(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 105095229 A
(43)申请公布日 2015.11.25

(21)申请号 201410177307.9

(22)申请日 2014.04.29

(71)申请人 国际商业机器公司
地址 美国纽约

(72)发明人 郭宏蕾 钱伟红 郭志立 包胜华
苏中 D·帕塞多

(74)专利代理机构 北京市中咨律师事务所 11247
代理人 周良玉 于静

(51)Int.Cl.
G06F 17/30(2006.01)
G06F 17/27(2006.01)

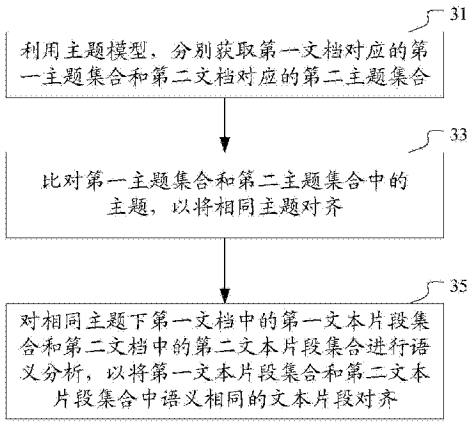
权利要求书2页 说明书12页 附图5页

(54)发明名称

训练主题模型的方法,对比文档内容的方法和相应的装置

(57)摘要

本发明公开了一种训练主题模型的方法和比对文档内容的方法以及相应的装置,上述训练主题模型的方法包括:提取文本片段的中心概念;为该中心概念构建特征向量,使得该特征向量包含中心概念在本体论中的关联信息;以及基于所构建的至少一个特征向量,训练主题模型。比对文档内容的方法包括:利用以上训练的主题模型,分别获取两个文档对应的两个主题集合;比对两个主题集合中的主题,将相同主题对齐;以及对相同主题下两个文档中的文本片段进行语义分析,以将语义相同的文本片段对齐。通过以上的方法和装置,可以基于中心概念的特征向量训练得到主题模型。利用这样的主题模型,可以实现文档语义内容的有效比对。



CN 105095229 A

1. 一种训练主题模型的方法，包括：

提取语料库文档中的文本片段的中心概念；

为所述中心概念构建特征向量，使得所述特征向量包含所述中心概念在本体论中的关联信息；以及

基于所构建的至少一个特征向量，训练主题模型。

2. 根据权利要求 1 所述的方法，其中所述关联信息通过以下方式获取：将所述中心概念映射到特定领域的本体树中，基于所述本体树中的信息获取所述关联信息。

3. 根据权利要求 2 所述的方法，其中所述关联信息包括所述中心概念的类信息，所述类信息包括以下中的一项或者多项：所述中心概念在所述本体树中的上位概念、下位概念和等价概念。

4. 根据权利要求 1 或 2 所述的方法，其中所述关联信息包括以下中的一项或多项：所述中心概念的领域信息，以及所述中心概念所对应的实体的属性特征信息。

5. 根据权利要求 1 所述的方法，其中所述特征向量还包括以下中的至少一项作为向量元素：与所述中心概念有关的搭配统计信息，以及所述中心概念在所述文本片段中的上下文信息。

6. 一种比对文档内容的方法，包括：

利用主题模型，分别获取第一文档对应的第一主题集合和第二文档对应的第二主题集合，其中所述主题模型基于为概念构建的特征向量而训练，所述特征向量包含所述概念在本体论中的关联信息；

比对所述第一主题集合和第二主题集合中的主题，以将相同主题对齐；以及

对相同主题下第一文档中的第一文本片段集合和第二文档中的第二文本片段集合进行语义分析，以将第一文本片段集合和第二文本片段集合中语义相同的文本片段对齐。

7. 根据权利要求 6 所述的方法，其中，获取第一文档对应的第一主题集合包括：

从第一文档的文本片段中提取出中心概念；

为所述中心概念构建特征向量，使得所述特征向量包含所述中心概念在本体论中的关联信息；

利用所述主题模型，基于所述特征向量，确定所述文本片段对应的主题；以及

将所述主题添加到第一主题集合。

8. 根据权利要求 7 所述的方法，其中所述关联信息包括以下中的至少一项：所述中心概念的领域信息，所述中心概念的类信息，以及所述中心概念所对应的实体的属性特征信息。

9. 根据权利要求 7 所述的方法，其中所述特征向量还包括以下中的至少一项作为向量元素：与所述中心概念有关的搭配统计信息，以及所述中心概念在所述文本片段中的上下文信息。

10. 根据权利要求 6-9 中任意一项所述的方法，其中所述第一文档和第二文档分别是用于描述两个地区在同一领域的法律法规的文档。

11. 一种训练主题模型的装置，包括：

概念提取单元，配置为提取语料库文档中的文本片段的中心概念；

向量构建单元，配置为，为所述中心概念构建特征向量，使得所述特征向量包含所述中

心概念在本体论中的关联信息；以及

训练单元，配置为基于所构建的至少一个特征向量，训练主题模型。

12. 根据权利要求 11 所述的装置，其中所述关联信息通过以下方式获取：将所述中心概念映射到特定领域的本体树中，基于所述本体树中的信息获取所述关联信息。

13. 根据权利要求 12 所述的装置，其中所述关联信息包括所述中心概念的类信息，所述类信息包括以下中的一项或者多项：所述中心概念在所述本体树中的上位概念、下位概念和等价概念。

14. 根据权利要求 11 或 12 所述的装置，其中所述关联信息包括以下中的一项或多项：所述中心概念的领域信息，以及所述中心概念所对应的实体的属性特征信息。

15. 根据权利要求 11 所述的装置，其中所述特征向量还包括以下中的至少一项作为向量元素：与所述中心概念有关的搭配统计信息，以及所述中心概念在所述文本片段中的上下文信息。

16. 一种比对文档内容的装置，包括：

主题获取单元，配置为利用主题模型，分别获取第一文档对应的第一主题集合和第二文档对应的第二主题集合，其中所述主题模型基于为概念构建的特征向量而训练，所述特征向量包含所述概念在本体论中的关联信息；

主题比对单元，配置为比对所述第一主题集合和第二主题集合中的主题，以将相同主题对齐；以及

文本片段分析单元，配置为对相同主题下第一文档中的第一文本片段集合和第二文档中的第二文本片段集合进行语义分析，以将第一文本片段集合和第二文本片段集合中语义相同的文本片段对齐。

17. 根据权利要求 16 所述的装置，其中所述主题获取单元包括：

概念提取模块，配置为从第一文档的文本片段中提取出中心概念；

向量构建模块，配置为为所述中心概念构建特征向量，使得所述特征向量包含所述中心概念在本体论中的关联信息；

主题确定模块，配置为利用所述主题模型，基于所述特征向量，确定所述文本片段对应的主题；以及

主题添加模块，配置为将所述主题添加到第一主题集合。

18. 根据权利要求 17 所述的装置，其中所述关联信息包括以下中的至少一项：所述中心概念的领域信息，所述中心概念的类信息，以及所述中心概念所对应的实体的属性特征信息。

19. 根据权利要求 17 所述的装置，其中所述特征向量还包括以下中的至少一项作为向量元素：与所述中心概念有关的搭配统计信息，以及所述中心概念在所述文本片段中的上下文信息。

20. 根据权利要求 16-19 中任意一项所述的装置，其中所述第一文档和第二文档分别是用于描述两个地区在同一领域的法律法规的文档。

训练主题模型的方法，对比文档内容的方法和相应的装置

技术领域

[0001] 本发明涉及文档内容分析，更具体地，涉及一种主题模型的构建和利用构建的主题模型对比文档内容。

背景技术

[0002] 在计算机信息处理领域中，许多应用和工具能够提供对文档内容进行分析和比对的功能。例如，搜索引擎可以对文档内容进行初步的语义分析，以确定该文档与搜索的关键词之间的相关性。还提供有一些版本管理工具，通过对不同版本的文档进行对比，来追踪、记录不同版本下文档内容的变化。

[0003] 然而，用户有时候需要对两篇内容相似的文档进行语义上的对比，以确定和区分语义上相似或相同的部分，以及语义上不相关的部分。例如，在一个例子中，两篇文档分别描述了两种相似的操作系统的功能特点；用户希望分析和对比这两篇文档，以获知这两种操作系统中有哪些相同的功能特点。在另一例子中，两篇文档分别描述了不同地区对于电池的使用和废弃的法律规定；用户希望通过对比这两篇文档，确定这两个地区对于电池的废弃的规定有什么不同。在以上的两个例子中，两篇文档虽然记录了相似的内容，但是其描述方式可能具有较大的差异。例如，两篇文档可能具有完全不同的文档结构，从不同角度和方面来描述同一主题，还可能采用不同的用语来表达同一概念。这为文档的分析和比对带来了困难。

