此外，与全文本相比，轻量级图形需要少得多的计算能力来走查(walk through)。这允许使用多得多的训练数据、缩短开发和学习周期，从而导致更准确的检索。实际的检索持续时间也可以被缩短。

[0020] 本方法与使用现实训练数据——诸如由专利当局和专利申请人提供的专利新颖

性检索数据和引文数据——兼容。本方法还允许高级的训练方案，诸如数据扩增(augmentation)，如稍后将详细讨论的。

[0021] 通过现实测试数据已经示出，专利文本的精简和简化图形表示，与现实训练数据相结合，产生相对高的检索准确度和高计算训练效率。

[0022] 从属权利要求涉及本发明的所选择的实施方案。

[0023] 接下来，参考附图更详细地讨论本发明的所选择的实施方案以及其优点。

附图说明

[0024] 图1A示出了在一般水平上一种示例性检索系统的块图。

[0025] 图1B示出了所述检索系统一个更详细的实施方案的块图，所述检索系统包括一系列基于神经网络的搜索引擎以及其训练器。

[0026] 图1C示出了根据一个实施方案的一种专利检索系统的块图。

[0027] 图2A示出了仅具有部分词关系/整体词关系的示例性嵌套图形的块图。

[0028] 图2B示出了具有部分词/整体词关系和下位词/上位词关系的示例性嵌套图形的块图。

[0029] 图3示出了一种示例性图形解析算法的流程图。

[0030] 图4A示出了使用专利检索/引文数据作为训练数据的专利检索神经网络训练的块图。

[0031] 图4B示出了使用来源于同一专利文档的权利要求-说明书(description)图形对作为训练数据的神经网络训练的块图。

[0032] 图4C示出了使用扩增的权利要求图形集合作为训练数据的神经网络训练的块图。

[0033] 图5例示了根据一个实施方案的示例性图形馈送用户接口(user interface)的功能。

具体实施方式

[0034] 定义

[0035] 在本文中的“自然语言单元”是指文本组块(chunk)，或在嵌入之后，文本组块的向量表示。所述组块可以是在以计算机可读形式存储的原始文本中出现一次或多次的单个词或多词子概念。所述自然语言单元可以被呈现为一组字符值(在计算机科学中通常称为“字符串”)，或数字地被呈现为多维向量值，或对这样的值的引用。

[0036] “自然语言块”是指包含自然语言单元的在语言上有意义的组合的数据实例(data instance)，例如一种语言(诸如英语)的一个或多个完整的或不完整的句子。所述自然语言块可以被表示为例如单个字符串并且被存储在文件(file)系统中的文件中和/或经由用户接口显示给用户。

[0037] “文档”是指包含自然语言内容并且与机器可读文档标识符相关联的机器可读实体，所述机器可读文档标识符相对于系统内的其他文档是唯一的。

[0038] “专利文档”是指专利申请或授权专利的自然语言内容。专利文档在本系统中与由诸如EPO、WIPO或USPTO或另一个国家或地区的另一个国家或地区专利局的公认专利当局分配的公开号和/或另一个机器可读唯一文档标识符相关联。术语“权利要求”是指专利文档

的权利要求——特别是独立权利要求——的基本内容。术语“全说明书”是指涵盖专利文档的说明书的至少一部分的专利文档内容。全说明书还可以涵盖专利文档的其他部分，诸如摘要或权利要求书。权利要求书和全说明书是自然语言块的示例。

[0039] “权利要求”在本文中被定义为在本专利申请的生效日会被欧洲专利局视为权利要求的自然语言块。特别地，“权利要求”是自然语言文档的用其中的机器可读整数数字标识的计算机可识别的块，所述机器可读整数数字例如在该块前面为字符串格式和/或作为标记文件格式(诸如xml或html格式)的相关信息(的一部分)。

[0040] “全说明书”在本文中被定义为计算机可识别的自然语言块，在还包含至少一个权利要求的专利文档内计算机可识别的，并且包含该文档的除权利要求之外的至少一个其他部分。另外，“全说明书”可以是通过标记文档格式(诸如xml或html格式)的相关信息可识别的。

[0041] 在本文中的“边缘关系”可以特别是从块提取的技术关系和/或从使用有关的自然语言单元的语义学导出的语义关系。特别地，所述边缘关系可以是

[0042] -部分词关系(也是:部分词/整体词关系);部分词:X是Y的一部分;整体词:Y具有X作为其自己的一部分;例如:“车轮”是“汽车”的部分词，

[0043] -下位词关系(也是:下位词/上位词关系);下位词:X是Y的下位;上位词:X是Y的上位;例如:“电动汽车”是“汽车”的下位词，或

[0044] -同义词关系:X与Y相同。

[0045] 在一些实施方案中，边缘关系被定义在递归图形的连续嵌套的节点之间，每个节点包含一个自然语言单元作为节点值。

[0046] 另一些可能的技术关系包括主题(thematic)关系，是指除上述关系之外，文本的一个子概念相对于一个或多个其他子概念所起的作用。至少一些主题关系可以被定义在连续嵌套的单元之间。在一个示例中，父单元的主题关系被定义在子单元中。主题关系的一个示例是作用分类(role class)“功能”。例如，“把手”的功能可以是“允许对对象的操纵”。这样的主题关系可以被存储作为“把手”单元的子单元，该“功能”作用与该子单元相关联。主题关系也可以是不具有预定义分类(或具有诸如“关系”的一般分类)的通用关系，但是用户可以自由定义该关系。例如，把手和杯子之间的通用关系可以是“[把手]用胶粘剂附接到[杯子]”。这样的主题关系可以被存储作为“把手”单元或“杯子”单元或二者的子单元，优选地彼此相互引用。

[0047] 如果关系单元链接到当通过数据处理器运行时产生包括特定的关系分类或子类中的关系的自然语言块的计算机可执行代码，该关系单元被认为定义该分类或子类中的关系。

[0048] “图形”或“数据图形”是指遵循一般非线性递归和/或网络数据模式的数据实例。本系统能够同时包含遵循相同数据模式并且其数据来源于不同来源和/或与不同来源相关的几个不同图形。实际上，图形可以任何合适的文本或二进制格式存储，这允许递归地存储数据项和/或存储数据项作为网络。图形特别是语义图形和/或技术图形(描述节点值之间的语义关系和/或技术关系)，而不是句法图形(其仅描述节点值之间的语言关系)。图形可以是树形图形。包括多个树的森林形图形在本文中被认为树形图形。特别地，所述图形可以是技术树形图形。

[0049] “数据模式”是指组织数据——特别是自然语言单元和与其相关联的数据，诸如所述单元之间的技术关系的信息——所根据的规则。

[0050] 自然语言单元的“嵌套”是指所述单元具有一个或多个子和一个或多个父的能力，如由数据模式确定的。在一个示例中，所述单元可以具有一个或多个子并且仅具有单个父。根单元不具有父，并且叶单元不具有子。同级单元具有相同的父。“连续嵌套”是指在父单元与其直接子单元之间的嵌套。

[0051] “递归”嵌套或数据模式是指允许嵌套包含数据项的自然语言单元的嵌套或数据模式。

[0052] “(自然语言)符号”是指较大的自然语言块中的一个词或词组块。符号还可以包含与词或词组块相关的元数据，诸如词性(POS)标签(label)或句法依赖性标注。一“组”自然语言符号特别是指可以根据预定规则或模糊逻辑基于其文本值、POS标签或依赖性标注或它们的任何组合来分组的符号。

[0053] 术语“数据存储装置”、“处理装置”和“用户接口装置”主要是指软件装置，即计算机可执行代码(指令)，所述计算机可执行代码(指令)可以被存储在非暂时性计算机可读介质上并且适于在由处理器执行时分别执行指定的功能，换言之，存储数字数据、允许用户与所述数据交互、以及处理所述数据。系统的所有这些部件可以被承载于一个软件中，所述软件由本地计算机运行或通过本地安装的网络浏览器由网络服务器运行，例如，由用于运行软件部件的合适的硬件支持。本文描述的方法是计算机实施的方法。

[0054] 所选择的实施方案的描述

[0055] 下文描述了一种自然语言检索系统，所述自然语言检索系统包括用于存储多个自然语言块和对应于所述块的数据图形的数字数据存储装置。所述存储装置可以包括一个或多个本地或云数据存储器。所述存储器可以是基于文件的或基于查询语言的。

[0056] 第一数据处理装置是适于将所述块转换为所述图形的转换器单元。每个图形包含多个节点，每个节点包含从所述块提取的自然语言单元作为节点值。边缘被限定在成对的节点之间，定义节点之间的技术关系。例如，所述边缘或它们中的一些可以定义两个节点之间的部分词关系。

[0057] 在一些实施方案中，所述图形中的包含特定自然语言单元值的至少一些节点的数目小于所述特定自然语言单元在对应的自然语言块中出现的数目。换言之，所述图形是原始文本的精简表示，该精简表示例如可使用稍后描述的符号识别和匹配方法来实现。通过允许每个节点有多个子节点，文本的基本技术内容(并且可选地语义内容)仍然可以保持在图形表示中。精简图形通过基于图形的神经网络算法处理也是高效的，由此与从直接文本表示学习相比它们能够更好并且更快地学习文本的基本内容。已经证明此方法在技术文本的比较方面并且特别是在基于权利要求检索专利全说明书以及权利要求的新颖性的自动评价方面特别强大。

