



(21) 申请号 201810898747.1

(22) 申请日 2018.08.08

(65) 同一申请的已公布的文献号

申请公布号 CN 110209808 A

(43) 申请公布日 2019.09.06

(73) 专利权人 腾讯科技(深圳)有限公司

地址 518057 广东省深圳市南山区高新区

科技中一路腾讯大厦35层

(72) 发明人 周鑫鹏 李阳 苏文杰

(74) 专利代理机构 深圳市深佳知识产权代理事

务所(普通合伙) 44285

专利代理师 王仲凯

(51) Int. Cl.

G06F 16/35 (2019.01)

(56) 对比文件

CN 104239436 A, 2014.12.24

CN 108170773 A, 2018.06.15

CN 103699663 A, 2014.04.02

US 2012254188 A1, 2012.10.04

审查员 何华

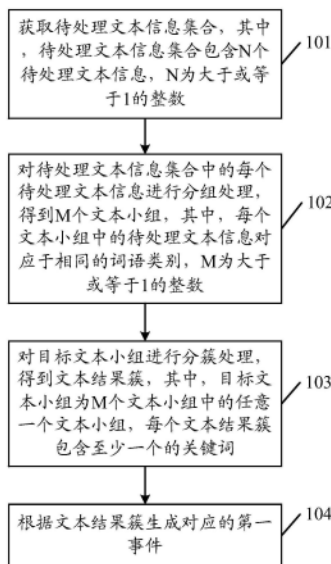
权利要求书5页 说明书26页 附图4页

(54) 发明名称

一种基于文本信息的事件生成方法以及相关装置

(57) 摘要

本发明实施例公开了一种基于文本信息的事件生成方法,包括:获取待处理文本信息集合,待处理文本信息集合包含N个待处理文本信息;对待处理文本信息集合中的每个待处理文本信息进行分组处理,得到M个文本小组,每个文本小组中的待处理文本信息对应于相同的词语类别;对目标文本小组进行分簇处理,得到文本结果簇,其中,目标文本小组为M个文本小组中的任意一个文本小组,每个文本结果簇包含至少一个的关键词;根据文本结果簇生成对应的第一事件。本发明实施例中还提供了一种事件生成装置。本发明实施例一方面节省了处理资源和计算资源,提升事件挖掘效率,另一方面,对分组后的待处理文本信息进行聚类,可以有效地增强事件生成的准确度。



1. 一种基于文本信息的事件生成方法,其特征在于,包括:

获取待处理文本信息集合,其中,所述待处理文本信息集合包含N个待处理文本信息,所述N为大于或等于1的整数;

对所述待处理文本信息集合中的每个待处理文本信息进行分组处理,得到M个文本小组,其中,每个文本小组中的待处理文本信息对应于相同的词语类别,所述M为大于或等于1的整数;

对目标文本小组进行分簇处理,得到文本结果簇,其中,所述目标文本小组为所述M个文本小组中的任意一个文本小组,每个文本结果簇包含至少一个的关键词;所述对目标文本小组进行分簇处理,得到文本结果簇,包括:从所述目标文本小组中获取待选关键词集合,其中,所述待选关键词集合中包含至少一个待选关键词;根据所述目标文本小组中的待处理文本信息数量,计算得到关键词个数;按照所述关键词个数从所述待选择关键词集合中选择P个目标关键词,其中,所述P为大于或等于1的整数,且所述P表示所述关键词个数;根据所述P个目标关键词生成至少一个所述文本结果簇;所述根据所述P个目标关键词生成至少一个所述文本结果簇,包括:根据所述P个目标关键词生成P个特征向量,其中,每个目标关键词对应于一个特征向量,每个特征向量中包含特征元素,所述特征元素用于表示目标关键词在待处理文本信息中的出现情况;根据所述P个特征向量确定聚类个数;根据所述聚类个数生成所对应的所述文本结果簇;所述根据所述P个特征向量确定聚类个数,包括:获取预设聚类个数;根据所述预设聚类个数以及所述P个特征向量,计算得到轮廓系数;根据所述预设聚类个数计算各个预设聚类簇之间的平均相似度;根据所述预设聚类个数计算每个所述预设聚类簇内待处理文本信息的相似度;根据所述各个预设聚类簇之间的平均相似度,以及每个预设聚类簇内待处理文本信息的相似度,计算得到离散系数;根据所述轮廓系数以及所述离散系数,计算簇数得分;若所述簇数得分满足簇数选取条件,则将所述预设聚类个数确定为所述聚类个数;