[0004] 现有的搜索引擎一般可以用于衡量一篇文档和给定关键词的相关性，有些搜索引擎的算法甚至可以从总体上衡量两篇文档的相关性。但是它们仍然无法对不同文档的各个部分进行语义上的分析和对齐。现有的版本管理工具仅对文档进行字面上的比对，无法提取其语义信息。面对不同文档结构、不同用语的两篇文档，版本管理工具无法实现语义上的对比和分析。因此，希望提出改进的方案，能够对文档进行语义上的分析和比对，以满足用户的需求。

发明内容

[0005] 考虑到现有技术中的不足，提出本发明，以提供一种基于本体论的主题模型，并利用这样的主题模型实现文档内容的比对。

[0006] 根据本发明的第一方面，提供了一种训练主题模型的方法，包括：提取语料库文档中的文本片段的中心概念；为所述中心概念构建特征向量，使得所述特征向量包含所述中心概念在本体论中的关联信息；以及基于所构建的至少一个特征向量，训练主题模型。

[0007] 根据本发明的第二方面，提供了一种比对文档内容的方法，包括：利用主题模型，分别获取第一文档对应的第一主题集合和第二文档对应的第二主题集合，其中所述主题模型基于为概念构建的特征向量而训练，所述特征向量包含所述概念在本体论中的关联信息；比对所述第一主题集合和第二主题集合中的主题，以将相同主题对齐；以及对相同主题下第一文档中的第一文本片段集合和第二文档中的第二文本片段集合进行语义分析，以

将第一文本片段集合和第二文本片段集合中语义相同的文本片段对齐。

[0008] 根据本发明第三方面，提供了一种训练主题模型的装置，包括：概念提取单元，配置为提取语料库文档中的文本片段的中心概念；向量构建单元，配置为，为所述中心概念构建特征向量，使得所述特征向量包含所述中心概念在本体论中的关联信息；以及训练单元，配置为基于所构建的至少一个特征向量，训练主题模型。

[0009] 根据本发明第四方面，提供了一种比对文档内容的装置，包括：主题获取单元，配置为利用主题模型，分别获取第一文档对应的第一主题集合和第二文档对应的第二主题集合，其中所述主题模型基于为概念构建的特征向量而训练，所述特征向量包含所述概念在本体论中的关联信息；主题比对单元，配置为比对所述第一主题集合和第二主题集合中的主题，以将相同主题对齐；以及文本片段分析单元，配置为对相同主题下第一文档中的第一文本片段集合和第二文档中的第二文本片段集合进行语义分析，以将第一文本片段集合和第二文本片段集合中语义相同的文本片段对齐。

[0010] 通过以上的方法和装置，可以训练得到有效反映主题与实体之间的语义关联的主题模型。利用这样的主题模型，可以确定出不同文档中的主题序列，进而对相同主题下的文本片段进行语义分析，实现文档语义内容的有效比对。

附图说明

[0011] 通过结合附图对本公开示例性实施方式进行更详细的描述，本公开的上述以及其它目的、特征和优势将变得更加明显，其中，在本公开示例性实施方式中，相同的参考标号通常代表相同部件。

[0012] 图 1 示出了适于用来实现本发明实施方式的示例性计算机系统 / 服务器 12 的框图；

[0013] 图 2 示出根据本发明一个实施例的训练主题模型的方法的流程图；

[0014] 图 3 示出根据本发明一个实施例的比对文档内容的方法的流程图；

[0015] 图 4 示出根据一个实施例获得第一主题集合的步骤；

[0016] 图 5A 示例性示出第一文档和第二文档的主题的对齐；

[0017] 图 5B 示例性示出图 5A 的例子中的文本片段的对齐；

[0018] 图 6 示出根据本发明一个实施例的训练主题模型的装置的示例性框图；以及

[0019] 图 7 示出根据本发明一个实施例的比对文档内容的装置的示例性框图。

具体实施方式

[0020] 下面将参照附图更详细地描述本发明的优选实施方式。虽然附图中显示了本发明的优选实施方式，然而应该理解，可以以各种形式实现本发明而不应被这里阐述的实施方式所限制。相反，提供这些实施方式是为了使本发明更加透彻和完整，并且能够将本发明的范围完整地传达给本领域的技术人员。

[0021] 图 1 示出了适于用来实现本发明实施方式的示例性计算机系统 / 服务器 12 的框图。图 1 显示的计算机系统 / 服务器 12 仅仅是一个示例，不应对本发明实施例的功能和使用范围带来任何限制。

[0022] 如图 1 所示，计算机系统 / 服务器 12 以通用计算设备的形式表现。计算机系统

/ 服务器 12 的组件可以包括但不限于 :一个或者多个处理器或者处理单元 16,系统存储器 28,连接不同系统组件 (包括系统存储器 28 和处理单元 16) 的总线 18。

[0023] 总线 18 表示几类总线结构中的一种或多种,包括存储器总线或者存储器控制器,外围总线,图形加速端口,处理器或者使用多种总线结构中的任意总线结构的局域总线。举例来说,这些体系结构包括但不限于工业标准体系结构 (ISA) 总线,微通道体系结构 (MAC) 总线,增强型 ISA 总线、视频电子标准协会 (VESA) 局域总线以及外围组件互连 (PCI) 总线。

[0024] 计算机系统 / 服务器 12 典型地包括多种计算机系统可读介质。这些介质可以是任何能够被计算机系统 / 服务器 12 访问的可用介质,包括易失性和非易失性介质,可移动的和不可移动的介质。

[0025] 系统存储器 28 可以包括易失性存储器形式的计算机系统可读介质,例如随机存取存储器 (RAM)30 和 / 或高速缓存存储器 32。计算机系统 / 服务器 12 可以进一步包括其它可移动 / 不可移动的、易失性 / 非易失性计算机系统存储介质。仅作为举例,存储系统 34 可以用于读写不可移动的、非易失性磁介质 (图 1 未显示,通常称为“硬盘驱动器”)。尽管图 1 中未示出,可以提供用于对可移动非易失性磁盘 (例如“软盘”) 读写的磁盘驱动器,以及对可移动非易失性光盘 (例如 CD-ROM,DVD-ROM 或者其它光介质) 读写的光盘驱动器。在这些情况下,每个驱动器可以通过一个或者多个数据介质接口与总线 18 相连。存储器 28 可以包括至少一个程序产品,该程序产品具有一组 (例如至少一个) 程序模块,这些程序模块被配置以执行本发明各实施例的功能。

[0026] 具有一组 (至少一个) 程序模块 42 的程序 / 实用工具 40,可以存储在例如存储器 28 中,这样的程序模块 42 包括——但不限于——操作系统、一个或者多个应用程序、其它程序模块以及程序数据,这些示例中的每一个或某种组合中可能包括网络环境的实现。程序模块 42 通常执行本发明所描述的实施例中的功能和 / 或方法。

[0027] 计算机系统 / 服务器 12 也可以与一个或多个外部设备 14(例如键盘、指向设备、显示器 24 等) 通信,还可与一个或者多个使得用户能与该计算机系统 / 服务器 12 交互的设备通信,和 / 或与使得该计算机系统 / 服务器 12 能与一个或多个其它计算设备进行通信的任何设备 (例如网卡,调制解调器等等) 通信。这种通信可以通过输入 / 输出 (I/O) 接口 22 进行。并且,计算机系统 / 服务器 12 还可以通过网络适配器 20 与一个或者多个网络 (例如局域网 (LAN),广域网 (WAN) 和 / 或公共网络,例如因特网) 通信。如图所示,网络适配器 20 通过总线 18 与计算机系统 / 服务器 12 的其它模块通信。应当明白,尽管图中未示出,可以结合计算机系统 / 服务器 12 使用其它硬件和 / 或软件模块,包括但不限于 :微代码、设备驱动器、冗余处理单元、外部磁盘驱动阵列、RAID 系统、磁带驱动器以及数据备份存储系统等。