[0058] 在一些实施方案中，包含特定自然语言单元的所有节点的数目是一。换言之，不存在重复节点。虽然这可以导致文本的原始内容的简化，但是至少在使用树形图形时，它导致可非常有效率地处理的并且仍然相对有表现力的适合于专利检索和新颖性评价的图形。

[0059] 在一些实施方案中，所述图形是至少用于在所述原始文本中发现的名词和名词组块的这样的精简图形。特别地，所述图形可以是根据其部分词关系布置的名词值(noun-

valued)节点的精简图形。在平均专利文档中,许多名词术语在整个文本中出现数十次或甚至数百次。通过本方案,这样的文档的内容可以被压缩到原始空间的一小部分,同时使它们对于机器学习而言更可行。

[0060] 在一些实施方案中,在至少一个原始自然语言块中出现多次的多个术语在对应的图形中恰好出现一次。

[0061] 精简图形表示也是有益的,因为在构建所述图形时同义词和共指关系(coreference)(在特定上下文中意指同一事物的表达)可以被考虑。这导致甚至更精简的图形。在一些实施方案中,以至少两种不同书面形式出现在至少一个原始自然语言块中的多个术语在对应的图形中恰好出现一次。

[0062] 第二数据处理装置是用于执行神经网络算法的神经网络训练器,所述神经网络算法能够迭代地遍历图形结构并且既能够从图形的内部结构又能够从其节点值学习,如由定义学习目标的损失函数以及训练数据案例(data case)定义的。所述训练器通常接收图形或从其导出的扩增图形的训练数据组合,如由所述训练算法指定的。所述训练器输出经训练的神经网络模型。

[0063] 已经发现如本文所描述的这种采用图形形式数据的监督机器学习方法在专利文档和科学文档之中查找技术上相关的文档方面异常强大。

[0064] 在一些实施方案中,所述存储装置还被配置为存储将所述块中的至少一些相互链接的引用数据。所述引用数据由所述训练器使用以导出训练数据,即以定义在训练中用作肯定训练案例或否定训练案例(即训练样本)的图形的组合。所述训练器的学习目标取决于此信息。

[0065] 第三数据处理装置是搜索引擎,所述搜索引擎适于通常通过用户接口或网络界面读取新鲜的图形或新鲜的自然语言块。如果需要,所述块在所述转换器单元中被转换为图形。所述搜索引擎使用所述经训练的神经网络模型,以用于基于所述新鲜的图形来确定自然语言块(或从其导出的图形)的子集。

[0066] 图1A示出了本系统的一个实施方案,所述系统特别适合于检索技术文档,诸如专利文档或科学文档。所述系统包括文档存储器10A,所述文档存储器包含多个自然语言文档。图形解析器12适于从文档存储器10A读取文档并且将它们转换为图形格式,这稍后被更详细地讨论。转换的图形被存储在图形存储器10B中。

[0067] 所述系统包括神经网络训练器单元14,所述神经网络训练器单元接收来自图形存储器的一组经解析的图形以及关于它们相互之间的关系的一些信息作为训练数据。在此情况下,提供了文档引用数据存储器10C,包括例如关于文档的引文数据和/或新颖性检索结果。训练器单元14运行基于图形的神经网络算法,所述神经网络算法产生用于基于神经网络的搜索引擎16的神经网络模型。引擎16使用来自图形存储器10B的图形作为目标搜索集合并且使用从用户接口18获得的用户数据——通常是文本或图形——作为参考。

[0068] 搜索引擎16可以是例如图形至向量搜索引擎,所述图形至向量搜索引擎被训练为查找最接近由用户数据形成的向量的与图形存储器10B的图形对应的向量。搜索引擎16还可以是分类器搜索引擎,诸如二元分类器搜索引擎,所述分类器搜索引擎将用户图形或从其导出的向量与从图形存储器10B获得的图形或从其导出的向量成对地进行比较。

[0069] 图1B示出了所述系统的一个实施方案,所述系统还包括文本嵌入单元13,所述文

本嵌入单元将图形的自然语言单元转换为多维向量格式。这是对来自图形存储器10B并且转换的图形以及通过用户接口18输入的图形进行的。通常，向量具有至少100个维度，诸如300个维度或更多个维度。

[0070] 在还示出于图1B中的一个实施方案中，神经网络搜索引擎16被分成形成一个系列的两部分。例如，引擎16包括图形嵌入引擎，所述图形嵌入引擎使用由神经网络训练器14的图形嵌入训练器14A使用来自文档引用数据存储器10C的引用数据训练的模型将图形转换为多维向量格式。在向量比较引擎16B中将用户图形与由图形嵌入引擎16A预先产生的图形进行比较。结果，查找到最接近用户图形的图形的缩小子集。图形的子集进一步通过图形分类器引擎16C与用户图形进行比较，以进一步缩小该组相关的图形。图形分类器引擎16C通过图形分类器训练器14C使用来自文档引用数据存储器10C的数据例如作为训练数据而被训练。此实施方案是有益的，因为通过向量比较引擎16B对预先形成的向量进行向量比较非常快，而图形分类器引擎可以访问图形的详细数据内容和结构，并且可以对图形进行准确比较以查找出它们之间的差异。图形嵌入引擎16A和向量比较引擎16B起图形分类器引擎16C的有效率的前置滤波器的作用，从而减少需要由图形分类器引擎16C处理的数据量。

[0071] 图形嵌入引擎可以将图形转换为具有至少100个维度、优选地200个维度或更多个维度并且甚至300个维度或更多个维度的向量。

[0072] 神经网络训练器14被分成两部分——图形嵌入部分和图形分类器部分，所述图形嵌入部分和图形分类器部分分别使用图形嵌入训练器14A和图形分类器训练器16C来训练。图形嵌入训练器14A形成基于神经网络的图形到向量模型，其目的是对于文本内容和内部结构相互相似的图形形成邻近向量(nearby vector)。图形分类器训练器14B形成分类器模型，所述分类器模型能够根据图形的文本内容和内部结构的相似度对图形对进行排序。

[0073] 从用户接口18获得的用户数据在在嵌入单元13中嵌入之后被馈送到图形嵌入引擎以用于向量化，在此之后向量比较引擎16B查找与图形存储器10B的图形对应的一组最接近的向量。该组最接近的图形被馈送到图形分类器引擎16C，所述图形分类器引擎使用经训练的图形分类器模型将它们与用户图形逐一进行比较，以获得准确的匹配。

[0074] 在一些实施方案中，图形嵌入引擎16A——如通过图形嵌入训练器14A训练的——输出向量，所述向量的角度相互越接近，图形在节点内容和节点结构方面就越相似，如使用依赖于其的学习目标从引用数据学习的。通过训练，从引用数据导出的肯定训练案例(描述相同概念的图形)的向量角可以被最小化，而否定训练案例(描述不同概念的图形)的向量角被最大化，或至少显著偏离零。

[0075] 图形向量可以被选择为具有例如200-1000个维度，诸如250-600个维度。

[0076] 已经发现这种监督机器学习模型能够有效率地评价由图形并且进一步由从其导出图形的自然语言块公开的技术概念的相似度。

[0077] 在一些实施方案中，图形分类器引擎16C——如通过图形分类器训练器14C训练的——输出相似度分数，所述相似度分数越高，被比较的图形在节点内容和节点结构方面就越相似，如使用依赖于其的学习目标从引用数据学习的。通过训练，从引用数据导出的肯定训练案例(描述相同概念的图形)的相似度得分可以被最大化，而否定训练案例(描述不同概念的图形)的相似度得分被最大化。

[0078] 余弦相似度是用于图形或从图形导出的向量的相似度的一个可能标准。

[0079] 应注意,图形分类器训练器14C或引擎16C不是强制性的,但是可以直接基于由图形嵌入引擎嵌入的向量之间的角度来评价图形相似度。出于此目的,可以使用本身已知的快速向量索引来查找给定的新鲜的图形向量的一个或多个邻近图形向量。

[0080] 训练器14和搜索引擎16或其子训练器14A、14C或子引擎16A、16C中的任何一个或两个所使用的神经网络可以是循环神经网络,特别是利用长短期记忆(Long Short-Term Memory,LSTM)单元的循环神经网络。在树结构图形的情况下,所述网络可以是Tree-LSTM网络,诸如Child-Sum-Tree-LSTM网络。所述网络可以具有一个或多个LSTM层和一个或多个网络层。所述网络可以使用注意机制,所述注意机制在训练和/或运行模型时使图形的各部分内部地或外部地相互有关。