根据所述文本结果簇生成对应的第一件事件。

2. 根据权利要求1所述的方法,其特征在于,所述对所述待处理文本信息集合中的每个待处理文本信息进行分组处理,得到M个文本小组,包括:

获取目标待处理文本信息,其中,所述目标待处理文本信息属于所述待处理文本信息集合中的任意一个待处理文本信息;

若所述目标待处理文本信息中包含实体信息,则将所述目标待处理文本信息划分至第一文本小组,其中,所述第一文本小组属于所述M个文本小组中的一个文本小组,所述实体信息用于表示一个固定信息。

3. 根据权利要求2所述的方法,其特征在于,所述获取目标待处理文本信息之后,所述方法还包括:

若所述目标待处理文本信息中不包含所述实体信息,则判断所述目标待处理文本信息中是否包含第一词语组合或第二词语组合,其中,所述第一词语组合包含地域信息和通名信息,所述第二词语组合包含地域信息和领域信息,所述地域信息用于表示地点名称,所述通名信息用于表示一类固定信息,所述领域信息用于表示一个领域;

若所述目标待处理文本信息中包含所述第一词语组合或所述第二词语组合,则将所述目标待处理文本信息划分至第二文本小组,其中,所述第二文本小组属于所述M个文本小组

中的一个文本小组。

4. 根据权利要求3所述的方法,其特征在于,所述判断所述目标待处理文本信息中是否包含第一词语组合或第二词语组合之后,所述方法还包括:

若所述目标待处理文本信息中不包含所述第一词语组合或所述第二词语组合,则判断所述目标待处理文本信息中是否包含所述地域信息;

若所述目标待处理文本信息中包含所述地域信息,则将所述目标待处理文本信息划分至第三文本小组,其中,所述第三文本小组属于所述M个文本小组中的一个文本小组。

5. 根据权利要求4所述的方法,其特征在于,所述判断所述目标待处理文本信息中是否包含所述地域信息之后,所述方法还包括:

若所述目标待处理文本信息中不包含所述地域信息,则判断所述目标待处理文本信息中是否包含第三词语组合或第四词语组合,其中,所述第三词语组合包含所述通名信息和事件信息,所述第四词语组合包含所述领域信息和所述事件信息,所述所述事件信息用于表示一个事件;

若所述目标待处理文本信息中包含所述第三词语组合或所述第四词语组合,则将所述目标待处理文本信息划分至第四文本小组,其中,所述第四文本小组属于所述M个文本小组中的一个文本小组;

若所述目标待处理文本信息中不包含所述第三词语组合或所述第四词语组合,则从所述待处理文本信息集合中剔除所述目标待处理文本信息。

6. 根据权利要求1所述的方法,其特征在于,所述根据所述预设聚类个数以及所述P个特征向量,计算得到轮廓系数,包括:

采用如下方式计算所述轮廓系数:

$$score_k^{sc} = \sum_{i=0}^P \frac{b(k,i) - a(k,i)}{\max(b(k,i), a(k,i))};$$

其中,所述 $score_k^{sc}$ 表示在所述预设聚类个数下得到的所述轮廓系数,所述P表示所述目标关键词的数目,所述k表示所述预设聚类个数,所述i表示第i个特征向量,所述a(k,i)表示在所述预设聚类个数为k时第i个特征向量到同一簇内其他特征向量的平均距离,所述b(k,i)表示在所述预设聚类个数为k时第i个特征向量到其他簇所有特征向量的平均距离;