[0028] 下面通过图 2 到图 4 描述根据本发明实施例的训练主题模型和对比文档内容的方法步骤。在这些实施例中,基于文档中各个中心概念在本体论中的关联信息来训练主题模型,使得主题模型能够体现各个概念的深层语义信息以及概念之间的关系。进一步地,基于如此训练的主题模型,就可以获得文档中包含的主题,并实现主题级对齐。接着,可以对各个主题下的文本片段进行语义分析,从而实现文本片段的对齐。

[0029] 如本领域技术人员所知,主题模型是对文字中隐含主题的一种建模方法,常常用于语义挖掘和语义分析。根据现有技术的主题模型,典型地,用一个特定的词频分布来刻画

一个主题。更具体来说，可以认为，一个文档中的每个词都是通过以一定概率选择某个主题，并从这个主题以一定概率选择某个词语来得到的。这一过程可以表示为：

[0030] P(词语|文档)=Σp(词语|主题)*p(主题|文档)

[0031] 或者，可以将上式表示为矩阵相乘的形式：

[0032] $C_{ij}=\Phi_{ik}\times\theta_{kj}$ (公式1)

[0033] 其中C_{ij}表示词语i在文档j中的出现概率，Φ_{ik}表示词语i在主题k中的出现概率，θ_{kj}表示主题k在文档j中的出现概率。由于每篇文档可以表示为一个词语的集合，因此，可以通过用词语i出现的次数除以文档j中词语的总数目来得到C_{ij}。也就是说，对于语料库中的文档，左边的矩阵C_{ij}是通过简单的计算就可获知的，而右侧的两个矩阵是未知的。这样，可以利用大量的文档和相应的矩阵C_{ij}，通过一系列训练，推理出右侧的“词语-主题”矩阵Φ_{ik}和“主题-文档”矩阵θ_{kj}。

[0034] 为了推理出上述两个矩阵，现有技术中提出了多种训练和推理方法，常用的有pLSA(概率潜在语义分析)方法和LDA(潜在狄利克雷分配)方法。pLSA方法采用期望最大化的算法对两个矩阵进行反复迭代计算，最后得到收敛的、趋近于真实的Φ_{ik}和θ_{kj}。LDA方法假定文档与主题之间服从狄利克雷分布，主题与词语之间服从多项式分布，并采用Gibbs采样方法进行采样和抽取，最终推断出上述两个矩阵。

[0035] 由此，在主题模型的训练过程中，以各个词语在各个文档中的出现频率作为输入，无需人为地对主题进行标注，就可以获得“词语-主题”矩阵Φ_{ik}和“主题-文档”矩阵θ_{kj}。由于矩阵Φ_{ik}示出词语i在主题k中的出现概率，通过该矩阵，就可以将主题表示为多个词语分布的集合。

[0036] 然而，本发明的提出者认为，以上的方法并未考虑到词语之间的语义关联关系，因此得到的主题模型不能体现出深层的语义信息。因此，在本发明的实施例中，结合文档中各个中心概念在本体论中的信息来进行主题模型的训练。

[0037] 具体地，图2示出根据本发明一个实施例的训练主题模型的方法的流程图。如图2所示，该实施例中训练主题模型的方法包括以下步骤：步骤21，提取语料库文档中的文本片段的中心(focused)概念；步骤23，为所述中心概念构建特征向量，使得所述特征向量包含所述中心概念在本体论中的关联信息；以及步骤25，基于所构建的至少一个特征向量，训练主题模型。下面详细各个步骤的执行过程。

[0038] 首先，在步骤21，提取语料库文档中的文本片段的中心概念。可以理解，语料库可以是用于模型训练的大量文档的集合。这些文档可以涉及各种不同领域以及不同主题。对于任意文档，可以将其划分为若干文本片段。上述文本片段可以是文档中自然形成的段落或句子，也可以是人为划分的一段文字，或者是其他形式。在一个典型例子中，上述文本片段是文档中的一个句子。

[0039] 对于以上所述的文本片段，在步骤21，可以利用语言学的分析，从中提取出中心概念。这里认为，中心概念是对文本片段所集中描述的实体的抽象表达。在语言学上，中心概念往往表现为句子中的核心名词等有限的形式。现有的计算机语言学分析已经能够区分句子的各个成分并确定词汇之间的修饰关系。因此，利用语言学分析，可以从文本片段中提取出至少一个中心概念。下面示例性示出作为文本片段的几个句子。

[0040] “任何人不得销售或者许诺销售，形状和大小类似于纽扣或者硬币、且汞含量高于

25 毫克的碱性锰电池”。（文本片段 1）

[0041] “制造商不得销售、散布、许诺销售除碱性锰纽扣电池之外的添加有汞的碱性锰电池，除非得到委员会的授权”。（文本片段 2）

[0042] “应该在电池包装上清楚标注电池的类型和制造商的名字”。（文本片段 3）

[0043] 通过语言学分析，可以从以上文本片段 1 和文本片段 2 中提取出词汇“碱性锰电池”作为中心概念，从文本片段 3 中提取出“电池包装”作为中心概念。

[0044] 接着，在步骤 23，为提取的中心概念构建特征向量，使得该特征向量包含所述中心概念在本体论中的关联信息。在该步骤中，需要结合本体论的知识来构建特征向量。

[0045] 如本领域技术人员所知，本体论 (Ontology) 原本是一个哲学概念，用于研究客观事物存在的本质。但近年来，随着信息技术的发展，这一理论被应用到计算机信息处理领域，并在人工智能、计算机语言以及数据库理论中发挥着重要的作用。

[0046] 在信息处理领域中，本体论可以用于对某个领域 (domain) 中的概念及其关系进行说明。具体地，本体论的基本元素是术语 (term) 或概念，其中具有某些相同属性的术语或概念可以构成类和子类。本体论还描述各个类和概念之间的关系。某个领域中这样的概念及其关系的总和可以称为该领域的本体。在形式上，一个领域的本体可以表现为描述该领域中各个概念的词汇表，该词汇表可以整理为树形结构，以示出各个概念之间的关系。这样的树形结构的词汇表又可以称为本体树。

[0047] 基于以上所述的本体论的体系知识，就可以对步骤 21 中提取的中心概念进行深层次的语义分析和信息挖掘。具体地，在步骤 23 中，首先将上述中心概念映射到某个领域的本体树中，然后基于该本体树中的信息获取该中心概念在本体论中的关联信息。

[0048] 在一个实施例中，上述关联信息包括上述中心概念的领域信息。如前所述，本体论依据不同领域来组织概念，形成本体树。因此，当将上述中心概念映射到某个本体树时，就可以将该本体树对应的领域作为该中心概念的领域。在此基础上，还可以确定该领域的上位领域。例如，假定从文本片段 1 中提取的中心概念“碱性锰电池”可以被映射到针对电池领域组织的本体树中，那么可以认为该中心概念所属的领域为电池领域。进一步地，可以确定该领域的上位领域，例如是电子学领域。

[0049] 在一个实施例中，上述关联信息包括上述中心概念的类信息。具体地，上述类信息可以包括以下中的一项或者多项：上述中心概念在对应的本体树中的上位概念、下位概念、等价概念（如果有的话）。这可以通过查询对应的本体树而获得。例如，对于从文本片段 1 中提取的中心概念“碱性锰电池”，可以基于对应的电池领域的本体树获知，其上位概念包括：化学电池、纽扣电池等，其下位概念包括碱性锰纽扣电池等。

[0050] 在一个实施例中，上述关联信息包括上述中心概念所对应的实体的属性特征信息。在一些情况下，本体论根据与概念对应的实体的属性特征对概念进行语义归类。依据不同的属性特征，同一概念可以归属于不同的语义特征类。此时，可以依据语义归类信息获取对应实体的属性信息。例如，对于前述的中心概念“碱性锰电池”，对应的实体的属性特征可能包括大小、重量、形状、成分等。在一个实施例中，这样的属性特征信息也可以从文本片段中提取。例如，文本片段 1 用“形状”和“大小”来限定中心概念“碱性锰电池”，那么就可以将这样的限定作为“碱性锰电池”的属性特征信息。

[0051] 在以上描述的关联信息的基础上，本领域技术人员还可以基于本体论的知识，获

取更多的与提取的中心概念有关的关联信息。此外，还可以获取更多与中心概念有关的其他信息作为向量元素，用于构建特征向量。

[0052] 例如，在一个实施例中，如果提取的中心概念是复合词汇，那么可以获取该复合词汇的内部词汇信息，并将其包含在特征向量中。例如，以上所述的中心概念“碱性锰电池”为复合词汇，可以拆分为内部词汇元素“碱性”、“锰”和“电池”。可以将这样的内部词汇信息作为向量元素包含在特征向量中。