[0081] 在下面在专利检索系统的上下文中描述本发明的一些另外的实施方案,由此被处理的文档是专利文档。上文所描述的一般实施方案和原理适用于所述专利检索系统。

[0082] 在某个实施方案中,所述系统被配置为在存储装置中存储自然语言文档,每个自然语言文档包含第一自然语言块和不同于所述第一自然语言块的第二自然语言块。训练器可以使用与第一文档的第一块对应的多个第一图形,并且对于每个第一图形,使用至少部分地基于与所述第一文档不同的第二文档的第二块的一个或多个第二图形,如由引用数据定义的。这样,神经网络模型从不同文档的不同部分之间的相互关系学习。另一方面,训练器可以使用与第一文档的第一块对应的多个第一图形,并且对于每个第一图形,使用至少部分地基于所述第一文档的第二块的第二图形。这样,神经网络模型可以从单个文档内的数据的内部关系学习。这两种学习方案可以由接下来详细描述的专利检索系统单独使用或同时使用。

[0083] 上文所讨论的精简图形表示特别适合于专利检索系统,即适合于权利要求图形和全说明书图形,特别是适合于全说明书图形。

[0084] 图1C示出了包括专利文档存储器10A的系统,所述专利文档存储器包含专利文档,所述专利文档至少包含计算机可识别的说明书部分和权利要求部分。图形解析器12被配置为通过权利要求图形解析器12A来解析权利要求,并且通过全说明书图形解析器12B来解析全说明书。经解析的图形被单独存储到权利要求和全说明书图形存储器10B。文本嵌入单元13准备用于在神经网络中处理的图形。

[0085] 引用数据可以包含公共专利申请和专利的检索和/或审查数据和/或专利文档之间的引文数据。在一个实施方案中,引用数据包含先前专利检索结果,即较早的专利文档被视为较晚提交的专利申请的新颖性和/或创造性障碍的信息。引用数据被存储在先前专利检索和/或引文数据存储器10C中。

[0086] 神经网络训练器14使用解析和嵌入的图形来形成特别是为了专利检索目的而训练的神经网络模型。这是通过使用专利检索和/或引文数据作为训练器14的输入来实现的。目的是例如使专利申请的权利要求图形和用作其新颖性障碍的专利文档的全说明书图形之间的向量角最小化或使其相似度分数最大化。这样,应用于多个(通常是数十万个或数百万个)权利要求,所述模型学习评价权利要求相对于现有技术的新颖性。对于通过用户接口18A获得的用户图形由搜索引擎16使用所述模型,以查找最有可能的新颖性障碍。结果可以被示出在检索结果视图界面18B中。

[0087] 图1C的系统可以利用一系列搜索引擎。所述引擎可以用从先前专利检索和/或引

文数据存储器10C获得的训练数据的相同或不同子集来训练。例如，人们可以使用通过大的或完整的引用数据集合(即肯定权利要求/全说明书对和否定权利要求/全说明书对)训练的图形嵌入引擎来从完整的现有技术数据集合过滤一组图形。然后在分类引擎中对照用户图形将被过滤的该组图形分类，所述分类引擎可以通过较小的例如专利分类特定的引用数据集合(即肯定权利要求/全说明书对和否定权利要求/全说明书对)来训练，以查找出图形的相似度。

[0088] 接下来，参考图2A和图2B描述特别适用于专利检索系统的树形图形结构。

[0089] 图2A示出了仅具有部分词关系作为边缘关系的树形图形。文本单元A-D作为线性递归节点10、12、14、16布置到所述图形内，起源于根节点10，并且文本单元E作为节点12的子、作为子节点18，如从所示出的自然语言块导出的。在本文中，从部分词/整体词表达“包括”、“具有”、“被包含在”和“包含”检测部分词关系。

[0090] 图2B示出了具有两个不同边缘关系的另一个树形图形，这两个不同边缘关系在此示例中是部分词关系(第一关系)和下位词关系(第二关系)。文本单元A-C被布置为具有部分词关系的线性递归节点10、12、14。文本单元D被布置为具有下位词关系的、父节点14的子节点26。文本单元E被布置为具有下位词关系的、父节点12的子节点24。文本单元F被布置为具有部分词关系的、节点24的子节点28。在本文中，从部分词/整体词表达“包括”、“具有”、“诸如”和“是例如”检测部分词关系和下位词关系。

[0091] 根据一个实施方案，第一数据处理装置适于通过首先从块中识别第一组自然语言符号(例如，名词和名词组块)和与所述第一组自然语言符号不同的第二自然语言符号(例如，部分词和整体词表达)来将所述块转换为图形。然后，利用第一组符号和第二组符号来执行匹配器，以用于形成第一组符号的匹配对(例如，来自“主体包括构件”的“主体”和“构件”)。最后，利用所述匹配对将第一组符号布置为所述图形的节点(例如，“主体”——(部分词边缘)——“构件”)。

[0092] 在一个实施方案中，在所述图形中使用至少部分词边缘，由此各个节点包含相互之间具有部分词关系的自然语言单元，如从所述块导出的。

[0093] 在一个实施方案中，在所述图形中使用下位词边缘，由此各个节点包含相互之间具有下位词关系的自然语言单元，如从自然语言块导出的。

[0094] 在一个实施方案中，在所述图形中使用边缘，所述图形的各个节点中的至少一个包含对同一图形中的一个或多个节点的引用以及附加地从相应的自然语言块导出的至少一个自然语言单元(例如，“在……以下”[节点id:X])。这样，图形空间被节省并且是简单的，例如树形、图形结构可以被维持，仍然允许所述图形中的有表现力的数据内容。

[0095] 在一些实施方案中，所述图形是树形图形，所述树形图形的节点值包含词或多词组块，所述词或多词组块是通常由图形转换单元利用词的词性和句法依赖性从所述自然语言块导出的或从其向量化形式导出的。

[0096] 图3详细示出了如何在第一数据处理装置中实施文本到图形转换的一个示例。首先，在步骤31中读取文本，并且从所述文本检测第一组自然语言符号(诸如名词)和第二组自然语言符号(诸如指示部分词性(meronymity)或整体词性(holonymity)的符号(例如“包括”)。这可以通过在步骤32中使所述文本符号化、33对符号进行词性(POS)标注、在步骤34中导出它们的句法依赖性来实现。使用该数据，可以在步骤35中确定名词组块，并且在步骤

36中确定部分词和整体词表达。在步骤37中，利用所述部分词和整体词表达形成名词组块的匹配对。名词组块对形成或可以被用来推断图形的部分词关系边缘。

[0097] 在一个实施方案中，如步骤38中所示出的，所述名词组块对被布置为树形图形，其中部分词是对应的整体词的子。在步骤39中将所述图形保存在图形存储器中以供进一步使用，如上文所讨论的。

[0098] 在一个实施方案中，图形形成步骤涉及使用诸如贝叶斯网络的概率图模型(PGM)，以用于推理优选的图形结构。例如，可以根据贝叶斯模型计算所述图形的不同边缘概率，在此之后使用所述边缘概率计算最可能的图形形式。

[0099] 在一个实施方案中，图形形成步骤包括将文本——通常是以符号化的、加POS标注和依赖性解析的形式——馈送到基于神经网络的技术解析器内，所述技术解析器从文本块查找相关组块并且提取它们的期望的边缘关系，诸如部分词关系和/或下位词关系。

[0100] 在一个实施方案中，所述图形是树形图形，所述树形图形包括根据树数据模式递归地布置的边缘关系，所述树数据模式是非循环的(acyclic)。这允许使用循环或非循环(non-recurrent)类型的有效率的基于树的神经网络模型。一个示例是Tree-LSTM模型。

[0101] 在另一个实施方案中，所述图形是允许循环(cycle)——即分支之间的边缘——的网络图形。这具有允许复杂的边缘关系被表达的益处。

[0102] 在又另一个实施方案中，所述图形是具有一个或多个边缘的长度(length)的线性和/或非线性分支的森林。线性分支具有的益处是：避免或显著简化树或网络构建步骤，并且最大量的源数据可用于神经网络。

[0103] 在每个模型中，边缘可能性——如果通过PGM模型获得——可以由神经网络存储和使用。

[0104] 应注意，如上文参考图3和本文档其他位置所描述的图形形成方法可以独立于在本文中所描述的其他方法和系统部分来实施，以形成和存储文档——特别是专利全说明书和权利要求书——的技术内容的技术精简表示。

[0105] 图4A-图4C示出了特别是为了专利检索目的而训练神经网络的不同但是相互不排斥的方法。

[0106] 对于一般情况，术语“专利文档”可以用“文档”(在所述系统中在其他文档之中具有唯一的计算机可读标识符)替换。“权利要求”可以用“第一计算机可识别的块”替换，并且“全说明书”可以用“至少部分地不同于所述第一块的第二计算机可识别的块”替换。

[0107] 在图4A的实施方案中，多个权利要求图形41A和对于每个权利要求图形对应的接近的现有技术全说明书图形42A，如通过引用数据相关的，由神经网络训练器44A用作训练数据。它们形成肯定训练案例，指示这样的图形之间的低向量角或高相似度得分将被实现。另外，对于每个权利要求图形，否定训练案例——即一个或多个遥远的现有技术图形——可以被用作训练数据的一部分。这样的图形之间的高向量角或低相似度得分将被实现。否定训练案例可以例如从完整的图形集合随机化。