所述根据所述预设聚类个数计算各个预设聚类簇之间的平均相似度,包括:

采用如下方式计算所述各个预设聚类簇之间的平均相似度:

$$\tilde{s}_i = \frac{\sum_{j=0}^{P_i} Len(\sum_{l=0}^{l=P, l \neq i} set(keyword_l)) \cap set(sentence_j))}{P_i};$$

其中,所述 $\tilde{s}_i$ 表示第i个预设聚类簇与其他预设聚类簇之间的平均相似度,所述 P_i 表示所述第i个预设聚类簇中待处理文本信息的数目,所述 $set(keyword_l)$ 表示所述第i个预设聚类簇的关键词列表,所述 $set(sentence_j)$ 表示所述第i个预设聚类簇的第j个待处理文本信息的词语列表,所述 $\cap$ 表示词语列表的交集,所述 $Len()$ 表示所述词语列表中的词语个数;

所述根据所述预设聚类个数计算每个所述预设聚类簇内待处理文本信息的相似度,包

括：

采用如下方式计算所述预设聚类簇内待处理文本信息的相似度：

$$s_i = \frac{\sum_{j=0}^{n_i} \text{Len}(\text{set}(\text{keyword}_i) \cap \text{set}(\text{sentence}_j))}{n_i};$$

其中，所述 s_i 表示所述第 i 个预设聚类簇待处理文本信息的相似度；

所述根据所述各个预设聚类簇之间的平均相似度，以及每个预设聚类簇内待处理文本信息的相似度，计算得到离散系数，包括：

采用如下方式计算所述离散系数：

$$\text{score}_k^{\text{doc}} = \frac{\sum_{i=0}^{i=k} s_i}{\sum_{i=0}^{i=k} \tilde{s}_i};$$

其中，所述 $\text{score}_k^{\text{doc}}$ 表示在所述预设聚类个数下得到的所述离散系数；

所述根据所述轮廓系数以及所述离散系数，计算簇数得分，包括：

采用如下方式计算所述簇数得分：

$$\text{score}_k = \text{score}_k^{\text{sc}} + \text{score}_k^{\text{doc}};$$

其中，所述 score_k 表示在所述预设聚类个数下得到的所述簇数得分。

7. 根据权利要求1所述的方法，其特征在于，所述根据所述 P 个特征向量确定聚类个数，替换为包括：

根据所述 P 个特征向量生成 Q 个待选择文本簇，其中，每个待选择文本簇中包含至少一个待处理文本信息，所述 Q 为大于或等于1的整数；

1) 获取 Q 个待选择文本簇中一个待选择文本簇的待处理文本信息数量；

2) 若所述待处理文本信息数量大于或等于数量阈值，则在上一个聚类计数值上增加1，得到本次聚类计数值；

重复执行步骤1) 和步骤2)，直至所述 Q 个待选择文本簇中的待处理文本信息数量均与所述数量阈值比对完成，根据所述本次聚类计数值确定所述聚类个数。

8. 根据权利要求1至7中任一项所述的方法，其特征在于，所述根据所述文本结果簇生成对应的第一事件之后，所述方法还包括：

获取第二事件；

根据所述第一事件和所述第二事件计算事件相似度，其中，所述第一事件对应于至少一个第一文本信息，所述第二事件对应于至少一个第二文本信息；

若所述相似度大于或等于预设门限，则确定所述第二事件与所述第一事件为相似事件，并将所述第二事件与所述第一事件进行合并；

若所述相似度小于所述预设门限，则为所述第二事件分配标识，其中，不同的标识用于表示不同的事件。

9. 根据权利要求8所述的方法，其特征在于，所述根据所述第一事件和所述第二事件计算事件相似度，包括：

采用如下方式计算所述事件相似度：

$$\alpha = \frac{|A \cap B|}{\min(|A|, |B|)};$$

其中,所述 α 表示所述事件相似度,所述A表示所述第一事件对应于所述第一文本信息的个数,所述B表示所述第二事件对应于所述第二文本信息的个数,所述 $A \cap B$ 表示所述第一文本信息与所述第二文本信息重合的文本信息个数。

