(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115374241 A
(43)申请公布日 2022.11.22

(21)申请号 202110540883.5

(22)申请日 2021.05.18

(71)申请人 京东科技控股股份有限公司
地址 100176 北京市大兴区北京经济技术开发区科创十一街18号C座2层221室

(72)发明人 孙宏超 陈蒙

(74)专利代理机构 北京路浩知识产权代理有限公司 11002
专利代理师 王毅

(51)Int.Cl.
G06F 16/33(2019.01)
G06F 16/335(2019.01)
G06F 16/35(2019.01)
G06F 40/194(2020.01)
G06F 40/30(2020.01)

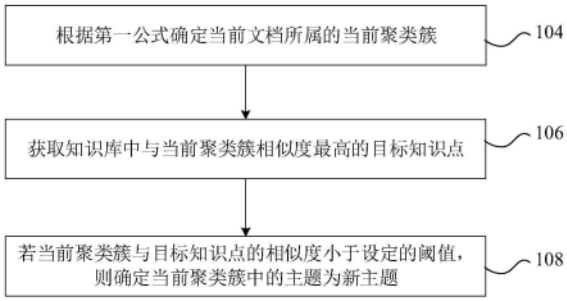
权利要求书2页 说明书12页 附图4页

(54)发明名称

基于语义的文档处理方法、装置、电子设备和存储介质

(57)摘要

本发明提供一种基于语义的文档处理方法、装置、电子设备和非暂态计算机可读存储介质，文档处理方法包括：根据第一公式确定所述当前文档所属的当前聚类簇，其中，所述第一公式根据预设的主题聚类模型、所述当前文档与聚类簇的单词的词频和语义权重得到，所述主题聚类模型采用狄利克雷过程混合模型和中国餐馆过程对文档聚类过程进行建模得到；获取知识库中与所述当前聚类簇相似度最高的目标知识点；若所述当前聚类簇与所述目标知识点的相似度小于设定的阈值，则确定所述当前聚类簇中的主题为新主题。本发明的技术方案通过引入语义信息来进行文档聚类分析，可以强化聚类能力，优化主题挖掘效果。



CN 115374241 A

1.一种基于语义的文档处理方法,其特征在于,所述文档处理方法包括:

根据第一公式确定所述当前文档所属的当前聚类簇,其中,所述第一公式根据预设的主题聚类模型、所述当前文档与聚类簇的单词的词频和语义权重得到,所述主题聚类模型采用狄利克雷过程混合模型和中国餐馆过程对文档聚类过程进行建模得到,所述主题聚类模型用于确定文档中的单词属于一个聚类簇的概率;

获取知识库中与所述当前聚类簇相似度最高的目标知识点;

若所述当前聚类簇与所述目标知识点的相似度小于设定的阈值,则确定所述当前聚类簇中的主题为新主题。

2.根据权利要求1所述的文档处理方法,其特征在于,所述根据第一公式确定所述当前文档所属的当前聚类簇,包括:

根据所述主题聚类模型、所述聚类簇的完整性、所述当前文档与聚类簇的同质性函数和所述当前文档与所述聚类簇的单词的语义权重确定所述第一公式;

根据所述第一公式获取所述当前文档属于所述聚类簇的概率;

根据所述概率确定所述当前文档所属的当前聚类簇。

3.根据权利要求2所述的文档处理方法,其特征在于,所述第一公式中,所述当前文档属于所述聚类簇的概率

$p(z_x = z \mid \vec{z},\ \vec{x},\ \alpha_0,\ \beta)$:

$$p(z_x = z \mid \vec{z},\ \vec{x},\ \alpha_0,\ \beta) = \left(\frac{m_z}{D-1+\alpha_0 D}\right) * \left(\frac{\prod_{w\in x}\prod_{j=1}^{N_x^w}(n_z^w \cdot ICF_w)+\beta+j-1}{\prod_{i=1}^{N_x} n_z + V\beta + i - 1}\right) * \left(1 + \sum_{w_i\in x \wedge w_j \in x} cw_{ij}\right)$$

其中,α_0为设定的参数,i和j均为自然数,mz表示聚类簇中包含的文档数量,D表示所有聚类簇中文档的数量,N_x^w表示在文档x中单词w的词频,N_x表示文档x的总词数,n_z^w表示在聚类簇z中单词w的词频,n_z表示聚类簇z的总词数,V表示词典大小,ICF_w表示单词w的在各个聚类簇之间的重要程度,其计算公式为$ICF(w \in x) = \log\left(\frac{|Z|}{|w\in Z|}\right)$,语义权重$cw_{ij}$的公式为:

$$cw_{ij} = \frac{\sum_{x'\subseteq z} n_{x'}^{w_i}}{\sum_{x'\subseteq z} n_{x'}^{w_i} + \sum_{x'\subseteq z} n_{x'}^{w_j}} (w_i,\ w_j) \in x'$$

。

4.根据权利要求1所述的文档处理方法,其特征在于,所述获取知识库中与所述当前聚类簇相似度最高的目标知识点之后,所述文档处理方法还包括:

若所述当前聚类簇与所述目标知识点的相似度大于等于所述阈值,则确定所述当前聚类簇属于所述目标知识点。

5.根据权利要求1所述的文档处理方法,其特征在于,所述确定所述当前聚类簇中的主题为新主题之后,所述文档处理方法还包括:

通过提取关键词确定所述新主题的类目。

6.根据权利要求1所述的文档处理方法，其特征在于，所述获取知识库中与所述当前聚类簇相似度最高的目标知识点之前，所述文档处理方法还包括：

通过欧式距离公式、余弦相似度公式或Jaccard相似系数计算所述当前聚类簇和其它聚类簇之间的相似度。

7.根据权利要求1所述的文档处理方法，其特征在于，所述采用狄利克雷过程混合模型和中国餐馆过程对文档聚类过程进行建模，包括：

根据参数为α_0的狄利克雷分布生成一个主题分布φ_i；

根据参数为β的狄利克雷分布生成一个单词分布θ_k；

根据参数为φ_i的多项分布，生成所述当前文档中单词的主题z_i；

根据参数为θ_{z_i}的多项分布，生成当前文档的单词$x_{i,j}$。

8.一种基于语义的文档处理装置，其特征在于，所述文档处理装置包括：

当前聚类簇确定单元，用于根据第一公式确定所述当前文档所属的当前聚类簇，其中，所述第一公式根据预设的主题聚类模型、所述当前文档与聚类簇的单词的词频和语义权重得到，所述主题聚类模型采用狄利克雷过程混合模型和中国餐馆过程对文档聚类过程进行建模得到，所述主题聚类模型用于确定文档中的单词属于一个聚类簇的概率；

目标知识点获取单元，用于获取知识库中与所述当前聚类簇相似度最高的目标知识点；

新主题确定单元，用于在若所述当前聚类簇与所述目标知识点的相似度小于设定的阈值时，确定所述当前聚类簇中的主题为新主题。

9.一种电子设备，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述程序时实现如权利要求1至7任一项所述的文档处理方法的步骤。

10.一种非暂态计算机可读存储介质，其上存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现如权利要求1至7任一项所述的文档处理方法的步骤。