[0108] 根据一个实施方案，在如由神经网络训练器44A实施的训练的至少一个阶段中，多个否定训练案例选自比所有可能的否定训练案例的平均值更难的所有可能训练案例的子集。例如，难的否定训练案例可以被选择，使得权利要求图形和说明书图形来自相同的专利分类(最高达预定分类水平)，或使得神经网络先前不能够正确地将说明书图形分类为否定

案例(以预定置信度)。

[0109] 根据一个实施方案——所述实施方案也可以独立于在本文中所描述的其他方法和系统部分来实施,通过提供多个专利文档来实施本基于神经网络的专利检索或新颖性评价系统的训练,每个所述专利文档具有计算机可识别的权利要求块和全说明书块,所述全说明书块包括专利文档的说明书的至少一部分。所述方法还包括提供神经网络模型和使用训练数据集合来训练所述神经网络模型,所述训练数据集合包括来自所述专利文档的用于形成经训练的神经网络模型的数据。训练包括使用来源于同一专利文档的权利要求块和全说明书块对作为所述训练数据集合的训练案例。

[0110] 通常,这些文档内部肯定训练案例形成训练的所有训练案例的一小部分,例如1%-25%,其余包含例如检索报告(审查员新颖性引文)训练案例。

[0111] 本机器学习模型通常被配置为将权利要求和全说明书转换为向量,并且所述模型的训练的学习目标可以是使同一专利文档的权利要求向量和全说明书向量之间的向量角最小化。另一个学习目标可以是使至少一些不同专利文档的权利要求向量和全说明书向量之间的向量角最大化。

[0112] 在图4B的实施方案中,来源于同一专利文档的多个权利要求图形41A和全说明书图形42A由神经网络训练器44B用作训练数据。一个权利要求的一个“自己的”全说明书通常形成一个完美的肯定训练案例。换言之,一个专利文档本身在技术上是其权利要求的理想的新颖性障碍。因此,这些图形对形成肯定训练案例,指示这样的图形之间的低向量角或高相似度分数将被实现。也在此场景下,可以使用引用数据和/或否定训练案例。

[0113] 测试已经示出,当用现实的基于新颖性检索的测试数据对进行测试时,简单地通过将来自同一文档的权利要求-说明书对添加到现实的基于新颖性检索的训练数据使现有技术的分类准确率提高15%以上。

[0114] 在典型情况下,在同一专利文档的全说明书中某处查找到一个权利要求的机器可读内容(自然语言单元,特别是词)的至少80%、通常至少90%、在很多情况下100%。因此,专利文档的权利要求和全说明书不仅通过认知内容和相同的唯一标识符(例如,公开号)而且还通过它们的字节级内容相互链接。

[0115] 根据一个实施方案——所述实施方案也可以独立于在本文中所描述的其他方法和系统部分来实施,本基于神经网络的专利检索或新颖性评价引擎的训练包括从至少一些原始权利要求或全说明书块导出至少一个部分地对应于原始块的缩减的数据实例,以及使用所述缩减的数据实例和所述原始权利要求或全说明书块一起作为所述训练数据集合的训练案例。

[0116] 在图4C的实施方案中,通过从原始权利要求图形41C’形成多个缩减的权利要求图形41C”-41C””来扩增肯定训练案例。缩减的权利要求图形是指这样的图形,其中

[0117] -至少一个节点被移除(例如,电话-显示器-传感器->电话-显示器)

[0118] -至少一个节点被移动到分支的更高(更一般)位置的另一个位置(例如,电话-显示器-传感器->电话-(显示器,传感器),和/或

[0119] -至少一个节点的自然语言单元值用更一般的自然语言单元值替换(电话-显示器-传感器->电子设备-显示器-传感器)。

[0120] 此种扩增方案允许用于神经网络的训练集合被扩大,从而产生更准确的模型。它

还允许仅用很少节点或用非常一般的术语针对所谓的琐碎发明(trivial invention)的新颖性进行有意义的检索和评价所谓的琐碎发明的新颖性,这在实际的专利新颖性检索数据中看到的至少不是很多。可以结合图4A和图4B的实施方案中的任一个或它们的组合来实施数据扩增。也在此场景下,可以使用否定训练案例。

[0121] 通过移除、移动或替换全说明书图形中的节点或它们的值,也可以扩增否定训练案例。

[0122] 树形图形结构,诸如基于部分词关系的图形结构,对于扩增方案是有益的,由于可以通过以直接并且鲁棒的方式删除节点或将节点移动到更高的树位置来扩增,因此仍然保留相干(coherent)逻辑。在此情况下,原始数据实例和缩减的数据实例都是图形。

[0123] 在一个实施方案中,缩减的图形是相对于原始图形或另一个缩减的图形已经删除至少一个叶节点的图形。在一个实施方案中,删除在图形的某个深度处的所有叶节点。

[0124] 还可以直接针对自然语言块实施本种扩增,特别是通过删除自然语言块的部分或将自然语言块的内容部分地改变为更一般的内容。

[0125] 每个原始实例的缩减的数据实例的数目可以是例如1-10000,特别是1-100。在具有2-50个扩增图形的权利要求扩增中实现良好的训练效果。

[0126] 在一些实施方案中,所述搜索引擎读取新鲜的自然语言块,诸如新鲜的权利要求,所述新鲜的权利要求通过转换器被转换为新鲜的图形,或通过用户接口直接读取新鲜的图形。接下来讨论适合于直接图形输入的用户接口。

[0127] 图5例示了用户接口的显示元件50上的示例性图形的表示和修改。显示元件50包括多个可编辑的数据小区(cell)A-F,其值在功能上连接到底层(underlying)图形的对应的自然语言单元(比如,对应地,单元A-F),并且被示出在各自的用户接口(UI)数据元件52、54、56、54’、56’、56”中。UI数据元件可以是例如文本字段,所述文本字段的值在启动所述元件之后可通过键盘编辑。UI数据元件52、54、65、54’、56’、56”根据它们在图形中的位置而被水平地和竖直地定位在显示元件50上。在本文中,水平位置对应于在图形中单元的深度。

[0128] 显示元件50可以是例如运行网络应用程序的网络浏览器的窗口、框架或面板,或在计算机中可执行的独立运行的程序的图形用户接口窗口。

[0129] 用户接口还包括移位引擎(shifting engine),所述移位引擎允许响应于用户输入而使自然语言单元在显示元件上水平地(竖直地)移动,并且以相应地修改图形。为了例示这,图5示出了数据小区F(元件56”)向左移位一级(箭头59A)。由于这,嵌套在元件54’下面的原始元件56”不复存在,并且形成嵌套在更高级元件52下面并且包括数据小区F(具有其原始值)的元件54”。如果此后数据元件54’向右移位两级(箭头59B),数据元件54’以及其子将向右移位并且嵌套在数据元件56下面作为数据元件56”和数据元件58。每个移位由底层图形中嵌套级的对应移位来反映。因此,当单元的子在用户接口中移位到不同的嵌套级时,单元的子将被保留在图形中。

[0130] 在一些实施方案中,UI数据元件包括自然语言辅助(helper)元件,所述辅助元件被示出与可编辑的数据小区有关,用于帮助用户输入自然语言数据。所述辅助元件的内容可以使用与有关的自然语言单元相关联的关系单元以及可选地其父元件的自然语言单元来形成。