10.一种事件生成装置,其特征在于,包括:

获取模块,用于获取待处理文本信息集合,其中,所述待处理文本信息集合包含N个待处理文本信息,所述N为大于或等于1的整数;

处理模块,用于对所述获取模块获取的所述待处理文本信息集合中的每个待处理文本信息进行分组处理,得到M个文本小组,其中,每个文本小组中的待处理文本信息对应于相同的词语类别,所述M为大于或等于1的整数;

所述处理模块,还用于对目标文本小组进行分簇处理,得到文本结果簇,其中,所述目标文本小组为所述M个文本小组中的任意一个文本小组,每个文本结果簇包含至少一个的关键词;

生成模块,用于根据所述处理模块处理得到的所述文本结果簇生成对应的第一事件;

所述处理模块,具体用于从所述目标文本小组中获取待选关键词集合,其中,所述待选关键词集合中包含至少一个待选关键词;根据所述目标文本小组中的待处理文本信息数量,计算得到关键词个数;按照所述关键词个数从所述待选择关键词集合中选择P个目标关键词,其中,所述P为大于或等于1的整数,且所述P表示所述关键词个数;根据所述P个目标关键词生成至少一个所述文本结果簇;所述根据所述P个目标关键词生成至少一个所述文本结果簇,包括:根据所述P个目标关键词生成P个特征向量,其中,每个目标关键词对应于一个特征向量,每个特征向量中包含特征元素,所述特征元素用于表示目标关键词在待处理文本信息中的出现情况;根据所述P个特征向量确定聚类个数;根据所述聚类个数生成所对应的所述文本结果簇;所述根据所述P个特征向量确定聚类个数,包括:获取预设聚类个数;根据所述预设聚类个数以及所述P个特征向量,计算得到轮廓系数;根据所述预设聚类个数计算各个预设聚类簇之间的平均相似度;根据所述预设聚类个数计算每个所述预设聚类簇内待处理文本信息的相似度;根据所述各个预设聚类簇之间的平均相似度,以及每个预设聚类簇内待处理文本信息的相似度,计算得到离散系数;根据所述轮廓系数以及所述离散系数,计算簇数得分;若所述簇数得分满足簇数选取条件,则将所述预设聚类个数确定为所述聚类个数。

11.一种事件生成装置,其特征在于,包括:存储器、收发器、处理器以及总线系统;

其中,所述存储器用于存储程序;

所述处理器用于执行所述存储器中的程序,包括如下步骤:

获取待处理文本信息集合,其中,所述待处理文本信息集合包含N个待处理文本信息,所述N为大于或等于1的整数;

对所述待处理文本信息集合中的每个待处理文本信息进行分组处理,得到M个文本小组,其中,每个文本小组中的待处理文本信息对应于相同的词语类别,所述M为大于或等于1的整数;