基于语义的文档处理方法、装置、电子设备和存储介质

技术领域

[0001] 本发明涉及数据处理技术领域，尤其涉及一种基于语义的文档处理方法、装置、电子设备和非暂态计算机可读存储介质。

背景技术

[0002] 零售场景下的对话机器人较多根据人工定义的知识体系来完成应答，但在实际应用中不可避免地会出现机器人应答不上的Query(问题)。使用Query聚类的方式可以对这些兜底Query进行分析处理，Query聚类指的是对机器人无法应答的Query，即转人工或者兜底应答的Query，进行自动聚类。Query聚类的目的是发现新的知识点或意图，从而缓解运营的压力，方便运营人员新增知识点。

[0003] Query聚类属于文本聚类，又称文档聚类。文本聚类在工业界属于一种使用较多的自然语言处理工具，是研究分类问题的一种统计分析方法，主要应用于数据挖掘和信息检索等方面。其具有无监督的优势，从而在大量实际业务场景中广泛应用。文本聚类等聚类分析可以作为一个单独的工具以发现数据中分布的一些深层的信息，并且概括出每一类的特点，或者把注意力放在某一个特定的类上以作进一步的分析；并且，聚类分析也可以作为数据挖掘算法中其他分析算法的一个预处理步骤。

[0004] 在相关技术中，存在按照字面信息聚类，但是语义信息不一致的情况，例如实际业务场景中会遇到把货值是什么意思，盘货是什么意思，导购是什么意思，换签是什么意等问题聚类到一起的情况，这些问题在字面上是相似的，但是语义上不适合聚类到一起。

[0005] 上下文的语义关系会影响文本聚类的能力，导致对文本的主题挖掘的效果较差。

发明内容

[0006] 本发明提供一种基于语义的文档处理方法、装置、电子设备和非暂态计算机可读存储介质，用以解决现有技术中聚类能力较差的缺陷，实现聚类能力的强化以及主题挖掘效果的优化。

[0007] 本发明提供一种基于语义的文档处理方法，包括：根据第一公式确定所述当前文档所属的当前聚类簇，其中，所述第一公式根据预设的主题聚类模型、所述当前文档与聚类簇的单词的词频和语义权重得到，所述主题聚类模型采用狄利克雷过程混合模型和中国餐馆过程对文档聚类过程进行建模得到，所述主题聚类模型用于确定文档中的单词属于一个聚类簇的概率；获取知识库中与所述当前聚类簇相似度最高的目标知识点；若所述当前聚类簇与所述目标知识点的相似度小于设定的阈值，则确定所述当前聚类簇中的主题为新主题。

[0008] 根据本发明提供的一种基于语义的文档处理方法，，所述根据第一公式确定所述当前文档所属的当前聚类簇，包括：根据所述主题聚类模型、所述聚类簇的完整性、所述当前文档与聚类簇的同质性函数和所述当前文档与所述聚类簇的单词的语义权重确定所述第一公式；根据所述第一公式获取所述当前文档属于所述聚类簇的概率；根据所述概率确

定所述当前文档所属的当前聚类簇。

[0009] 根据本发明提供的一种基于语义的文档处理方法，所述第一公式中，所述当前文档属于所述聚类簇的概率$p(z_x = z \mid \vec{z},\ \vec{x},\ \alpha_0,\ \beta)$：

[0010]
$$p(z_x = z \mid \vec{z},\ \vec{x},\ \alpha_0,\ \beta) = \left(\frac{m_z}{D-1+\alpha_0 D}\right) * \left(\frac{\prod_{w\in x}\prod_{j=1}^{N_x^w}(n_z^w \cdot ICF_w)+\beta+j-1}{\prod_{i=1}^{N_x} n_z + V\beta + i - 1}\right) * \left(1 + \sum_{w_i\in x \wedge w_j \in x} cw_{ij}\right)$$

[0011] 其中，α_0为设定的参数，i和j均为自然数，m_z表示聚类簇中包含的文档数量，D表示所有聚类簇中文档的数量，N_x^w表示在文档x中单词w的词频，N_x表示文档x的总词数，n_z^w表示在聚类簇z中单词w的词频，n_z表示聚类簇z的总词数，V表示词典大小，ICF_w表示单词w的在各个聚类簇之间的重要程度，其计算公式为$ICF(w \in x) = \log\left(\frac{|Z|}{|w\in Z|}\right)$；其中，语义权重$cw_{ij}$的公式为：

[0012]
$$cw_{ij} = \frac{\sum_{x'\subseteq z} n_{x'}^{w_i}}{\sum_{x'\subseteq z} n_{x'}^{w_i} + \sum_{x'\subseteq z} n_{x'}^{w_j}} (w_i,\ w_j) \in x'$$

。

[0013] 根据本发明提供的一种基于语义的文档处理方法，所述获取知识库中与所述当前聚类簇相似度最高的目标知识点之后，所述文档处理方法还包括：若所述当前聚类簇与所述目标知识点的相似度大于等于所述阈值，则确定所述当前聚类簇属于所述目标知识点。

[0014] 根据本发明提供的一种基于语义的文档处理方法，所述确定所述当前聚类簇中的主题为新主题之后，所述文档处理方法还包括：通过提取关键词确定所述新主题的类目。

[0015] 所述获取知识库中与所述当前聚类簇相似度最高的目标知识点之前，所述文档处理方法还包括：通过欧式距离公式、余弦相似度公式或Jaccard相似系数计算所述当前聚类簇和其它聚类簇之间的相似度。

[0016] 根据本发明提供的一种基于语义的文档处理方法，所述采用狄利克雷过程混合模型和中国餐馆过程对文档聚类过程进行建模，包括：根据参数为α_0的狄利克雷分布生成一个主题分布φ_i；根据参数为β的狄利克雷分布生成一个单词分布θ_k；根据参数为φ_i的多项分布，生成所述当前文档中单词的主题z_i；根据参数为θ_{z_i}的多项分布，生成当前文档的单词$x_{i,j}$。

[0017] 本发明还提供一种基于语义的文档处理装置，包括：当前聚类簇确定单元，用于根据第一公式确定所述当前文档所属的当前聚类簇，其中，所述第一公式根据预设的主题聚类模型、所述当前文档与聚类簇的单词的词频和语义权重得到，所述主题聚类模型采用狄利克雷过程混合模型和中国餐馆过程对文档聚类过程进行建模得到，所述主题聚类模型用于确定文档中的单词属于一个聚类簇的概率；目标知识点获取单元，用于获取知识库中与所述当前聚类簇相似度最高的目标知识点；新主题确定单元，用于在若所述当前聚类簇与

所述目标知识点的相似度小于设定的阈值时，确定所述当前聚类簇中的主题为新主题。

[0018] 本发明还提供一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时实现如上述任一种所述基于语义的文档处理方法的步骤。

[0019] 本发明还提供一种非暂态计算机可读存储介质，其上存储有计算机程序，该计算机程序被处理器执行时实现如上述任一种所述基于语义的文档处理方法的步骤。

[0020] 本发明提供的基于语义的文档处理方法、装置、电子设备和非暂态计算机可读存储介质，通过增强的狄利克雷过程混合模型对文档聚类过程进行建模，并结合文档和聚类簇的单词的词频和语义权重判断当前文档所属的当前聚类簇中的主题是否新主题，增强了模型的聚类能力，实现了结合语义对文档聚类过程进行建模和主题挖掘，具有较好的主题挖掘效果。

附图说明

[0021] 为了更清楚地说明本发明或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作一简单地介绍，显而易见地，下面描述中的附图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的附图。