[0131] 代替如图5中所例示的基于图形的用户接口,所述用户接口可以允许输入块文本,

诸如独立权利要求。文本块然后被馈送到图形解析器,以获得可在检索系统的另外的阶段使用的图形。

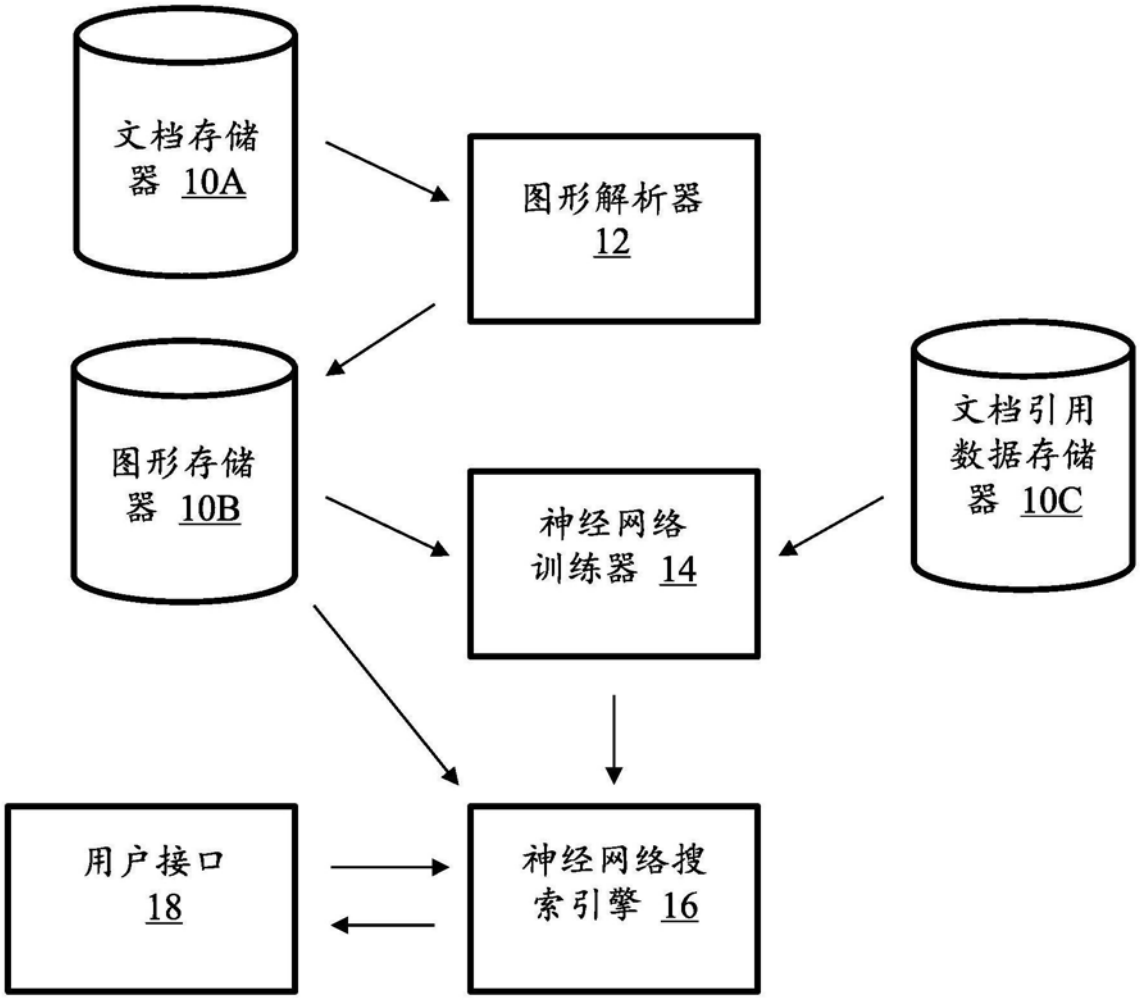

文档存储器 10A
图形解析器 12
图形存储器 10B
文档引用数据存储器 10C
神经网络训练器 14
用户接口 18
神经网络搜索引擎 16

图1A

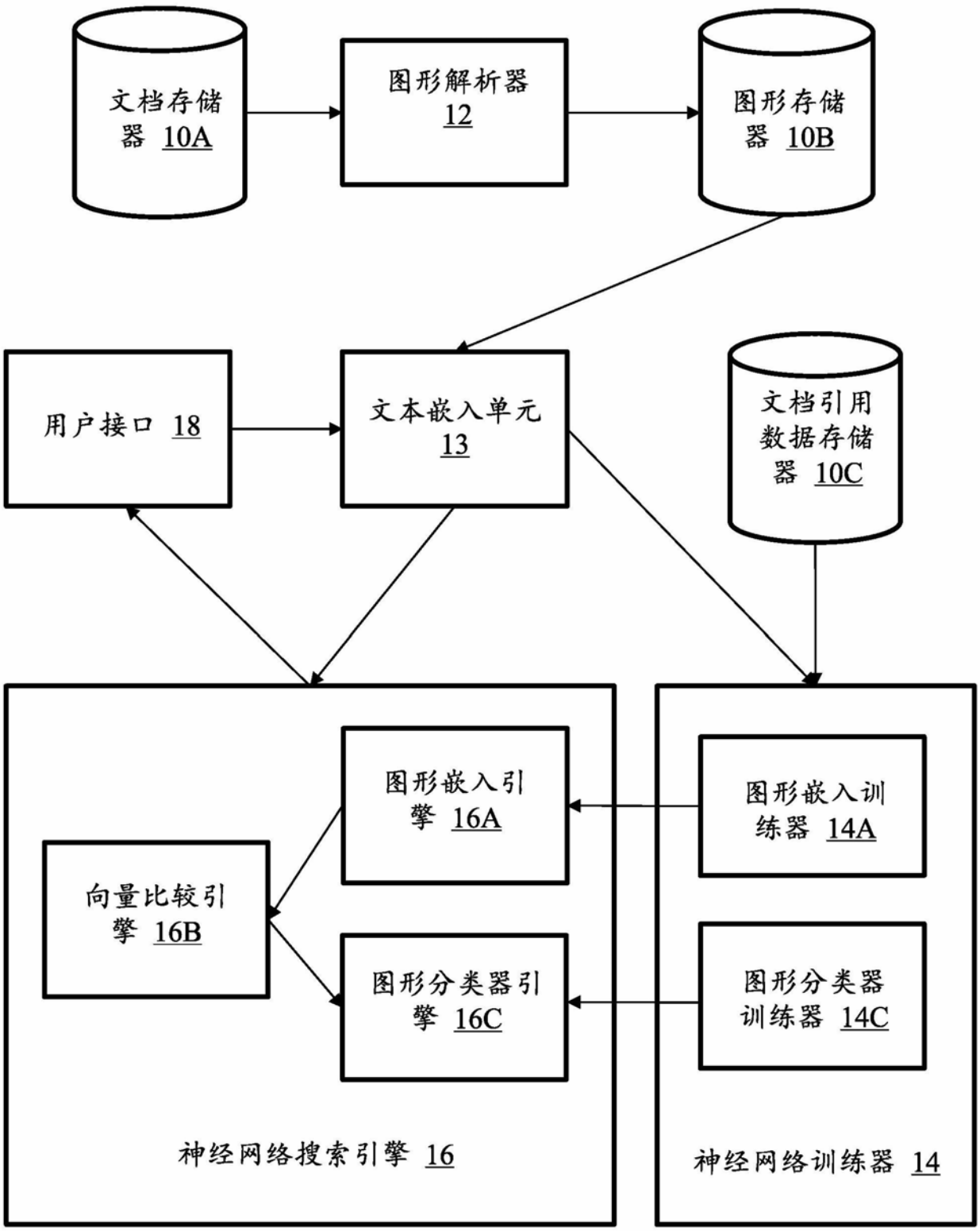

文档存储器 10A
图形解析器 12
图形存储器 10B
用户接口 18
文本嵌入单元 13
文档引用数据存储器 10C
图形嵌入引擎 16A
向量比较引擎 16B
图形分类器引擎 16C
神经网络搜索引擎 16
图形嵌入训练器 14A
图形分类器训练器 14C
神经网络训练器 14