[0053] 在一个实施例中，可以获取与中心概念有关的搭配统计信息作为特征向量的向量元素。可以理解，通过预先学习大量文档，可以获得与词汇搭配相关的信息。或者，可以在根据本发明实施例的步骤扫描文档片段的同时，对词汇搭配信息进行统计，形成与词汇搭配有关的信息。利用这样的信息，可以直接获得与中心概念有关的搭配统计信息。如此获得的与中心概念有关的搭配统计信息可以示出，例如，中心概念经常与哪些词汇搭配在一起同时出现。例如，在一个例子中，上述搭配统计信息包括，以较高概率（例如高于预先确定的第一阈值）与上述中心概念在同一文档片段中一同出现的其他概念。可选地，上述统计信息还可以包括，与上述中心概念“互斥”的概念，也就是，以较低概率（例如低于预先确定的第二阈值）与上述中心概念一同出现的其他概念。例如，对于上述中心概念“碱性锰电池”，利用与词汇搭配有关的统计信息可以确定，常常与其一同出现的概念有“汞”、“含量”等，与其互斥的概念有“镍镉电池”、“锌碳电池”等。这些信息可以作为中心概念“碱性锰电池”的搭配统计信息。

[0054] 在一个实施例中，还可以获取所提取的中心概念在所述文本片段中的上下文信息作为特征向量的向量元素。在一个例子中，所述上下文信息包括，在所述文本片段中，所述中心概念附近（例如，距离小于某阈值）的其他概念。在另一例子中，上述上下文信息包括，所述文本片段中的关键动词或动词短语。在另一例子中，上述上下文信息还包括，所述文本片段中的其他关键名词或名词短语。例如，对于文本片段 1 中的中心概念“碱性锰电池”，可以从文本片段 1 中提取动词“销售”和“许诺销售”作为上下文信息。

[0055] 利用以上所述的多种信息，可以为文本片段的中心概念构建特征向量。相比于现有技术的方法，步骤 23 构建的特征向量反映了更深层次的信息。例如，对于以上举例的文本片段 1，现有技术的主题模型训练方法在扫描该文本片段时，仅仅从中提取各个词语，例如 < 任何，人，销售，许诺，形状，…，>，并将这些词语的词频用于构建公式 1 左侧的矩阵 C_{ij}。然而，根据以上结合图 2 描述的步骤 23，在提取出中心概念 A 的基础上（在文本片段 1 的情况下，A ＝“碱性锰电池”），可以针对该中心概念 A 构建如下的特征向量 V：

[0056] V ＝ (A，A 的内部词汇信息，A 的领域，A 的上位领域，A 的上位概念，A 的下位概念，A 的属性特征，A 的搭配统计信息，A 的上下文中的关键短语）。

[0057] 可以理解，尽管以上结合例子描述了可以用于构建特征向量的多种信息，并示例性地给出了特征向量 V 的表达形式，但是，本领域技术人员可以根据需要，选择以上信息中的一种或多种来构建特征向量。并且，本领域技术人员还可以在以上举例描述的信息的基础上进行进一步扩展、修改或组合，得到更多或其他信息用于构建特征向量。特征向量的表达形式、元素数目、元素类型均不限于以上的举例。如此构建的特征向量反映了中心概念的多维度的信息，进而更全面地反映了对应的文本片段所聚焦的实体的特点。

[0058] 以上描述了从文本片段提取中心概念的步骤 21，和针对中心概念构建特征向量的

步骤 23。通过反复执行步骤 21 和 23，可以从语料库文档的多个文本片段中提取出多个中心概念，并分别为该多个中心概念构建多个特征向量。在此基础上，可以执行步骤 25，基于所构建的至少一个特征向量，更典型地，基于多个特征向量的集合，训练主题模型。模型训练的过程可以采用多种已知的方法。

[0059] 在一个实施例中，采用聚类的方式进行主题模型的训练。具体地，依据向量之间的距离对获得的多个特征向量进行聚类，使得距离接近（例如低于某个距离阈值）的特征向量被聚类在一起。可以采用现有技术中的多种聚类算法来实现上述聚类过程，由此得到多个聚类。可以认为，通过上述方式得到的每个聚类对应于一个主题。

[0060] 在一个实施例中，可以将主题表示为所对应的聚类的中心。由于一个聚类由多个特征向量构成，在一个例子中，可以将聚类中所包含的多个特征向量映射到相应维度的向量空间中，进而利用已知的方式确定该多个特征向量在该向量空间中的中心“位置”，并用该中心“位置”所对应的向量来表征该聚类所对应的主题。由此，可以将主题表示为与特征向量维度相同的向量的形式，该向量也可以称为主题向量。可以理解，在其他实施方式中，还可以采用其他方式来计算或表达与聚类所对应的主题。

[0061] 在一个实施例中，采用矩阵计算的方式进行主题模型的训练。这与现有技术的训练方式相似。具体地，可以用获得的至少一个特征向量构成矩阵，其示出特征向量的各个元素在各个文档中的分布。将该矩阵作为训练的数据源，其作用类似于公式 (1) 中的矩阵 C_{ij}。采用现有技术中的各种推理方法，例如 pLSA 和 LDA 方式，可以类似地训练获得矩阵 Φ_{ik} 作为主题矩阵。通过该矩阵，可以同样地将主题表示为主题向量的形式。这与利用聚类方法的训练结果是一致的。

[0062] 本领域技术人员还可以采用其他方式基于特征向量的集合来训练主题模型。

[0063] 由于特征向量的构建考虑了中心概念在本体论中的信息，从而反映出中心概念所描述的实体的本质特点，因此，基于这样的特征向量所训练得到的主题模型可以更好地反映出主题与实体的本质关联。

[0064] 在训练了主题模型的基础上，可以利用训练的主题模型进行文档内容的比对。图 3 示出根据本发明一个实施例的比对文档内容的方法的流程图。如图 3 所示，该实施例中对比文档内容的方法包括以下步骤：步骤 31，利用主题模型，分别获取第一文档对应的第一主题集合和第二文档对应的第二主题集合，其中所述主题模型基于为概念构建的特征向量而训练，所述特征向量包含所述概念在本体论中的关联信息；步骤 33，比对所述第一主题集合和第二主题集合中的主题，以将相同主题对齐；以及步骤 35，对相同主题下第一文档中的第一文本片段集合和第二文档中的第二文本片段集合进行语义分析，以将第一文本片段集合和第二文本片段集合中语义相同的文本片段对齐。下面详细各个步骤的执行过程。

[0065] 首先，在步骤 31，利用主题模型，分别获取第一文档对应的第一主题集合和第二文档对应的第二主题集合。可以理解，该主题模型是根据图 2 的方法训练得到的主题模型，训练的基础是针对多个概念构建的多个特征向量，并且每个特征向量包含对应的概念在本体论中的关联信息。在如此训练得到的主题模型下，每个主题可以表示为特征向量的元素值的分布。下面结合第一文档，描述利用主题模型获取主题集合的过程。

[0066] 图 4 示出根据一个实施例获得第一主题集合的步骤。可以理解，为了便于对第一文档进行分析，可以将第一文档划分为多个文本片段。可以理解，上述文本片段可以是文档

中自然形成的段落或句子，也可以是人为划分的一段文字，或者是其他形式。在一个典型例子中，上述文本片段是文档中的一个句子。

[0067] 在此基础上，首先在步骤 41，从第一文档的文本片段中提取出中心概念。中心概念的提取可以利用现有的语言学的分析来实现。该步骤的具体执行过程类似于图 2 的步骤 21。

[0068] 接着，在步骤 43，针对提取的中心概念，构建特征向量，使得所述特征向量包含所述中心概念在本体论中的关联信息。具体而言地，与步骤 23 类似地，首先将上述中心概念映射到特定领域的本体树中，然后基于该本体树中的信息获取该中心概念在本体论中的关联信息。

[0069] 在一个实施例中，上述关联信息包括上述中心概念的领域信息。

[0070] 在一个实施例中，上述关联信息包括上述中心概念的类信息。具体地，上述类信息可以包括以下中的一项或者多项：上述中心概念在对应的本体树中的上位概念、下位概念、等价概念（如果有的话）。