[0022] 图1是本发明提供的基于语义的文档处理方法的流程示意图之一；

[0023] 图2是本发明提供的狄利克雷过程混合模型的构造过程的示意图；

[0024] 图3是本发明提供的中国餐馆过程的构造过程的示意图；

[0025] 图4是本发明提供的主题聚类模型的建模方法的流程示意图；

[0026] 图5是本发明提供的文档聚类的概率的获取过程的示意图；

[0027] 图6是本发明提供的基于语义的文档处理方法的流程示意图之二；

[0028] 图7是本发明提供的基于语义的文档处理装置的结构示意图；

[0029] 图8是本发明提供的电子设备的结构示意图。

具体实施方式

[0030] 为使本发明的目的、技术方案和优点更加清楚，下面将结合本发明中的附图，对本发明中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0031] 在本发明一个或多个实施例中使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本发明一个或多个实施例。在本发明一个或多个实施例和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式，除非上下文清楚地表示其他含义。还应当理解，本发明一个或多个实施例中使用的术语“和/或”是指并包含一个或多个相关联的列出项目的任何或所有可能组合。

[0032] 应当理解，尽管在本发明一个或多个实施例中可能采用术语第一、第二等来描述各种信息，但这些信息不应限于这些术语。这些术语仅用来将同一类型的信息彼此区分开。例如，在不脱离本发明一个或多个实施例范围的情况下，第一也可以被称为第二，类似地，

第二也可以被称为第一。取决于语境，如在此所使用的词语“如果”可以被解释成为“在.....时”或“当.....时”或“响应于确定”。

[0033] 在相关技术中，聚类分析在算法分类上可以分为基于向量空间的文本聚类方法和基于主题模型的文本聚类方法。

[0034] 基于向量空间的文本聚类方法将文本表示为多维空间中的一个点，通过衡量相似性的方式，比较在一个聚类中的模式之间是否比不在同一聚类中的模式之间具有更多的相似性。基于向量空间的聚类分析的算法可以分为划分法(K-Means)、层次法(HAC)、基于密度的方法(DBSCAN)、基于图谱的方法(AP和Spectral)。在实际操作时，基于向量的聚类分析方法一般使用TF-IDF或者BERT等方式进行文本信息表示，再使用余弦相似或者其他向量相似方法计算距离信息，并根据一定聚类逻辑进行聚类操作。

[0035] 基于主题模型的方法包括隐狄利克雷分布(Latent Dirichlet Allocation)，动态主题模型(Dynamic topic models)，概率潜在语义分析(Probabilistic latent semantic analysis)。在实际应用时，基于主题模型的聚类分析方法一般通过词袋的方式，将每一个文档看作一个词频向量，通过此方式完成文本信息的量化，再基于多项分布与狄利克雷分布的共轭关系，构建了词与主题，主题与文档两个概率过程，并通过吉布斯采样的方式进行参数优化。

[0036] 在相关技术中，存在按照字面信息聚类，但是语义信息不一致的情况，例如实际业务场景中会遇到把货值是什么意思，盘货是什么意思，导购是什么意思，换签是什么意等问题聚类到一起的情况，这些问题在字面上是相似的，但是语义上不适合聚类到一起。

[0037] 基于向量空间的聚类分析中信息表示上使用TF-IDF表示后聚类无法挖掘语义信息，使用BERT等方式进行语义表示太耗费机器资源，对流式文本处理延时较高。并且，现有的聚类方法大多需要指定聚类簇的数量，不能较好的产生自适应簇的数量。

[0038] 基于主题模型的聚类分析中使用的词袋方式进行的文本信息表示，没有考虑文本的上下文的语义关系，进而会影响文本聚类的能力，导致对文本的主题挖掘的效果较差。

[0039] 为解决该问题，本发明实施例提供一种基于语义的文档处理方案。下面结合图1至图8对本发明示例实施方式进行详细说明。

[0040] 如图1所示的是本发明实施例的基于语义的文档处理方法的流程图。本发明实施例提供的方法可以由任意具备计算机处理能力的电子设备执行，例如终端设备和/或服务器。如图1所示，该基于语义的文档处理方法包括：

[0041] 步骤104，根据第一公式确定当前文档所属的当前聚类簇，其中，第一公式根据预设的主题聚类模型、当前文档与聚类簇的单词的词频和语义权重得到，主题聚类模型采用狄利克雷过程混合模型和中国餐馆过程对文档聚类过程进行建模得到，主题聚类模型用于确定文档中的单词属于一个聚类簇的概率。

[0042] 具体地，聚类簇是聚类分析得到的数据对象的集合，这些对象与同一个聚类簇中的对象彼此相似即具有同质性，与其它聚类簇中的对象相异。基于主题聚类模型得到的聚类簇为基于主题的聚类簇，每个聚类簇对应一个主题。语义权重指的是单词的语义的重要程度。

[0043] 狄利克雷过程混合模型(Dirichlet Process Mixture Model，简称DPMM)是一种非参数的贝叶斯模型。中国餐馆过程是一种时间离散的(discrete-time)随机过程

(stochastic process),类似于中国餐馆中坐在餐桌旁的顾客。当前文档即从实际业务场景中获取的未命中问题的或者经过质检挖掘的人工日志的文本,也是当前需要进行聚类分析的文本。文档聚类分析是一种一种使用较多的自然语言处理工具,是研究分类问题的一种统计分析方法,主要应用于数据挖掘和信息检索等方面。主题聚类模型是基于狄利克雷过程混合模型和中国餐馆过程对文档聚类过程进行建模得到的,是一种以文档的主题为聚类目标将不同的文档进行聚类的聚类模型。

[0044] 例如,对于一个文档“没有及时发出”,可以得到该文档的当前当前聚类簇,该当前聚类簇的主题为“发出”。

[0045] 步骤106,获取知识库中与当前聚类簇相似度最高的目标知识点。

[0046] 具体地,相似度指的不同的聚类簇的相似性。知识库是存储已经存在的所有知识点的数据库。一般通过计算聚类簇对应的主题之间的距离来计算二者的相似度。如果距离小,那么相似度大;如果距离大,那么相似度小。

[0047] 步骤108,若当前聚类簇与目标知识点的相似度小于设定的阈值,则确定当前聚类簇中的主题为新主题。

[0048] 具体地,当前聚类簇是当前文档所属的聚类簇,若当前聚类簇与目标知识点的相似度大于等于设定的阈值,则当前聚类簇与目标知识点的相似度较高,可以认为当前聚类簇中的主题与目标知识点的主题属于同一主题。反之,若当前聚类簇与目标知识点的相似度较低,则可以认为当前聚类簇中的主题与目标知识点的主题不同,当前聚类簇中的主题为新主题。

[0049] 例如,对文档“没有及时发出”来说,目标知识点的主题为“发货”,当前聚类簇与目标知识点的相似度为80%,大于设定的阈值75%,则认为当前聚类簇中的主题与目标知识点的主题属于同一主题。

[0050] 本发明实施例中的技术方案中,基于主题聚类模型,根据文档与聚类簇之间的单词的词频和语义权重判断确定当前所属的当前聚类簇,通过词的共现关系来引入语义信息,强化了模型的聚类能力,进而在对未命中Query进行聚类分析时可以更好地挖掘到新意图、新知识点,从而缓解运营的压力,方便运营人员新增知识点。

[0051] 在本发明实施例中的狄利克雷过程混合模型可以理解为一种聚类方法,但是不需要指定类别数量,它可以从数据中推断聚类簇的数量。

[0052] 在狄利克雷过程混合模型中,包括以下定义内容:

[0053] $G\sim DP(\alpha_0,G_0)$,

[0054] $\theta_i|G\sim G$,

[0055] $x_i|\theta_i\sim F(\theta_i)$。

[0056] 其中,i为自然数,$F(\theta_i)$是一个混合分布,包含若干个单独的分布。$F(\theta_i)$中的数据包括$x_1,x_2,\cdots,x_n$,其中$n>2$且n为自然数。这些数据可以是多元的、相互独立的,来自于一些未知的分布中。其值可以是实数型的,也可以是类别型的。

[0057] G为由狄利克雷过程产生的测度空间上的概率分布,θ_i为混合分布中某个分布的参数,$x_i|\theta_i\sim F(\theta_i)$为根据参数$\theta_i$产生的某个分布的观测值,如果$x_i$和$x_j$来自同一分布,则$\theta_i=\theta_j$,其中,j为自然数。

[0058] 如图2所示,在构造狄利克雷过程混合模型时,根据参数为α_0和G_0的狄利克雷过程

生成测度空间Θ上的概率分布G，根据参数为G得到混合分布中某个分布的参数θ_i，根据参数θ_i产生某个分布的观测值x_i。

[0059] $F(\theta_i)$中的参数θ_i来自于狄利克雷过程(Dirichlet Process，简称DP)G，即该分布的参数的先验是一个狄利克雷过程。其集中参数为α_0，基分布为G_0。

[0060] 对于空间上的任意一个有限分割，如果空间上的一个随机概率分布在这个分割中各部分上的测度服从一个狄利克雷分布，那么我们就称随机概率分布服从狄利克雷过程。

[0061] 狄利克雷过程是一种非参数的贝叶斯模型的随机过程，其定义如下：假设G_0是测度空间Θ上的随机概率分布，参数α_0是正实数，对于测度空间上的任意有限划分$A_1,\ldots,A_r$，其中，r为自然数，如果存在$(G(A_1),\ldots,G(A_r))\sim Dir(\alpha_0G_0(A_1),\ldots,\alpha_0G_0(A_r))$，则认为G服从由基分布$G_0$和参数$\alpha_0$构成的狄利克雷过程，可以表示为$G\sim DP(\alpha_0,G_0)$。

[0062] 因为狄利克雷分布和多项分布具有共轭的特性，所以当测度空间的划分满足多项分布时，即$\theta_1,\ldots,\theta_n\sim Mult$，则狄利克雷过程对测度空间有限划分的后验分布也服从狄利克雷过程，即$(G(A_1),\ldots,G(A_r))|\theta_1,\ldots,\theta_n\sim Dir(\alpha_0G_0(A_1)+n_1,\ldots,\alpha_0G_0(A_r)+n_r)$，

[0063] 其中，公式中的$n_1,\ldots,n_r$表示的是观测值属于A_k的数目。

[0064] 多项分布是指单次试验中的随机变量的取值不再是0-1的，而是有多种离散值可能(1,2,3...,k)。比如投掷6个面的骰子实验，n次实验结果服从k＝6的多项分布。这里，n和k均为自然数。

[0065] 在贝叶斯概率理论中，如果后验概率和先验概率满足同样的分布律，那么，先验分布和后验分布被叫做共轭分布，同时，先验分布叫做似然函数的共轭先验分布。在观测到的数据符合多项分布，参数的先验分布和后验分布都是狄利克雷分布的情况，就是狄利克雷-多项分布共轭。

[0066] 狄利克雷过程混合模型的数据是来自于一个混合分布，即n个数据是来自于k个分布，其中k为自然数，且$k\leqslant n$。因此，对于来自同一个分布的观测值，其分布θ_i的参数应该是一样的。但，如果我们的先验G是连续分布，那么，抽样结果不可能出现相同的值，而由于狄利克雷过程的特性，我们可以抽取相同的参数θ且使得其与基分布相似。同时，由于该模型数据增长速度远大于混合模型个数的增长，因此，对于一定的样本数据，我们可以得到远小于样本的类别个数，因此该模型可以作为非参数聚类方法。

[0067] 狄利克雷过程定义中的采样方式可以采用中国餐馆过程CRP(Chinese Restaurant Process)进行建模，其建模过程如下：

[0068] 如图3所示，假设有个餐馆，可以容纳无限张桌子，每张桌子的参数为$\varphi_1,\ \ldots,\ \varphi_k$，分配桌子给每一个顾客$z_1,\ldots,z_n$，$z_i$表示分配给第i位顾客的桌子，每位新到的顾客可以开一张新的桌子，其概率为$\frac{\alpha_0}{(1+\alpha_0)}$，或者坐到已经有顾客的桌子上，其概率为$\frac{n_k}{\alpha_0+n}$，$n_k$为第k张桌子上已有的顾客的数量。

[0069] CRP过程中只能实现相同数据的分块，无法实现相似数据之间的聚类，相似数据之间的聚类问题可以通过狄利克雷过程混合模型来解决。根据CRP过程，可以得到狄利克雷过程混合模型中的聚类簇的数量。

[0070] 在一种假定的情景中，一家中国餐厅拥有无限数量的圆形桌子，每个桌子都有无

限的容量。此时，一个顾客坐在第一张桌子旁。下一个顾客可以选择坐在与该顾客相同的桌子旁，或者选择坐在下一张桌子旁。这样的情况会一直持续下去，每个顾客只能有两种选择，一种选择坐在已有其他顾客的桌子旁，另一种选择是选择一张未被占用的桌子，其中，选择已有顾客桌子的概率与该桌子已经存在的顾客数量成比例，即与含有少量顾客的桌子相比，他们更愿意坐在含有更多顾客的桌子旁。此过程的结果具有可交换性，这意味着顾客所就坐的顺序不会影响最终分配的可能性。中国餐馆过程的该性质极大地简化了文本的主题挖掘中的许多问题，例如，通过狄利克雷过程混合模型得到主题的数量可以通过中国餐馆过程确定。

[0071] 基于上述任一实施例，图4为本发明实施例提供的基于语义的文档处理方法的流程示意图，如图4所示，对文档聚类过程进行建模具体包括：

[0072] 步骤1021，根据参数为α_0的狄利克雷分布生成一个主题分布φ_i。

[0073] 具体地，在步骤1021中，根据$\varphi_i \sim \mathrm{Dir}(\alpha_0)$，采样$\varphi_i$。

[0074] 步骤1022，根据参数为β的狄利克雷分布生成一个单词分布θ_k。

[0075] 具体地，在步骤1022中，根据θ_k～Dir(β)，采样θ_k。

[0076] 步骤1023，根据参数为φ_i的多项分布，生成当前文档中单词的主题z_i；

[0077] 具体地，在步骤1023中，根据$z_i \sim \mathrm{Mult}(\varphi_i)$，采样一个聚类簇。这里，每个聚类簇对应一个主题。

[0078] 步骤1024，根据参数为θ_{z_i}的多项分布，生成当前文档的单词$x_{i,j}$。

[0079] 具体地，在步骤1024中，根据$x_{i,j} \sim \mathrm{Mult}(\theta_{z_i})$，依次生成$x_i$文档中的每一个词。

[0080] 如图5所示，在狄利克雷过程混合模型中，首先使用两个狄利克雷先验参数α_0和β为当前文档随机抽取出某个主题φ和单词θ。然后从该文档所对应的多项分布φ中抽取一个主题z，然后从主题z所对应的多项分布θ中抽取一个单词x_i。然后将这个过程重复n次，就产生了文档d。其中，i取值为1,2,...,n，n为自然数。这样，当前文档产生主题的概率，主题产生单词的概率不再是两个确定的值，而是随机变量。