对目标文本小组进行分簇处理,得到文本结果簇,其中,所述目标文本小组为所述M个文本小组中的任意一个文本小组,每个文本结果簇包含至少一个的关键词;所述对目标文本小组进行分簇处理,得到文本结果簇,包括:从所述目标文本小组中获取待选关键词集合,其中,所述待选关键词集合中包含至少一个待选关键词;根据所述目标文本小组中的待处理文本信息数量,计算得到关键词个数;按照所述关键词个数从所述待选择关键词集合中选择P个目标关键词,其中,所述P为大于或等于1的整数,且所述P表示所述关键词个数;根据所述P个目标关键词生成至少一个所述文本结果簇;所述根据所述P个目标关键词生成至少一个所述文本结果簇,包括:根据所述P个目标关键词生成P个特征向量,其中,每个目标关键词对应于一个特征向量,每个特征向量中包含特征元素,所述特征元素用于表示目标关键词在待处理文本信息中的出现情况;根据所述P个特征向量确定聚类个数;根据所述聚类个数生成所对应的所述文本结果簇;所述根据所述P个特征向量确定聚类个数,包括:获取预设聚类个数;根据所述预设聚类个数以及所述P个特征向量,计算得到轮廓系数;根据所述预设聚类个数计算各个预设聚类簇之间的平均相似度;根据所述预设聚类个数计算每个所述预设聚类簇内待处理文本信息的相似度;根据所述各个预设聚类簇之间的平均相似度,以及每个预设聚类簇内待处理文本信息的相似度,计算得到离散系数;根据所述轮廓系数以及所述离散系数,计算簇数得分;若所述簇数得分满足簇数选取条件,则将所述预设聚类个数确定为所述聚类个数;

根据所述文本结果簇生成对应的第一事件;

所述总线系统用于连接所述存储器以及所述处理器,以使所述存储器以及所述处理器进行通信。

12.一种计算机可读存储介质,包括指令,当其在计算机上运行时,使得计算机执行如权利要求1至9中任一项所述的方法。

一种基于文本信息的事件生成方法以及相关装置

技术领域

[0001] 本发明涉及计算机处理领域,尤其涉及一种基于文本信息的事件生成方法以及相关装置。

背景技术

[0002] 随着互联网技术的迅猛发展和计算机的普及,用户越来越容易在网站上获得大量正在发生的事件新闻。大量新闻媒体将新闻放到网站,这对用户能否快速掌握主要事件成为一个挑战。因此,新闻事件的挖掘是一个非常意义的研究课题。

[0003] 基于新闻流的事件挖掘是指从海量的实时新闻数据中挖掘出蕴含其中的事件,并对挖掘的事件构建一个完备的事件库。目前,已构建了一种事件挖掘系统,首先对每条文本信息进行词性标注和命名实体识别,然后从文本信息中抽取事件时间及事件短语,再依据事件短语对事件进行分类,最后依据事件实体和事件时间的共现信息生成事件。

[0004] 然而,由于每天更新的新闻量非常大,往往有几十万条,采用上述介绍的事件挖掘系统对每条新闻的文本信息进行处理,需要耗费相当多的处理资源和计算资源,导致事件挖掘时延较高。

发明内容

[0005] 本发明实施例提供了一种基于文本信息的事件生成方法以及相关装置,一方面节省了处理资源和计算资源,提升事件挖掘效率,另一方面,对分组后的待处理文本信息进行聚类,可以有效地增强事件生成的准确度。

[0006] 有鉴于此,本发明的第一方面提供了一种基于文本信息的事件生成方法,包括:

[0007] 获取待处理文本信息集合,其中,所述待处理文本信息集合包含N个待处理文本信息,所述N为大于或等于1的整数;

[0008] 对所述待处理文本信息集合中的每个待处理文本信息进行分组处理,得到M个文本小组,其中,每个文本小组中的待处理文本信息对应于相同的词语类别,所述M为大于或等于1的整数;

[0009] 对目标文本小组进行分簇处理,得到文本结果簇,其中,所述目标文本小组为所述M个文本小组中的任意一个文本小组,每个文本结果簇包含至少一个的关键词;

[0010] 根据所述文本结果簇生成对应的第一事件。

[0011] 本发明的第二方面提供了一种事件生成装置,包括:

[0012] 获取模块,用于获取待处理文本信息集合,其中,所述待处理文本信息集合包含N个待处理文本信息,所述N为大于或等于1的整数;