[0071] 在一个实施例中，上述关联信息包括上述中心概念所对应的实体的属性特征信息。

[0072] 可选地，还可以获取上述中心概念的内部词汇信息，并将其包含在特征向量中。

[0073] 在一个实施例中，还可以获取与提取的中心概念有关的搭配统计信息作为特征向量的向量元素。

[0074] 在一个实施例中，还可以获取所提取的中心概念在文本片段中的上下文信息作为特征向量的向量元素。

[0075] 上述信息的获取方式和具体例子可以参照对图 2 的步骤 23 的描述。本领域技术人员可以根据需要，选择以上信息中的一种或多种来构建特征向量。并且，本领域技术人员还可以在以上举例描述的信息的基础上进行进一步扩展、修改或组合，得到更多或其他信息用于构建特征向量。不过，应理解的是，步骤 43 构建的特征向量是用于基于主题模型确定文本片段的主题，因此，该特征向量应该与训练主题模型所基于的特征向量在向量维度、元素上保持一致。也就是说，训练主题模型时采用什么方式构建特征向量，在利用主题模型确定主题时，也应该采用相同方式来构建特征向量。

[0076] 在构建了特征向量的基础上，在步骤 45，利用主题模型，基于所述特征向量确定文本片段的主题。可以理解，在已经训练得到主题模型的情况下，通过对特征向量进行与主题模型对应的计算，可以直接确定出文本片段的主题。在一个实施例中，主题模型中的主题表示为主题向量的形式。此时，可以将上述特征向量与主题模型下各个主题的主题向量进行比较，将匹配的主题确定为该文本片段对应的主题。可以理解，主题向量和特征向量具有相同的维度。因此，可以通过向量距离的计算和比较，确定出与构建的特征向量距离最短的主题向量。进而，将上述主题向量对应的主题确定为匹配的主题，也就是上述文本片段对应的主题。

[0077] 接着，在步骤 47，将上述主题添加到第一主题集合。

[0078] 通过针对第一文档中的各个文本片段反复执行步骤 41 到 47，可以确定出各个文本片段的主题，由此获得与第一文档对应的第一主题集合。

[0079] 以上结合第一文档描述了获取主题集合的方法。显然，该方法同样地适用于第二

文档。通过针对第二文档中的各个文本片段类似地执行步骤 41 到 47，可以获得与第二文档对应的第二主题集合。

[0080] 在分别获得了第一文档和第二文档的主题集合的基础上，在步骤 33，比对所述第一主题集合和第二主题集合中的主题，将相同主题对齐。可以理解，各个主题具有相应的主题标识或标签。一旦确定了某个文本片段的主题，就可以将为该文本片段添加相应的主题标签。相应地，第一主题集合包含了第一文档的各个文本片段的主题标签，第二主题集合包含了第二文档的各个文本片段的主题标签。通过比对这些主题标签，可以容易地确定两个主题集合中相同的主题，并将相同主题对齐。

[0081] 图 5A 示例性示出第一文档和第二文档的主题的对齐。在图 5A 的例子中，假定第一文档包含文本片段 S1，S2，S3，…Sn，第二文档包含文本片段 P1，P2，P3，…Pm。通过利用主题模型，第一文档中的文本片段 S1-S3 对应于主题 T1，S4 和 S5 对应于主题 T2，S6，S8 和 S10 对应于主题 T3，S7 和 S9 对应于主题 T4，等等。于是，第一主题集合包含 T1，T2，T3，T4 等主题。类似地，假定第二文档中的文本片段 P1-P3 对应于主题 T5，而 P4，P6 对应于主题 T1，P5，P7 和 P8 对应于主题 T3，P9-P11 对应于主题 T6，等等。通过比对主题标签，可以容易地确定出两个主题集合中的相同主题 T1 和 T3，并将相同主题进行对齐。

[0082] 接着，在步骤 35，对相同主题下第一文档中的第一文本片段集合和第二文档中的第二文本片段集合进行语义分析，以将第一文本片段集合和第二文本片段集合中语义相同的文本片段对齐。可以理解，由于文本片段与主题之间的对应关系，对于第一主题集合和第二主题集合中的某个相同主题，可以容易地获得第一文档中与该主题对应的第一文本片段集合，和第二文档中与该主题对应的的第二文本片段集合。例如，在图 5A 的例子中，第一主题集合和第二主题集合具有相同的主题 T1。在第一文档中，与该主题 T1 对应的文本片段包括 S1-S3，这些文本片段构成第一文本片段集合；在第二文档中，与主题 T1 对应的文本片段包括 P4 和 P6，这两个文本片段构成第二文本片段集合。

[0083] 对于上述第一文本片段集合和第二文本片段集合，可以分别对其进行语义分析和比较，以检测各个文本片段语义的差异。可以采用现有技术中的多种语义分析方法来执行上述过程。在一个实施例中，采用词语级别的语义分析来分析各个文本片段。在该分析过程中，主要考虑文本片段中出现的各个词语。在一个实施例中，对各个文本片段采用概念级别的对比，包括实体的对比，领域术语的对比等。在另一实施例中，还可以考虑文本片段中的概念在本体论中的相似度，以进一步确定文本片段在语义上的相似度。通过这样的语义分析，可以获取第一文本片段集合和第二文本片段集合中语义相同的文本片段，实现语义片段的对齐。

[0084] 图 5B 示例性示出图 5A 的例子中的文本片段的对齐。如前所述，在相同主题 T1 下，第一文档中的文本片段 S1-S3 构成第一文本片段集合，第二文档中的文本片段 P4 和 P6 构成第二文本片段集合。图 5B 具体示出了这些文本片段的内容。可以看到，尽管属于同样的主题，这些文本片段在语义上存在差异。通过语义分析可以确定，第一文档中的文本片段 S2 和第二文档中的文本片段 P4 具有相同的语义，因此，在步骤 35，可以将这两个文本片段对齐。

[0085] 通过以上描述可以看到，根据图 3 的方法，首先利用主题模型分别获得两个文档的主题，从而对齐相同主题；然后利用语义分析对齐同一主题下的文本片段，最终实现了两

个文档内容的对比。利用上述方法，即使两个文档具有完全不同的文档结构，采用了不同的术语体系，按照不同顺序进行了描述，仍然可以对两个文档的实质内容进行比对。

[0086] 图 3 的方法特别适用于第一文档和第二文档分别用于描述两个地区在同一领域的法律法规的情况。由于法律法规通常针对实体进行描述，因此，对于描述法律法规的文档，可以容易地获得其中各个概念的本体论信息，进而应用基于本体论信息的主题模型来实现主题对齐。例如，背景技术中描述了这样的场景：两篇文档分别描述了不同地区对于电池的使用和废弃的法律规定；用户希望通过对比这两篇文档，确定这两个地区对于电池的废弃的规定有什么不同。对于这样的场景，利用本发明实施例的方法，可以有效地实现主题的对齐和文本片段的对齐，使得用户可以容易地找出两个地区对同一问题进行规定的相应条款，进而确定出两个地区的规定有何不同。

[0087] 基于同一发明构思，本发明还提供了训练主题模型的装置，和比对文档内容的装置。

[0088] 图 6 示出根据本发明一个实施例的训练主题模型的装置的示例性框图。如图 6 所示，该训练主题模型的装置 600 包括：概念提取单元 61，配置为提取语料库文档中的文本片段的中心概念；向量构建单元 63，配置为，为所述中心概念构建特征向量，使得所述特征向量包含所述中心概念在本体论中的关联信息；以及训练单元 65，配置为基于所构建的至少一个特征向量，训练主题模型。

[0089] 在一个实施例中，上述向量构建单元 63 配置为，将上述中心概念映射到特定领域的本体树中，基于该本体树中的信息获取该中心概念在本体论中的关联信息。

[0090] 在一个实施例中，上述关联信息包括所述中心概念的类信息，所述类信息包括以下中的一项或者多项：上述中心概念在其映射到的本体树中的上位概念、下位概念和等价概念。

[0091] 在一个实施例中，上述关联信息包括以下中的一项或多项：上述中心概念的领域信息，以及上述中心概念所对应的实体的属性特征信息。

[0092] 根据一个实施例，上述向量构建单元 63 还配置为，获取以下信息中的至少一项作为特征向量的向量元素：与所述中心概念有关的搭配统计信息，以及所述中心概念在所述文本片段中的上下文信息。