[0081] 假定共有M篇文章，M为自然数。每篇文章下的主题分布是一个从参数为的狄利克雷先验分布中采样得到的多项分布，每片文章的单词分布是一个从参数为的狄利克雷先验分布中采样得到的多项分布。对于某篇文章中的第n个词，首先从该文章中出现的每个主题的多项分布中选择或采样一个主题，然后再在这个主题对应的词的多项分布中选择或采样一个词。不断重复这个随机生成过程，直到M篇文章全部生成完成。综上，M篇文档会对应于M个独立的狄利克雷-多项共轭结构。

[0082] 假定给定文档d中有多个主题z_1、z_2、z_3，它们的主题分布{P(z_i|d)，i＝1,2,3}可能是{0.4,0.5,0.1}，也可能是{0.2,0.2,0.6}，即这些主题被d选中的概率都不再认为是确定的值，而主题分布到底是哪个取值集合我们不确定，但其先验分布是狄利克雷分布，所以可以从无穷多个主题分布中按照狄利克雷先验随机抽取出某个主题分布。

[0083] 在完成主题聚类模型建模之后，即可以使用建模的关系推断某个文档属于某个主题的概率。为了引入语义信息，在步骤104中，可以根据聚类簇的完整性、当前文档与聚类簇的同质性函数和当前文档与聚类簇的单词的语义权重获取当前文档属于聚类簇的概率，并

根据该概率确定当前文档所属的当前聚类簇。

[0084] 具体地，当前文档属于不同聚类簇的概率可能各不相同，取概率最大时对应的聚类簇为当前文档所属的聚类簇。

[0085] 这里，同质性函数与当前文档和聚类簇的单词的词频相关。

[0086] 具体地，在步骤104中，根据以下公式(1)获取当前文档属于聚类簇的概率 $p(z_x = z \mid \vec{z},\ \vec{x},\ \alpha_0,\ \beta)$：

[0087]
$$p(z_x = z \mid \vec{z},\ \vec{x},\ \alpha_0,\ \beta) = \left(\frac{m_z}{D-1+\alpha_0 D}\right) * \left(\frac{\prod_{w\in x}\prod_{j=1}^{N_x^w}(n_z^w \cdot ICF_w)+\beta+j-1}{\prod_{i=1}^{N_x} n_z+V\beta+i-1}\right) * \left(1+\sum_{w_i\in x \wedge w_j\in x} cw_{ij}\right) \quad (1)$$

[0088] 其中，α_0为设定的参数，i和j均为自然数，m_z表示聚类簇中包含的文档数量，D表示所有聚类簇中文档的数量，N_x^w表示在文档x中单词w的词频，N_x表示文档x的总词数，n_z^w表示在聚类簇z中单词w的词频，n_z表示聚类簇z的总词数，V表示词典大小，ICF_w表示单词w的在各个聚类簇之间的重要程度，其计算公式为$ICF(w \in x) = \log\left(\frac{|Z|}{|w\in Z|}\right)$，语义权重$cw_{ij}$的公式为：

[0089]
$$cw_{ij} = \frac{\sum_{x'\subseteq z} n_{x'}^{w_i}}{\sum_{x'\subseteq z} n_{x'}^{w_i} + \sum_{x'\subseteq z} n_{x'}^{w_j}} (w_i,\ w_j) \in x'$$

。

[0090] 其中，词典可以理解为一种单词数据库，储存有所有的单词。词典的大小即为词典所有单词的个数。

[0091] 在以上公式(1)中，等号右侧第一项的数值可以表示聚类簇的完整性。可以看出，聚类簇中包含的文档数量越多，等号右侧第一项的数值越高，该聚类簇的完整性越高。

[0092] 等号右侧第二项可以看做一个同质性函数，其数值可以用来衡量聚类簇和文档之间的同质性。等号右侧第二项的数值越高，聚类簇和文档之间的同质性越高。聚类簇的同质性表征的是聚类的紧凑化程度。聚类簇的同质性函数的取值范围为[0,1]，其值越大意味着聚类的紧凑化程度越高。

[0093] 等号右侧的第三项表示聚类簇和文档之间的单词的语义权重。等号右侧第三项的数值越高，聚类簇和文档之间的单词的语义权重越高。

[0094] 在步骤104中，可以得到当前文档属于不同聚类簇的概率。在步骤104中，取具有最大值的概率对应的聚类簇作为当前文档所属的当前聚类簇。

[0095] 例如，当前文档属于不同聚类簇的概率各不相同，当前文档属于某聚类簇的概率为0.8，而当前文档属于其它聚类簇的概率均小于0.8，则概率0.8对应的聚类簇为当前文档所属的当前聚类簇。

[0096] 在步骤106之前，遍历所有聚类簇，获取计算当前聚类簇的主题与其它聚类簇的主题之间的相似度。这里的所有聚类簇，指的是现有的知识库中所有的聚类簇。

[0097] 这里，当前聚类簇与其它聚类簇之间的相似度即为当前聚类簇的主题与其它聚类簇的主题之间的相似度。

[0098] 计算两个聚类簇的主题的相似度，可以通过欧式距离公式、余弦相似度距离公式或者Jaccard相似系数计算二者之间的相似度。

[0099] 当前聚类簇与哪个聚类簇相似度越高，越有可能归属于哪个聚类簇。但是并不能说，当前聚类簇与哪个聚类簇相似度最高，就一定归属于这个聚类簇。这个时候，需要找到知识库中与当前聚类簇相似度最高的聚类簇，并进行下一步的处理。

[0100] 在步骤106中，取知识库中与当前聚类簇相似度最高的聚类簇作为目标知识点。

[0101] 该当前聚类簇可能属于该目标知识点，也可能不属于该目标知识点。若当前聚类簇不属于目标知识点，则是一个新的聚类簇。

[0102] 在判断当前聚类簇是否属于目标知识点时，可以设定一个阈值，并根据二者的相似度是否大于等于这个阈值，判断当前聚类簇是否属于目标知识点。

[0103] 具体地，若所述当前聚类簇与所述目标知识点的相似度大于等于设定的阈值，则确定所述当前聚类簇属于所述目标知识点。

[0104] 这里，设定的阈值与步骤108中的阈值可以相同，也可以不相同。

[0105] 例如，可以设定步骤108中的阈值为0.78，若当前聚类簇和目标知识点的相似度小于0.78，则当前聚类簇为新的聚类簇。同时，设定当前聚类簇和目标知识点的相似度的阈值0.79，只有当二者的相似度大于等于0.79，才能判断当前聚类簇属于目标知识点。

[0106] 还可以设定两个阈值均为0.75，此时，若当前聚类簇和目标知识点的相似度小于0.75，则当前聚类簇为新的聚类簇。若当前聚类簇和目标知识点的相似度大于等于0.75，则当前聚类簇属于目标知识点。

[0107] 在步骤108之后，文档处理方法还包括通过提取关键词确定新主题的类目。

[0108] 如图6所示，基于发明实施例的基于语义的文档分析方法，可以将未命中问题以及人工日志的内容作为输入文档进行知识挖掘。

[0109] 具体地，未命中问题和经过质检挖掘的人工日志经过聚类分析后，可以得到文档处于不同的聚类簇的概率，根据概率可以得到每个文档所属的聚类簇。将该聚类簇与现有的知识体系进行相似度计算，找到与当前聚类簇最相似的知识点。如果当前聚类簇与最相似的知识点的相似度大于等于设定的阈值，则认为当前聚类簇属于该知识点，如果当前聚类簇与最相似的知识点的相似度小于某个阈值，则认为当前聚类簇属于新知识点。此时，可以通过提取关键词或者人工检查的方式来定义新知识点的类目。