[0013] 处理模块,用于对所述获取模块获取的所述待处理文本信息集合中的每个待处理文本信息进行分组处理,得到M个文本小组,其中,每个文本小组中的待处理文本信息对应于相同的词语类别,所述M为大于或等于1的整数;

[0014] 所述处理模块,还用于对目标文本小组进行分簇处理,得到文本结果簇,其中,所

述目标文本小组为所述M个文本小组中的任意一个文本小组,每个文本结果簇包含至少一个的关键词;

[0015] 生成模块,用于根据所述处理模块处理得到的所述文本结果簇生成对应的第一事件。

[0016] 本发明的第三方面提供了一种事件生成装置,包括:存储器、收发器、处理器以及总线系统;

[0017] 其中,所述存储器用于存储程序;

[0018] 所述处理器用于执行所述存储器中的程序,包括如下步骤:

[0019] 获取待处理文本信息集合,其中,所述待处理文本信息集合包含N个待处理文本信息,所述N为大于或等于1的整数;

[0020] 对所述待处理文本信息集合中的每个待处理文本信息进行分组处理,得到M个文本小组,其中,每个文本小组中的待处理文本信息对应于相同的词语类别,所述M为大于或等于1的整数;

[0021] 对目标文本小组进行分簇处理,得到文本结果簇,其中,所述目标文本小组为所述M个文本小组中的任意一个文本小组,每个文本结果簇包含至少一个的关键词;

[0022] 根据所述文本结果簇生成对应的第一事件;

[0023] 所述总线系统用于连接所述存储器以及所述处理器,以使所述存储器以及所述处理器进行通信。

[0024] 本发明的第四方面提供了一种计算机可读存储介质,所述计算机可读存储介质中存储有指令,当其在计算机上运行时,使得计算机执行上述各方面所述的方法。

[0025] 从以上技术方案可以看出,本发明实施例具有以下优点:

[0026] 本发明实施例中,提供了一种基于文本信息的事件生成方法,首先获取待处理文本信息集合,然后对待处理文本信息集合中的每个待处理文本信息进行分组处理,得到M个文本小组,每个文本小组中的待处理文本信息对应于相同的词语类别,再对目标文本小组进行分簇处理,得到文本结果簇,其中,目标文本小组为M个文本小组中的任意一个文本小组,每个文本结果簇包含至少一个的关键词,最后根据文本结果簇生成对应的第一事件。通过上述方式,在待处理文本信息集合包含大量待处理文本信息的情况下,能够根据词语类别对这些待处理文本信息先进行分组,然后进行聚类处理,避免了同时处理大量的待处理文本信息,同时提升了聚类的可信度,一方面节省了处理资源和计算资源,提升事件挖掘效率,另一方面,对分组后的待处理文本信息进行聚类,可以有效地增强事件生成的准确度。

附图说明

[0027] 图1为本发明实施例中事件生成系统的一个结构示意图;

[0028] 图2为本发明实施例中基于文本信息的事件生成系统一个架构示意图;

[0029] 图3为本发明实施例中基于文本信息的事件生成方法一个实施例示意图;

[0030] 图4为本发明实施例中增量更新事件的一个实施例示意图;

[0031] 图5为本发明实施例中事件融合的一个实施例示意图;

[0032] 图6为本发明实施例中事件生成装置的一个实施例示意图;

[0033] 图7为本发明实施例中事件生成装置的另一个实施例示意图;

[0034] 图8为本发明实施例中事件生成装置的一个结构示意图。

具体实施方式

[0035] 本发明实施例提供了一种基于文本信息的事件生成方法以及相关装置,一方面节省了处理资源和计算资源,提升事件挖掘效率,另一方面,对分组后的待处理文本信息进行聚类,可以有效地增强事件生成的准确度。