[0093] 在一个实施例中，上述训练单元 65 配置为，采用向量聚类的方式训练主题模型，将主题模型下的主题表示为主题向量。

[0094] 图 7 示出根据本发明一个实施例的比对文档内容的装置的示例性框图。如图 7 所示，该实施例中比对文档内容的装置总体上表示为装置 700，并包括：主题获取单元 71，配置为利用主题模型，分别获取第一文档对应的第一主题集合和第二文档对应的第二主题集合，其中所述主题模型基于为概念构建的特征向量而训练，所述特征向量包含所述概念在本体论中的关联信息；主题比对单元 73，配置为比对所述第一主题集合和第二主题集合中的主题，以将相同主题对齐；以及文本片段分析单元 75，配置为对相同主题下第一文档中的第一文本片段集合和第二文档中的第二文本片段集合进行语义分析，以将第一文本片段集合和第二文本片段集合中语义相同的文本片段对齐。

[0095] 在一个实施例中，上述主题获取单元 71 包括（未示出）：概念提取模块，配置为从第一文档的文本片段中提取出中心概念；向量构建模块，配置为为所述中心概念构建特征

向量，使得所述特征向量包含所述中心概念在本体论中的关联信息；主题确定模块，配置为利用所述主题模型，基于所述特征向量确定所述文本片段对应的主题；以及主题添加模块，配置为将所述主题添加到第一主题集合。

[0096] 在一个实施例中，上述关联信息包括以下中的至少一项：上述中心概念的领域信息，上述中心概念的类信息，以及上述中心概念所对应的实体的属性特征信息。

[0097] 根据一个实施例，上述向量构建模块还配置为，获取以下信息中的至少一项作为特征向量的向量元素：与所述中心概念有关的搭配统计信息，以及所述中心概念在所述文本片段中的上下文信息。

[0098] 根据一个实施例，上述语义分析包括以下中的至少一项：词语层级的语义分析，概念层级的语义分析，以及基于文本片段中的概念在本体论中的相似度的语义分析。

[0099] 在一个实施例中，第一文档和第二文档分别是用于描述两个地区在同一领域的法律法规的文档。

[0100] 通过以上的方法和装置，可以训练得到更好地反映出主题与实体的语义关联的主题模型。利用这样的主题模型，可以确定出不同文档中的相同主题，进而对相同主题下的文本片段进行语义分析，实现文档本质内容的有效比对。

[0101] 本发明可以是系统、方法和/或计算机程序产品。计算机程序产品可以包括计算机可读存储介质，其上载有用于使处理器实现本发明的各个方面的计算机可读程序指令。

[0102] 计算机可读存储介质可以是可以保持和存储由指令执行设备使用的指令的有形设备。计算机可读存储介质例如可以是——但不限于——电存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或者上述的任意合适的组合。计算机可读存储介质的更具体的例子（非穷举的列表）包括：便携式计算机盘、硬盘、随机存取存储器（RAM）、只读存储器（ROM）、可擦式可编程只读存储器（EPROM或闪存）、静态随机存取存储器（SRAM）、便携式压缩盘只读存储器（CD-ROM）、数字多功能盘（DVD）、记忆棒、软盘、机械编码设备、例如其上存储有指令的打孔卡或凹槽内凸起结构、以及上述的任意合适的组合。这里所使用的计算机可读存储介质不被解释为瞬时信号本身，诸如无线电波或者其他自由传播的电磁波、通过波导或其他传输媒介传播的电磁波（例如，通过光纤电缆的光脉冲）、或者通过电线传输的电信号。

[0103] 这里所描述的计算机可读程序指令可以从计算机可读存储介质下载到各个计算/处理设备，或者通过网络、例如因特网、局域网、广域网和/或无线网下载到外部计算机或外部存储设备。网络可以包括铜传输电缆、光纤传输、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或者网络接口从网络接收计算机可读程序指令，并转发该计算机可读程序指令，以供存储在各个计算/处理设备中的计算机可读存储介质中。

[0104] 用于执行本发明操作的计算机程序指令可以是汇编指令、指令集架构（ISA）指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、或者以一种或多种编程语言的任意组合编写的源代码或目标代码，所述编程语言包括面向对象的编程语言—诸如Smalltalk、C++等，以及常规的过程式编程语言—诸如“C”语言或类似的编程语言。计算机可读程序指令可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机

或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络—包括局域网 (LAN) 或广域网 (WAN)—连接到用户计算机，或者，可以连接到外部计算机（例如利用因特网服务提供商来通过因特网连接）。在一些实施例中，通过利用计算机可读程序指令的状态信息来个性化定制电子电路，例如可编程逻辑电路、现场可编程门阵列 (FPGA) 或可编程逻辑阵列 (PLA)，该电子电路可以执行计算机可读程序指令，从而实现本发明的各个方面。

[0105] 这里参照根据本发明实施例的方法、装置（系统）和计算机程序产品的流程图和/或框图描述了本发明的各个方面。应当理解，流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合，都可以由计算机可读程序指令实现。

[0106] 这些计算机可读程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器，从而生产出一种机器，使得这些指令在通过计算机或其它可编程数据处理装置的处理器执行时，产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中，这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作，从而，存储有指令的计算机可读介质则包括一个制造品，其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

[0107] 也可以把计算机可读程序指令加载到计算机、其它可编程数据处理装置、或其它设备上，使得在计算机、其它可编程数据处理装置或其它设备上执行一系列操作步骤，以产生计算机实现的过程，从而使得在计算机、其它可编程数据处理装置、或其它设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

[0108] 附图中的流程图和框图显示了根据本发明的多个实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段或指令的一部分，所述模块、程序段或指令的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个连续的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或动作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

[0109] 以上已经描述了本发明的各实施例，上述说明是示例性的，并非穷尽性的，并且也不限于所披露的各实施例。在不偏离所说明的各实施例的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择，旨在最好地解释各实施例的原理、实际应用或对市场中的技术的技术改进，或者使本技术领域的其它普通技术人员能理解本文披露的各实施例。

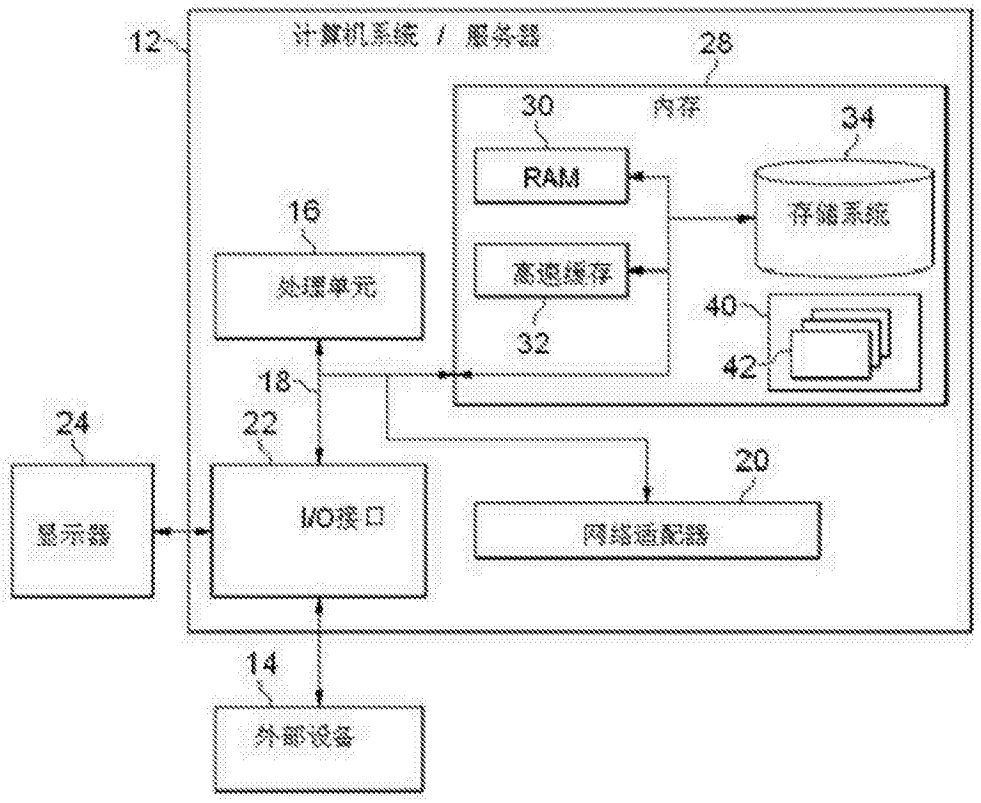

12
计算机系统 / 服务器
28
30
内存
34
RAM
存储系统
16
处理单元
高速缓存
32
40
42
18
24
22
20
显示器
I/O接口
网络适配器
14
外部设备