[0110] 具体地，在对文档聚类过程进行建模后，可以通过分类、匹配、主题提取等方式对文档进行处理，即为该文档打标签。之后，将文档主题属于已有知识点或者新知识点进行输出。

[0111] 本发明提供的基于语义的文档处理方法，通过增强的狄利克雷过程混合模型对文档聚类过程进行建模，并结合文档和聚类簇的单词的词频和语义权重进行概率计算，以判断当前文档所属的当前聚类簇中的主题是否新主题，增强了模型的聚类能力，实现了结合语义对文档聚类过程进行建模和主题挖掘，具有较好的主题挖掘效果。

[0112] 下面对本发明提供的基于语义的文档处理装置进行描述，下文描述的基于语义的文档处理装置与上文描述的基于语义的文档处理方法可相互对应参照。

[0113] 如图7所示，基于语义的文档处理装置包括：

[0114] 当前聚类簇确定单元704，用于根据第一公式确定当前文档所属的当前聚类簇，其

中，第一公式根据预设的主题聚类模型、当前文档与聚类簇的单词的词频和语义权重得到，主题聚类模型采用狄利克雷过程混合模型和中国餐馆过程对文档聚类过程进行建模得到，主题聚类模型用于确定文档中的单词属于一个聚类簇的概率。

[0115] 目标知识点获取单元706，用于获取知识库中与当前聚类簇相似度最高的目标知识点；

[0116] 新主题确定单元708，用于在若当前聚类簇与目标知识点的相似度小于设定的阈值时，确定当前聚类簇中的主题为新主题。

[0117] 本发明实施例中的技术方案中，基于主题聚类模型，根据文档与聚类簇之间的单词的频率和语义权重进行文档属于某个聚类簇的概率计算，通过词的共现关系来引入语义信息，强化了模型的聚类能力，进而在对未命中Query进聚类分析时可以更好地挖掘到新意图、新知识点的主题，从而缓解运营的压力，方便运营人员新增知识点。

[0118] 其中，建模单元702还可以包括：

[0119] 第一生成子单元，用于根据参数为α_0的狄利克雷分布生成一个主题分布φ_i。

[0120] 第二生成子单元，用于根据参数为β的狄利克雷分布生成一个单词分布θ_k。

[0121] 第三生成子单元，用于根据参数为φ_i的多项分布，生成当前文档中单词的主题z_i；

[0122] 第四生成子单元，用于根据参数为θ_{z_i}的多项分布，生成当前文档的单词$x_{i,j}$。

[0123] 概率获取单元704还用于根据聚类簇的完整性、当前文档与聚类簇的同质性函数和当前文档与聚类簇的单词的语义权重获取当前文档属于聚类簇的概率。

[0124] 概率获取单元704根据以下公式(1)获取当前文档属于聚类簇的概率$p(z_x = z \mid \vec{z},\ \vec{x},\ \alpha_0,\ \beta)$:

[0125]
$$p(z_x = z \mid \vec{z},\ \vec{x},\ \alpha_0,\ \beta) = \left(\frac{m_z}{D-1+\alpha_0 D}\right) * \left(\frac{\prod_{w\in x}\prod_{j=1}^{N_x^w}(n_z^w \cdot ICF_w)+\beta+j-1}{\prod_{i=1}^{N_x} n_z+V\beta+i-1}\right) * \left(1+\sum_{w_i\in x \wedge w_j\in x} cw_{ij}\right) \quad (1)$$

[0126] 其中，α_0为设定的参数，i和j均为自然数，m_z表示聚类簇中包含的文档数量，D表示所有聚类簇中文档的数量，N_x^w表示在文档x中单词w的词频，N_x表示文档x的总词数，n_z^w表示在聚类簇z中单词w的词频，n_z表示聚类簇z的总词数，V表示词典大小，ICF_w表示单词w的在各个聚类簇之间的重要程度，其计算公式为$ICF(w \in x) = \log\left(\frac{|Z|}{|w\in Z|}\right)$，语义权重$cw_{ij}$的公式为：

[0127]
$$cw_{ij} = \frac{\sum_{x'\subseteq z} n_{x'}^{w_i}}{\sum_{x'\subseteq z} n_{x'}^{w_i} + \sum_{x'\subseteq z} n_{x'}^{w_j}} (w_i,\ w_j) \in x'$$

。

[0128] 新主题确定单元708还用于在获取知识库中与当前聚类簇相似度最高的目标知识点之后，若当前聚类簇与目标知识点的相似度大于等于阈值，则确定当前聚类簇属于目标知识点。

[0129] 基于语义的文档处理装置还可以包括相似度计算单元，用于在获取知识库中与当

前聚类簇相似度最高的目标知识点之前，遍历所有聚类簇，获取计算当前聚类簇的主题与其它聚类簇的主题之间的相似度。

[0130] 基于语义的文档处理装置还可以包括分类单元，用于在确定当前聚类簇中的主题为新主题后，通过提取关键词确定新主题的类目。

[0131] 由于本发明的示例实施例的基于语义的文档处理装置的各个功能模块与上述基于语义的文档处理方法的示例实施例的步骤对应，因此对于本发明装置实施例中未披露的细节，请参照本发明上述的基于语义的文档处理方法的实施例。

[0132] 本发明提供的基于语义的文档处理装置，通过增强的狄利克雷过程混合模型对文档聚类过程进行建模，并结合文档和聚类簇的单词的词频和语义权重进行概率计算，以判断当前文档所属的当前聚类簇中的主题是否新主题，增强了模型的聚类能力，实现了结合语义对文档聚类过程进行建模和主题挖掘，具有较好的主题挖掘效果。

[0133] 图8示例了一种电子设备的实体结构示意图，如图8所示，该电子设备可以包括：处理器(processor)810、通信接口(Communications Interface)820、存储器(memory)830和通信总线840，其中，处理器810，通信接口820，存储器830通过通信总线840完成相互间的通信。处理器810可以调用存储器830中的逻辑指令，以执行基于语义的文档处理方法，该方法包括：根据第一公式确定所述当前文档所属的当前聚类簇，其中，所述第一公式根据预设的主题聚类模型、所述当前文档与聚类簇的单词的词频和语义权重得到，所述主题聚类模型采用狄利克雷过程混合模型和中国餐馆过程对文档聚类过程进行建模得到，所述主题聚类模型用于确定文档中的单词属于一个聚类簇的概率；获取知识库中与所述当前聚类簇相似度最高的目标知识点；若所述当前聚类簇与所述目标知识点的相似度小于设定的阈值，则确定所述当前聚类簇中的主题为新主题。

[0134] 此外，上述的存储器830中的逻辑指令可以通过软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U盘、移动硬盘、只读存储器(ROM，Read-Only Memory)、随机存取存储器(RAM，Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0135] 另一方面，本发明还提供一种计算机程序产品，所述计算机程序产品包括存储在非暂态计算机可读存储介质上的计算机程序，所述计算机程序包括程序指令，当所述程序指令被计算机执行时，计算机能够执行上述各方法所提供的基于语义的文档处理方法，该方法包括：根据第一公式确定所述当前文档所属的当前聚类簇，其中，所述第一公式根据预设的主题聚类模型、所述当前文档与聚类簇的单词的词频和语义权重得到，所述主题聚类模型采用狄利克雷过程混合模型和中国餐馆过程对文档聚类过程进行建模得到，所述主题聚类模型用于确定文档中的单词属于一个聚类簇的概率；获取知识库中与所述当前聚类簇相似度最高的目标知识点；若所述当前聚类簇与所述目标知识点的相似度小于设定的阈值，则确定所述当前聚类簇中的主题为新主题。