图1B

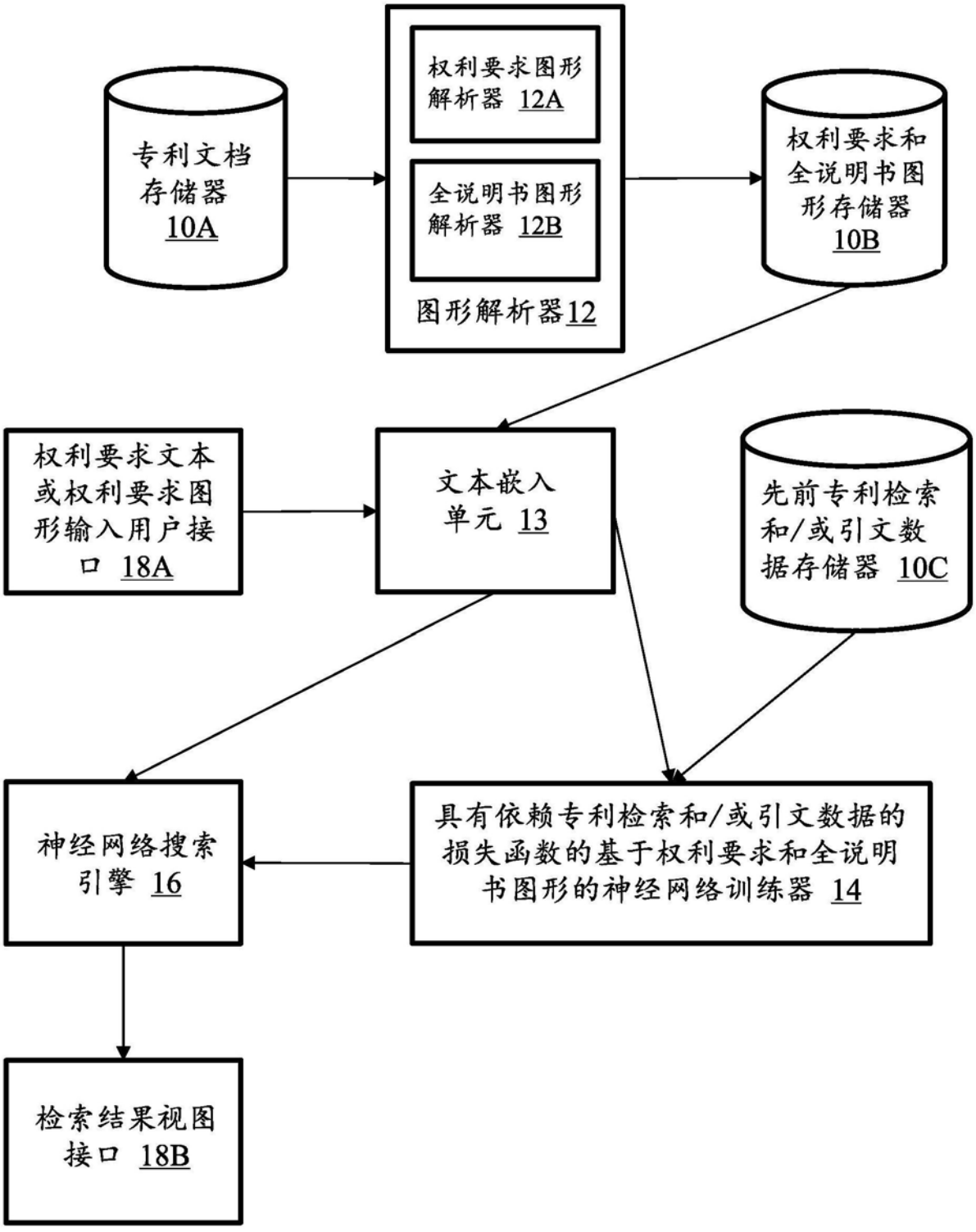

专利文档存储器 10A
权利要求图形解析器 12A
全说明书图形解析器 12B
图形解析器12
权利要求和全说明书图形存储器 10B
权利要求文本或权利要求图形输入用户接口 18A
文本嵌入单元 13
先前专利检索和/或引文数据存储器 10C
神经网络搜索引擎 16
具有依赖专利检索和/或引文数据的损失函数的基于权利要求和全说明书图形的神经网络训练器 14
检索结果视图接口 18B

图1C

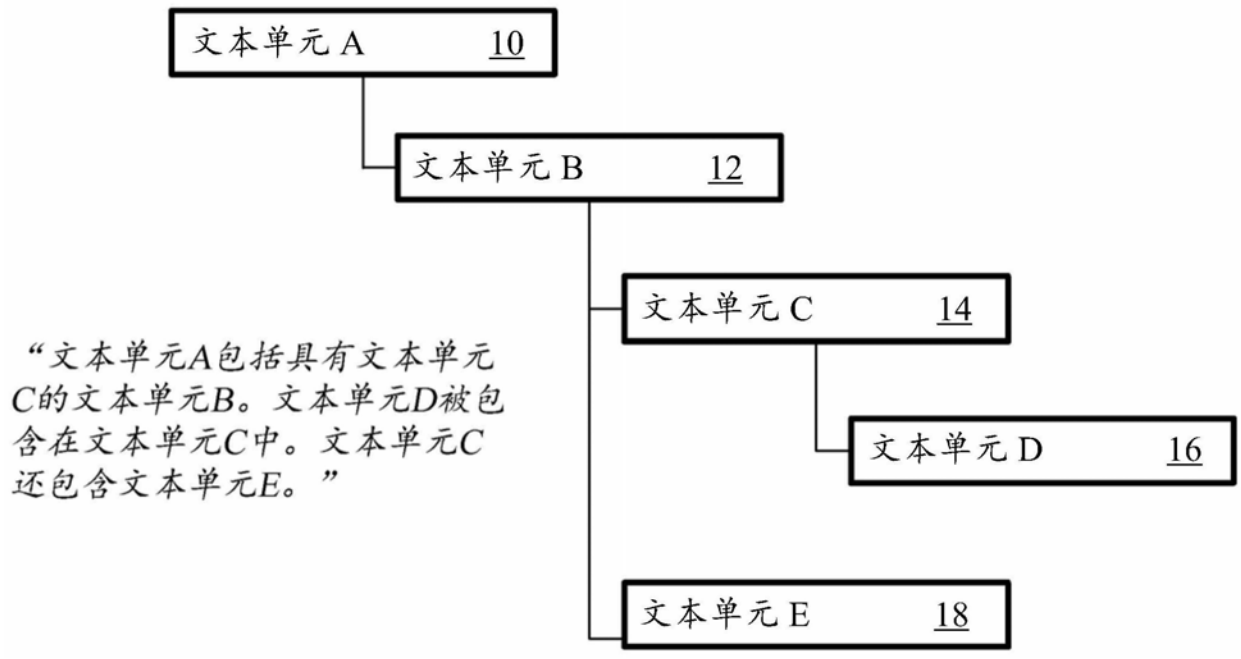

文本单元 A 10
文本单元 B 12
文本单元 C 14
文本单元 D 16
文本单元 E 18
“文本单元A包括具有文本单元C的文本单元B。文本单元D被包含在文本单元C中。文本单元C还包含文本单元E。”

图2A

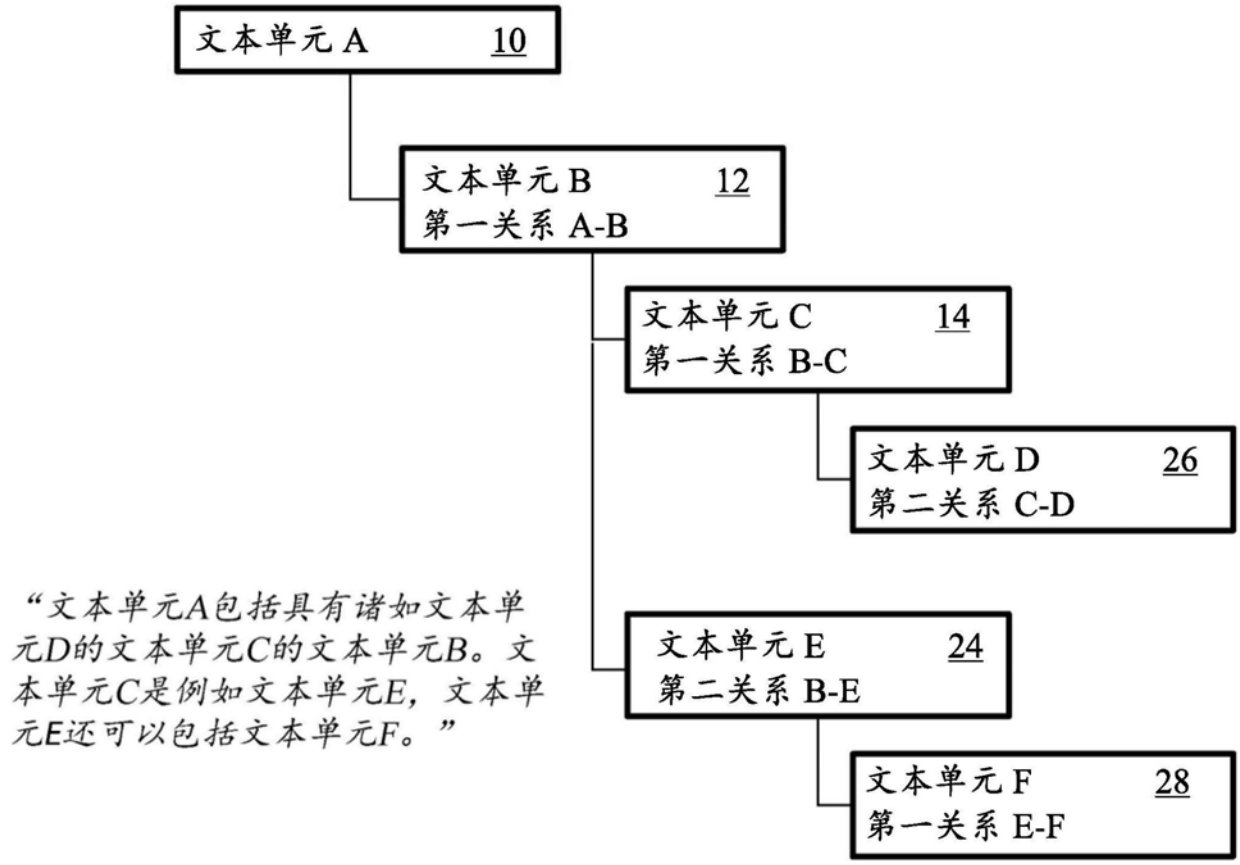

文本单元 A 10
文本单元 B 12
第一关系 A-B
文本单元 C 14
第一关系 B-C
文本单元 D 26
第二关系 C-D
文本单元 E 24
第二关系 B-E
文本单元 F 28
第一关系 E-F
“文本单元A包括具有诸如文本单元D的文本单元C的文本单元B。文本单元C是例如文本单元E，文本单元E还可以包括文本单元F。”