图 1

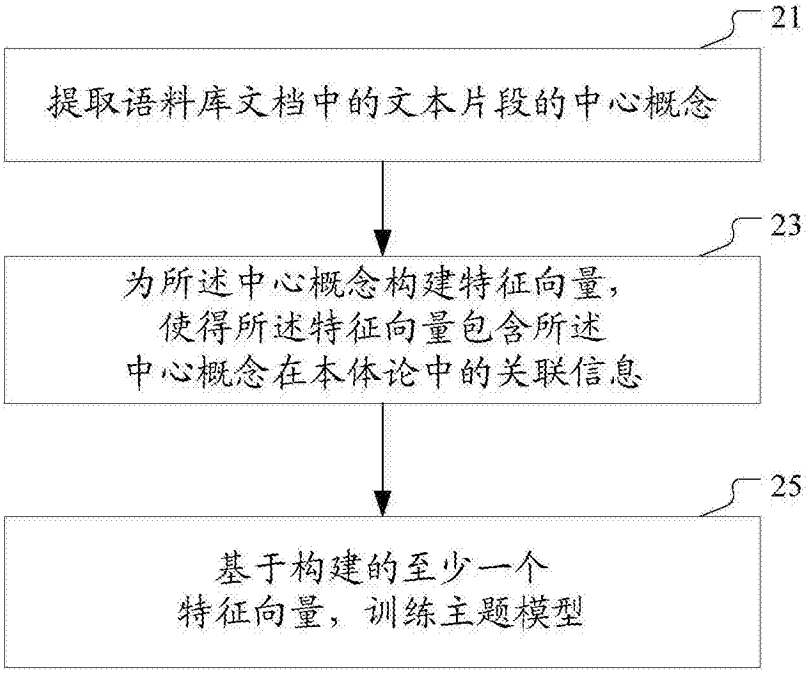

21
提取语料库文档中的文本片段的中心概念
23
为所述中心概念构建特征向量，
使得所述特征向量包含所述
中心概念在本体论中的关联信息
25
基于构建的至少一个
特征向量，训练主题模型

图 2

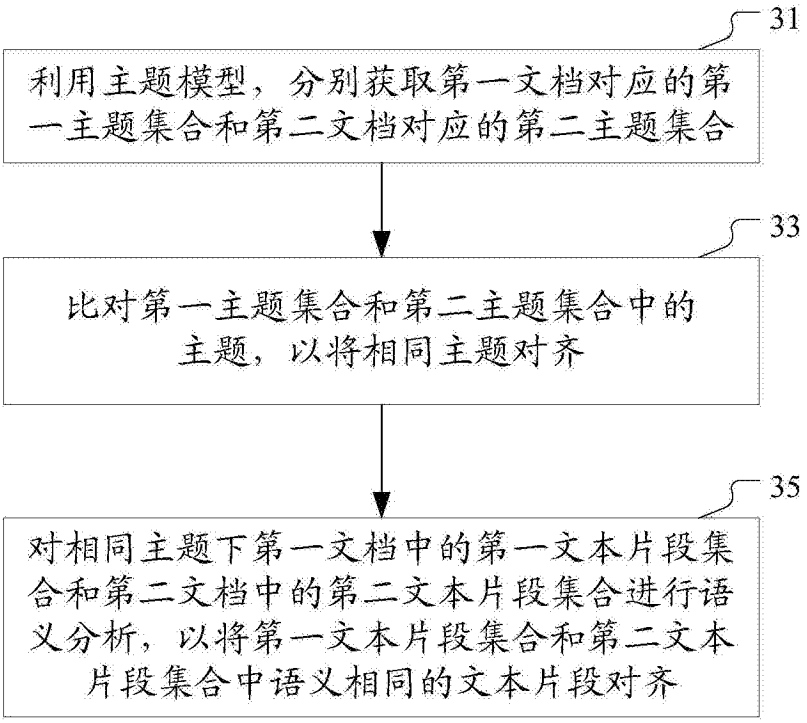

31
利用主题模型，分别获取第一文档对应的第一主题集合和第二文档对应的第二主题集合
33
比对第一主题集合和第二主题集合中的主题，以将相同主题对齐
35
对相同主题下第一文档中的第一文本片段集合和第二文档中的第二文本片段集合进行语义分析，以将第一文本片段集合和第二文本片段集合中语义相同的文本片段对齐

图 3

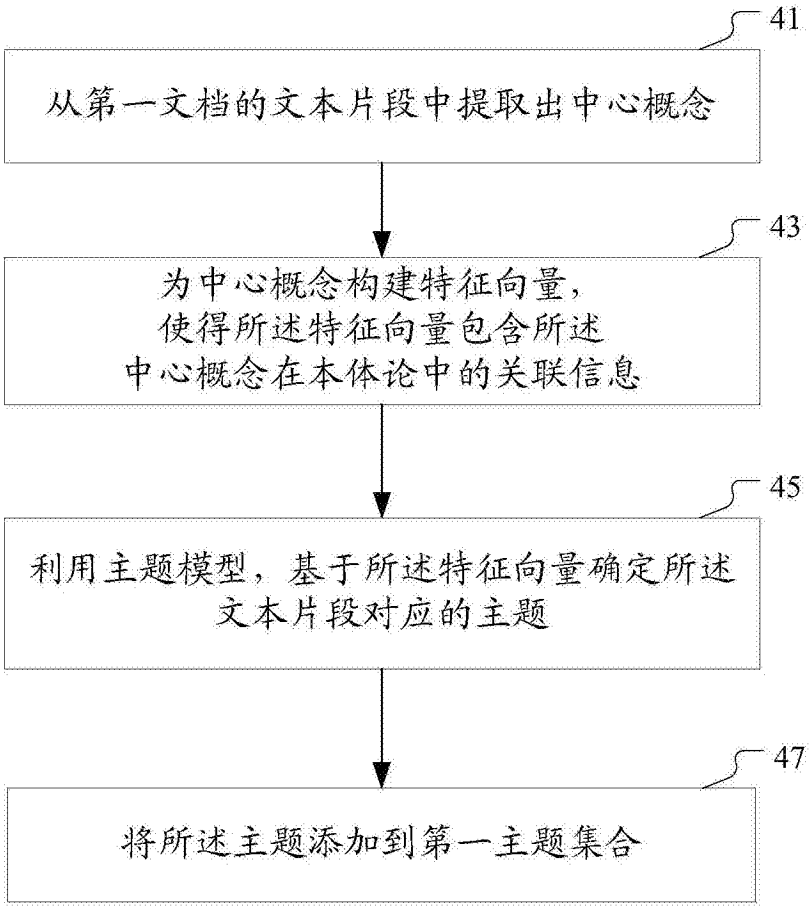

41
从第一文档的文本片段中提取出中心概念
43
为中心概念构建特征向量，
使得所述特征向量包含所述
中心概念在本体论中的关联信息
45
利用主题模型，基于所述特征向量确定所述
文本片段对应的主题
47
将所述主题添加到第一主题集合