[0136] 又一方面，本发明还提供一种非暂态计算机可读存储介质，其上存储有计算机程

序，该计算机程序被处理器执行时实现以执行上述各提供的基于语义的文档处理方法，该方法包括：根据第一公式确定所述当前文档所属的当前聚类簇，其中，所述第一公式根据预设的主题聚类模型、所述当前文档与聚类簇的单词的词频和语义权重得到，所述主题聚类模型采用狄利克雷过程混合模型和中国餐馆过程对文档聚类过程进行建模得到，所述主题聚类模型用于确定文档中的单词属于一个聚类簇的概率；获取知识库中与所述当前聚类簇相似度最高的目标知识点；若所述当前聚类簇与所述目标知识点的相似度小于设定的阈值，则确定所述当前聚类簇中的主题为新主题。

[0137] 以上所描述的装置实施例仅仅是示意性的，其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性的劳动的情况下，即可以理解并实施。

[0138] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到各实施方式可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件。基于这样的理解，上述技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品可以存储在计算机可读存储介质中，如ROM/RAM、磁碟、光盘等，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行各个实施例或者实施例的某些部分所述的方法。

[0139] 最后应说明的是：以上实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

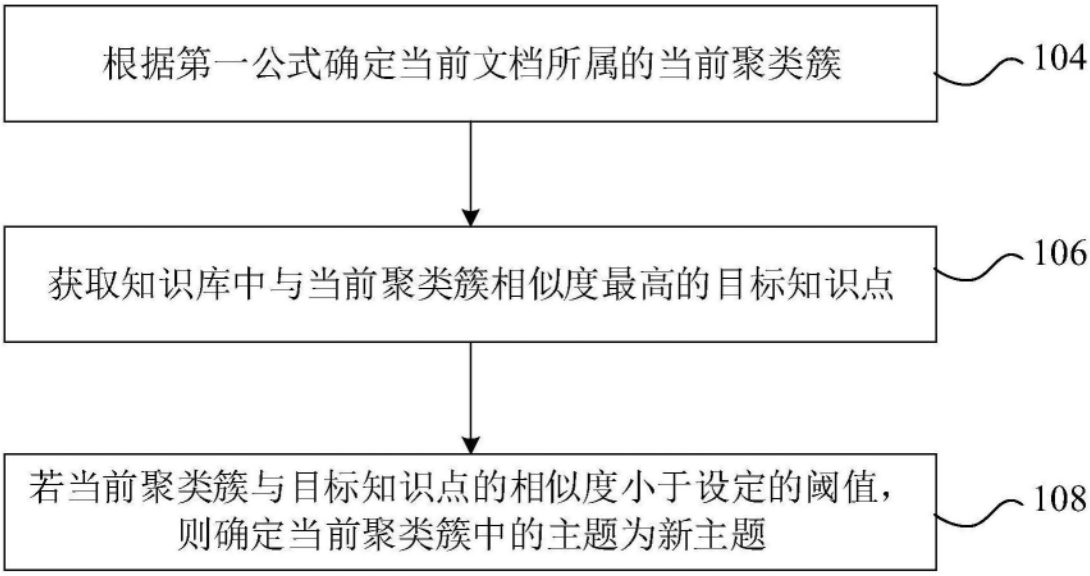

根据第一公式确定当前文档所属的当前聚类簇
104
获取知识库中与当前聚类簇相似度最高的目标知识点
106
若当前聚类簇与目标知识点的相似度小于设定的阈值，则确定当前聚类簇中的主题为新主题
108