图2B

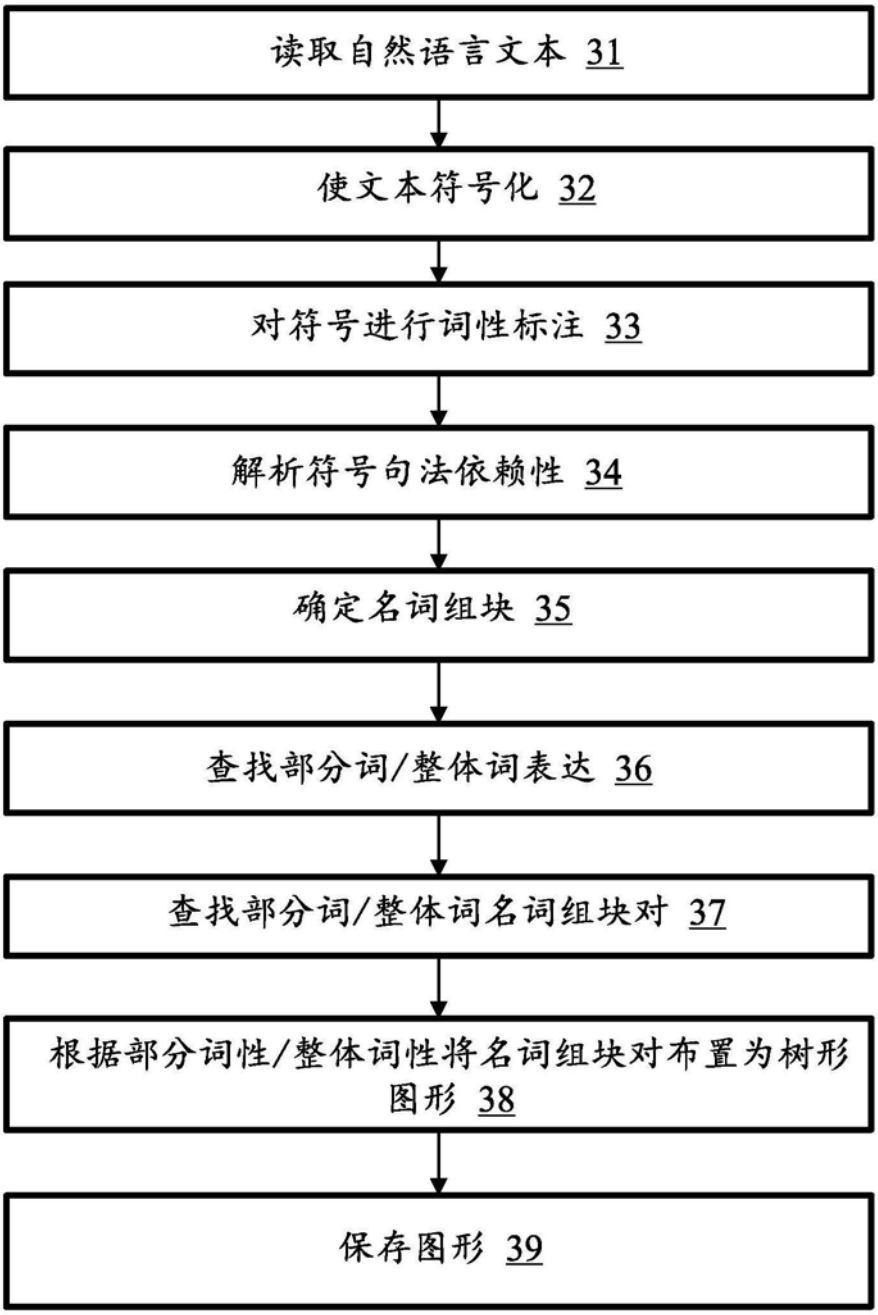

读取自然语言文本 31
使文本符号化 32
对符号进行词性标注 33
解析符号句法依赖性 34
确定名词组块 35
查找部分词/整体词表达 36
查找部分词/整体词名词组块对 37
根据部分词性/整体词性将名词组块对布置为树形图形 38
保存图形 39

图3

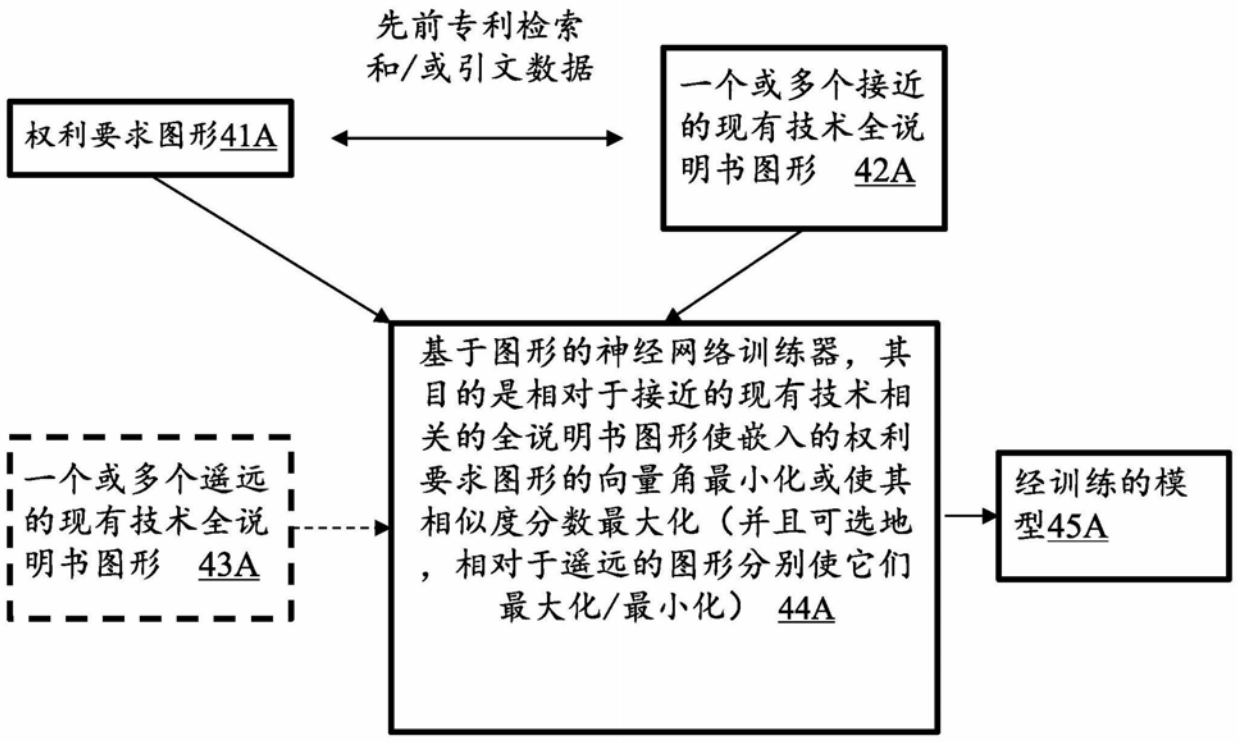

先前专利检索
和/或引文数据
权利要求图形41A
一个或多个接近的现有技术全说明书图形 42A
基于图形的神经网络训练器，其目的是相对于接近的现有技术相关的全说明书图形使嵌入的权利要求图形的向量角最小化或使其相似度分数最大化（并且可选地，相对于遥远的图形分别使它们最大化/最小化） 44A
一个或多个遥远的现有技术全说明书图形 43A
经训练的模型45A

图4A

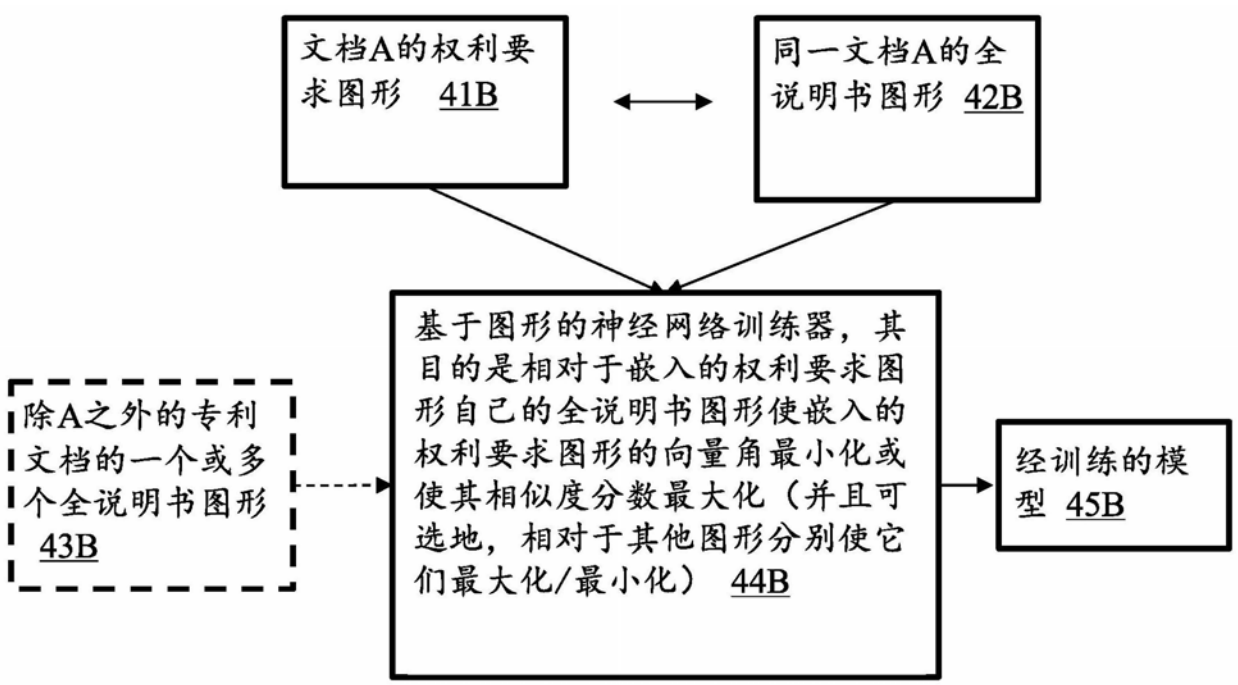

文档A的权利要求图形 41B
同一文档A的全说明书图形 42B
基于图形的神经网络训练器，其目的是相对于嵌入的权利要求图形自己的全说明书图形使嵌入的权利要求图形的向量角最小化或使其相似度分数最大化（并且可选地，相对于其他图形分别使它们最大化/最小化） 44B
除A之外的专利文档的一个或多个全说明书图形 43B
经训练的模型 45B