图 4

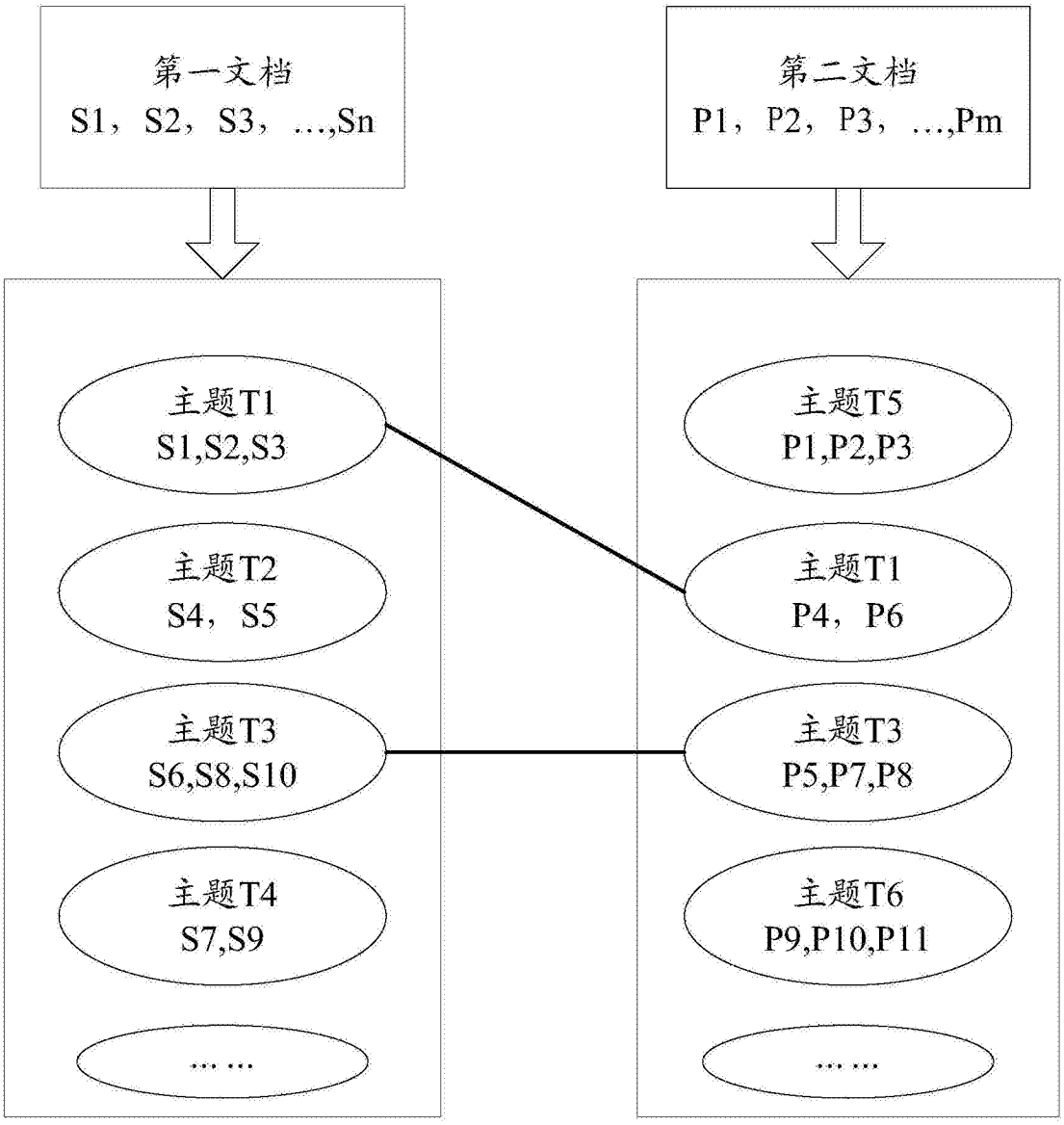

第一文档
S1，S2，S3，…,Sn
第二文档
P1，P2，P3，…,Pm
主题T1
S1,S2,S3
主题T2
S4，S5
主题T3
S6,S8,S10
主题T4
S7,S9
……
主题T5
P1,P2,P3
主题T1
P4，P6
主题T3
P5,P7,P8
主题T6
P9,P10,P11
……

图 5A

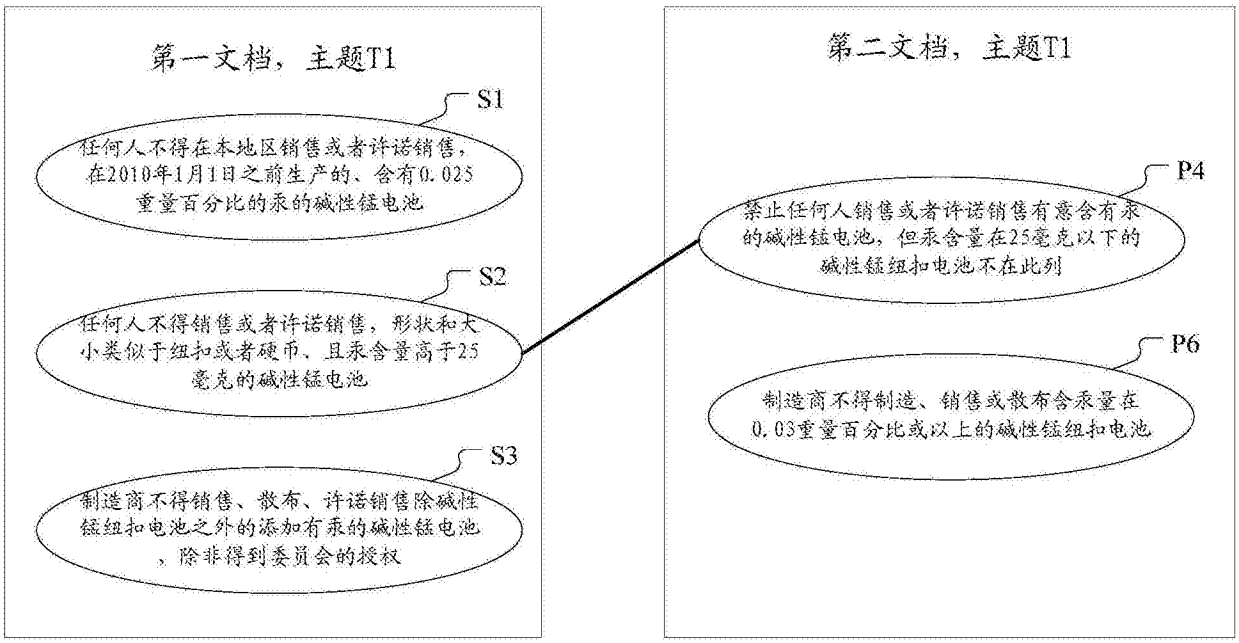

第一文档，主题T1
S1
任何人不得在本地区销售或者许诺销售，在2010年1月1日之前生产的、含有0.025重量百分比的汞的碱性锰电池
S2
任何人不得销售或者许诺销售，形状和大小类似于纽扣或者硬币、且汞含量高于25毫克的碱性锰电池
S3
制造商不得销售、散布、许诺销售除碱性锰纽扣电池之外的添加有汞的碱性锰电池，除非得到委员会的授权
第二文档，主题T1
P4
禁止任何人销售或者许诺销售有意含有汞的碱性锰电池，但汞含量在25毫克以下的碱性锰纽扣电池不在此列
P6
制造商不得制造、销售或散布含汞量在0.03重量百分比或以上的碱性锰纽扣电池

图 5B

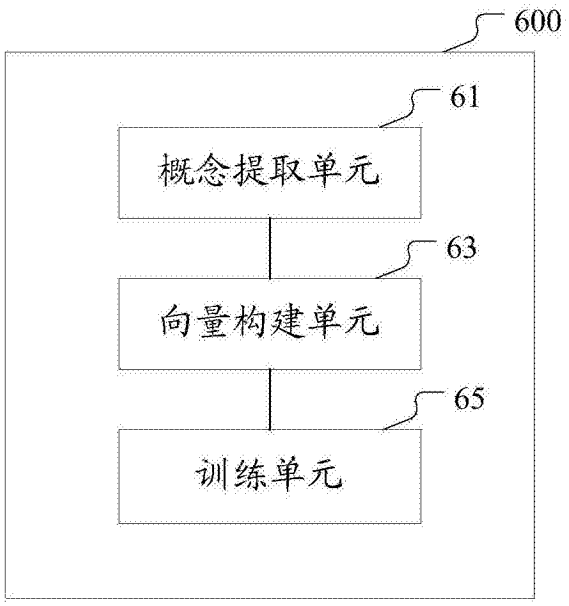

600
61
概念提取单元
63
向量构建单元
65
训练单元

图 6

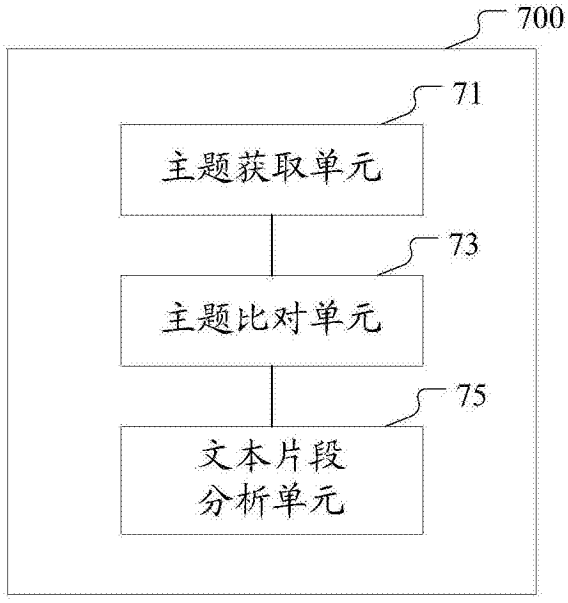

700
71
主题获取单元
73
主题比对单元
75
文本片段
分析单元

图 7