图1

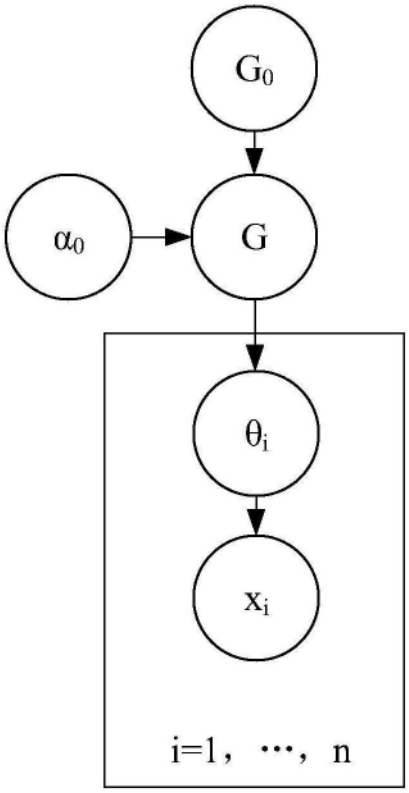

G_0
α_0
G
θ_i
x_i
i=1，…，n

图2

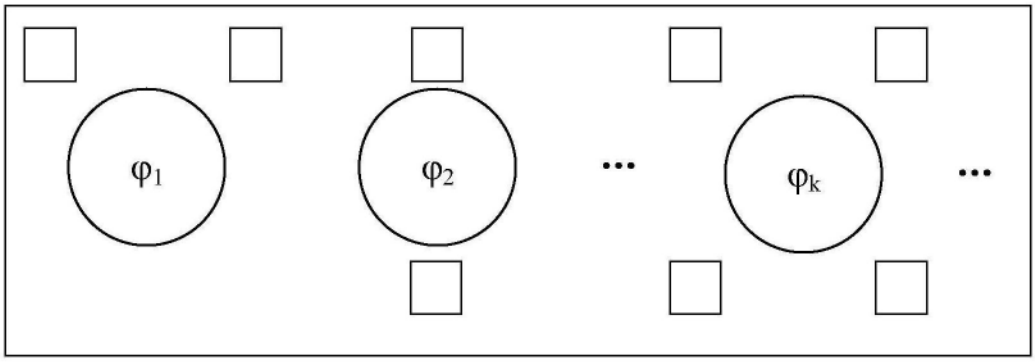

φ1
φ2
φk

图3

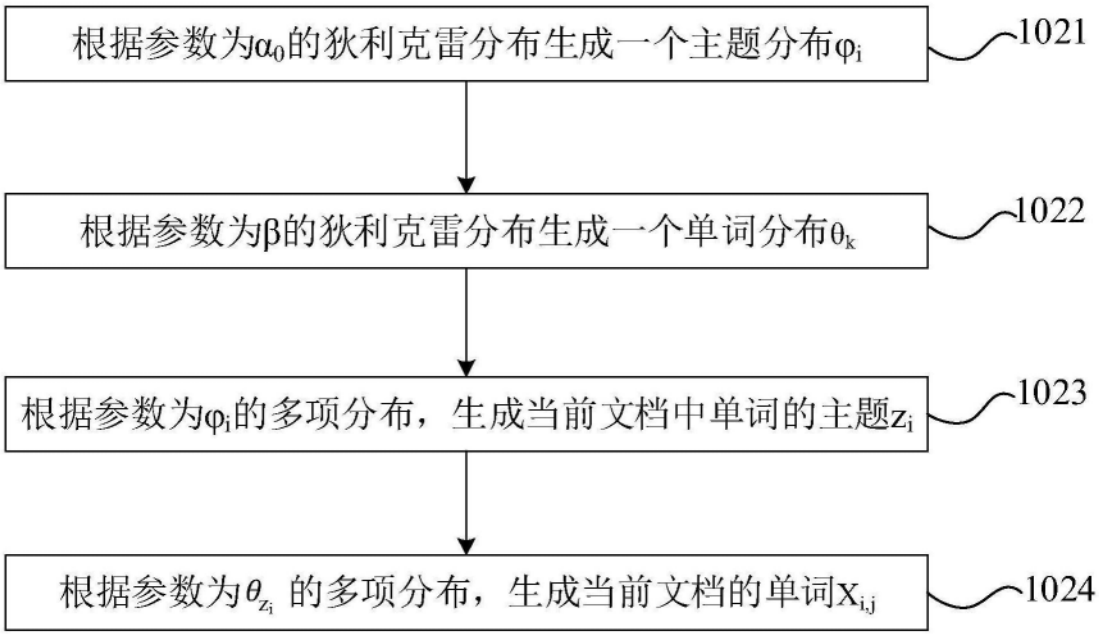

根据参数为α0的狄利克雷分布生成一个主题分布φi
1021
根据参数为β的狄利克雷分布生成一个单词分布θk
1022
根据参数为φi的多项分布，生成当前文档中单词的主题zi
1023
根据参数为θzi 的多项分布，生成当前文档的单词Xi,j
1024

图4

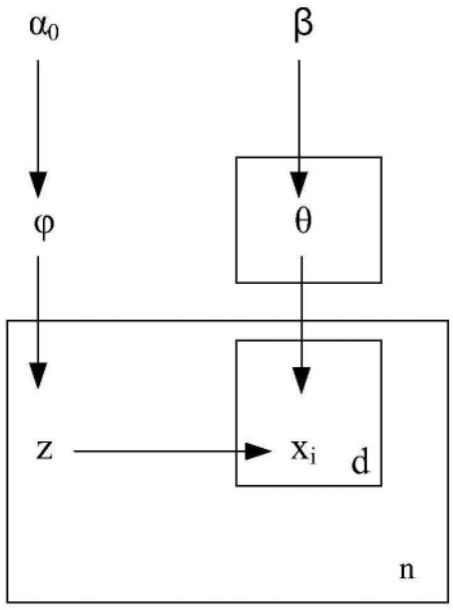

α0
β
φ
θ
z
xi
d
n

图5

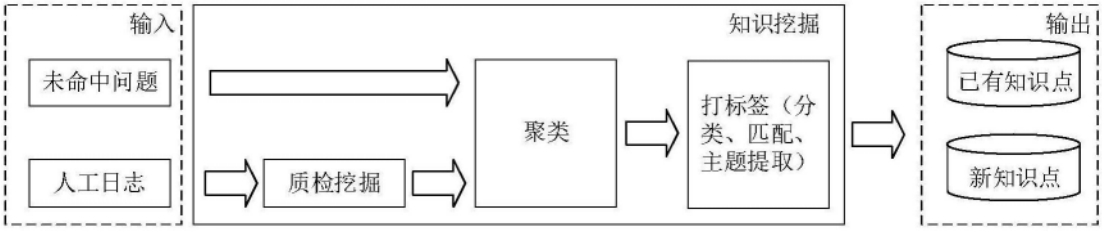

输入
未命中问题
人工日志
知识挖掘
质检挖掘
聚类
打标签（分类、匹配、主题提取）
输出
已有知识点
新知识点

图6

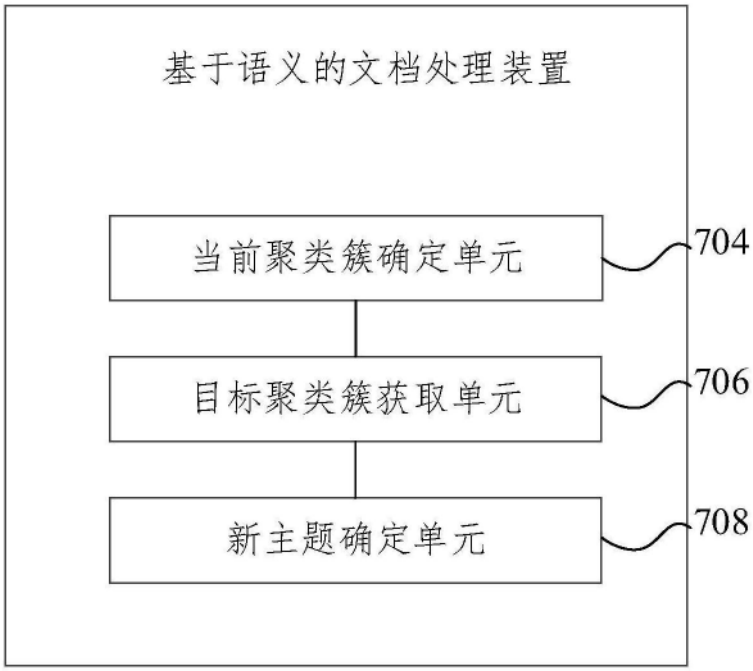

基于语义的文档处理装置
当前聚类簇确定单元
704
目标聚类簇获取单元
706
新主题确定单元
708

图7

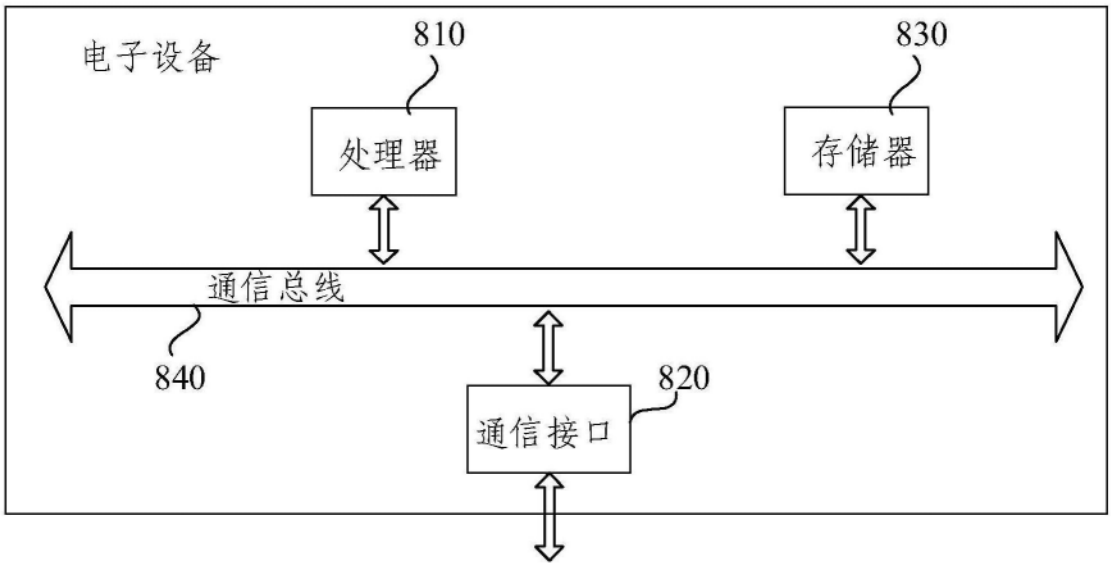

电子设备
810
处理器
830
存储器
通信总线
840
820
通信接口

图8