图4B

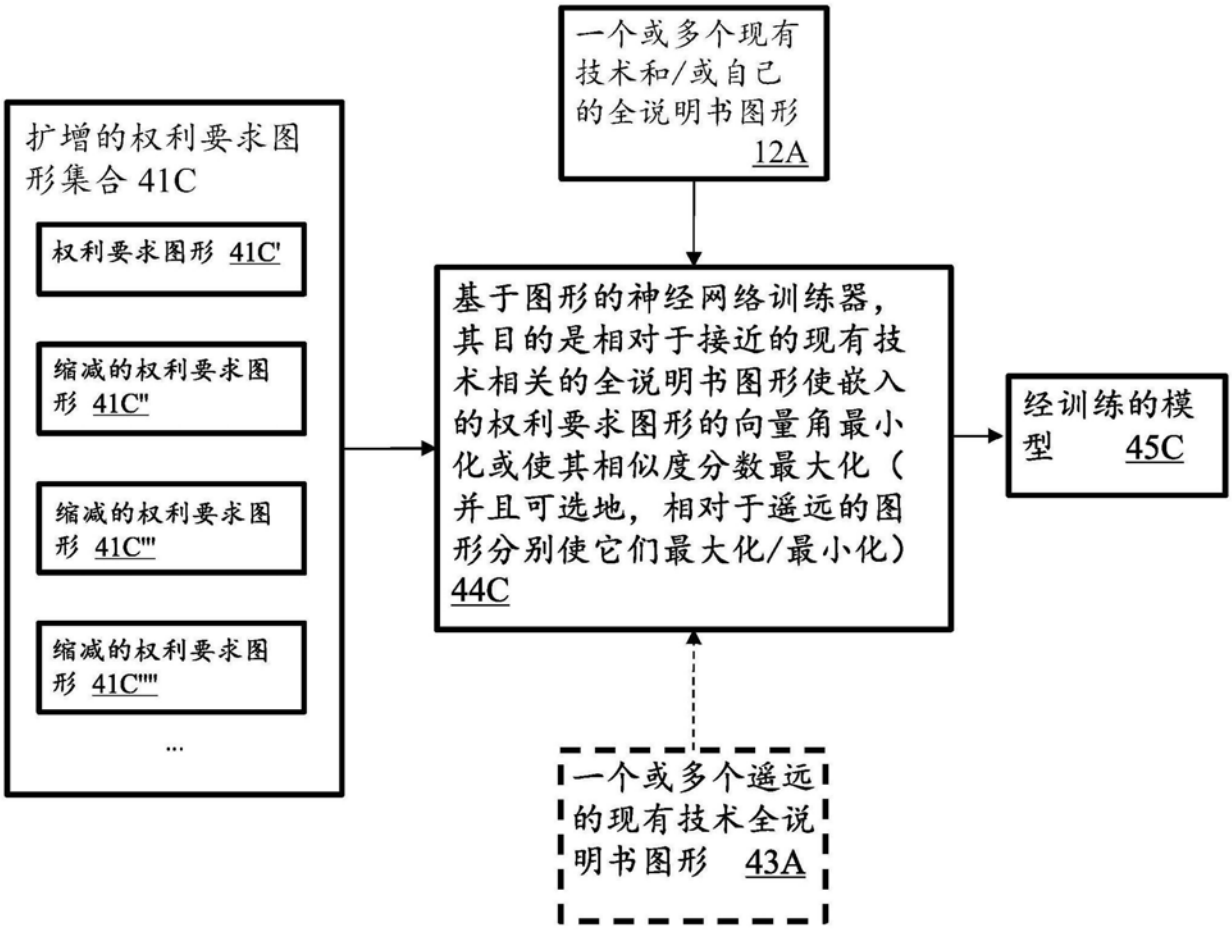

一个或多个现有技术和/或自己的全说明书图形 12A
扩增的权利要求图形集合 41C
权利要求图形 41C'
缩减的权利要求图形 41C''
缩减的权利要求图形 41C'''
缩减的权利要求图形 41C''''
...
基于图形的神经网络训练器，其目的是相对于接近的现有技术相关的全说明书图形使嵌入的权利要求图形的向量角最小化或使其相似度分数最大化（并且可选地，相对于遥远的图形分别使它们最大化/最小化）44C
经训练的模型 45C
一个或多个遥远的现有技术全说明书图形 43A

图4C

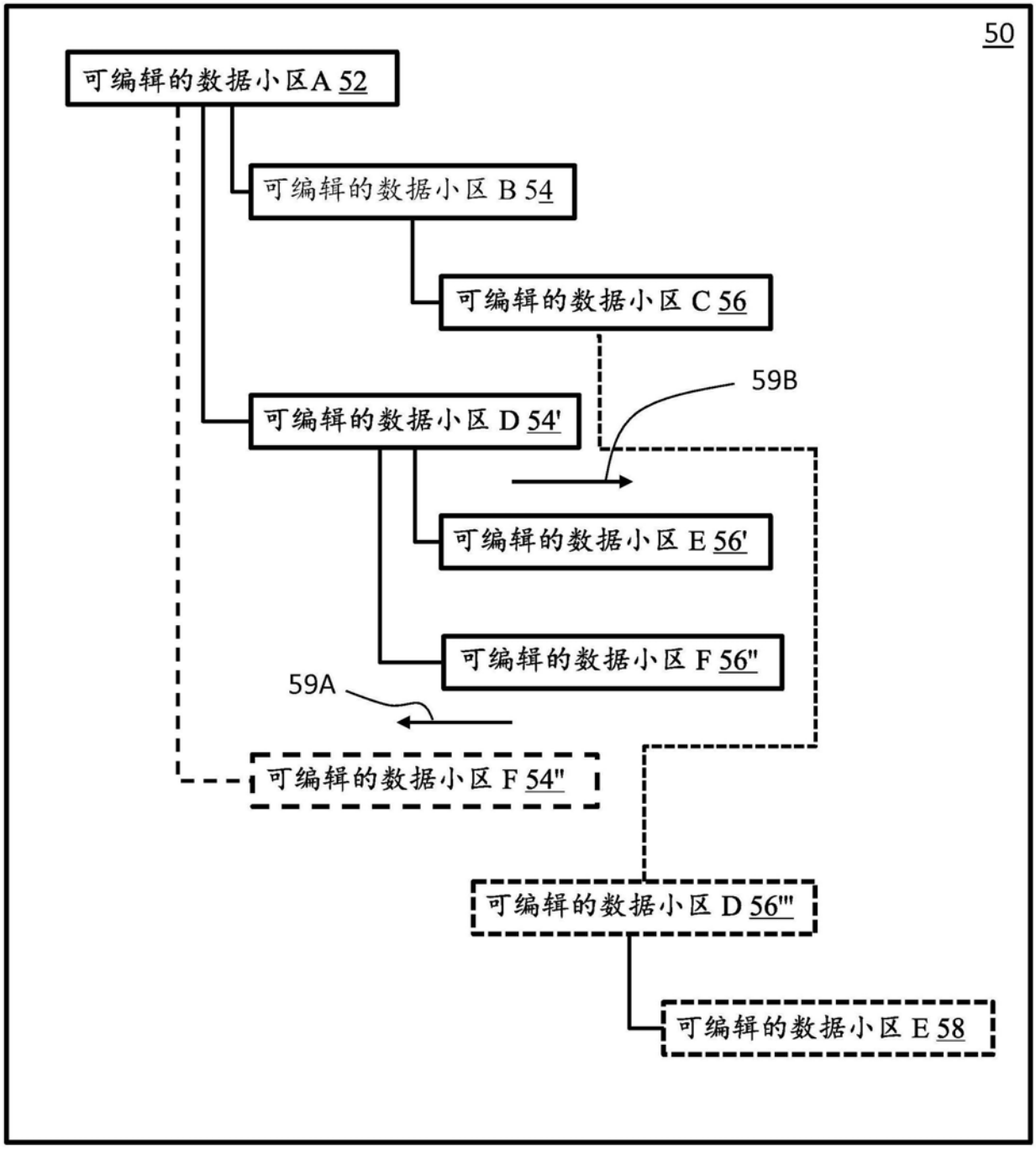

50
可编辑的数据小区A 52
可编辑的数据小区 B 54
可编辑的数据小区 C 56
59B
可编辑的数据小区 D 54'
可编辑的数据小区 E 56'
可编辑的数据小区 F 56"
59A
可编辑的数据小区 F 54"
可编辑的数据小区 D 56'"
可编辑的数据小区 E 58

图5