[0036] 本发明的说明书和权利要求书及上述附图中的术语“第一”、“第二”、“第三”、“第四”等(如果存在)是用于区别类似的对象,而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换,以便这里描述的本发明的实施例例如能够以除了在这里图示或描述的那些以外的顺序实施。此外,术语“包括”和“具有”以及他们的任何变形,意图在于覆盖不排他的包含,例如,包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元,而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0037] 应理解,本发明可以应用于事件挖掘的场景,基于新闻流的事件挖掘是指从海量的实时新闻数据中挖掘出蕴含其中的事件,并对挖掘的事件构建一个完备的事件库。其中,事件库中的每个事件对应一个全局唯一的事件标识。所挖掘的事件可以以结构化数据的形式表示,比如事件的实体、时间、地点和关键词等事件信息。事件挖掘属于信息抽取的子任务,近几年的事件挖掘方法从技术上可分为以下三种:

[0038] 第一种为基于数据驱动的事件抽取方法;

[0039] 数据驱动的方法是自然语言处理(Natural Language Processing,NLP)领域常用的方法,这种方法通过构建模型模拟数据之间的统计关系。基于数据驱动的文本信息挖掘方法可包括基于概率论的统计推理模型、概率模型、信息论理论以及几何代数的相关模型。常用的技术有词频统计,词义消歧,语言模型和聚类等。目前,已提出一种探测偶然事件和突发地方性事件的方法,他们通过在由搜索查询问题中的到的基于时间序列的实体上使用双层的层次聚类的方法探测事件主题及事件变化。又提出了一种从每日新闻中抽取事件的方法,他们依据实体和文档关系建立无向二分图,通过实体特征和文档的关系对文档和实体进行加权,然后对文档聚类,聚类结果得分由文档得分计算得到,得分高的结果为当天的新闻事件。后续还提出了一个开放领域的事件抽取和分类系统“TwiCal”,在“TwiCal”中对于一条新的评论数据,通过自然语言处理方法抽取事件短语及时间信息,依据事件短语对事件分类,通过事件实体和时间共现信息判断抽取的事件短语是否是一个有意义的事件。基于“TwiCal”又提出了一种无监督的贝叶斯模型模型挖掘事件,对文档、事件及实体之间分布建立模型,通过吉布斯采样更新模型参数。

[0040] 在基于狄利克雷过程的事件混合模型中,在模型中事件个数不需提前指定,而是通过模型在数据中学习得到,同时,为了解决实体词表示多样化问题,在模型中使用了词向量来表示词语。在一种半监督的文本聚类方法的中,首先通过使用卷积神经网络(Convolutional Neural Network,CNN)网络将文本进行特征化表示,然后对文本进行聚类。其中,聚类方法可以为K均值(k-means)聚类,且将聚类融合到神经网络结果中,网络参数及聚类参数同时学习,得到文本聚类模型。

[0041] 第二种为基于知识模版的事件挖掘方法;

[0042] 该方法依据基于专家知识构建的模版进行事件挖掘。通过预定义的模版或语义学规则从语料库中挖掘需求的数据信息。如通过预设的地点、目的和行为关键词模版从文档中挖掘事件信息。或者通过种子事件从文本中挖掘事件模版,然后通过使用事件模版探测新事件和事件词。

[0043] 第三种为基于混合模型的事件挖掘方法;

[0044] 基于数据的事件挖掘方法和基于知识的事件挖掘方法在使用时各有优势,但有些情形,单纯使用一种方法不能达到预期的效果,有学者将两种方法结合使用,构建事件挖掘的混合模型。可以通过使用词语句法规则和无向图的方法从议会记录中挖掘事件,或者将词语语义模版与聚类相结合从新闻文本中挖掘事件。

[0045] 相对于基于数据驱动的事件抽取方法,基于知识模版的事件挖掘方法需要更少的标注数据,其次,通过使用词法,句法及语义元素信息定义可构建有效的事件挖掘模版,并且数据结果易于解释和追踪。当挖掘特定的事件信息时,基于模版的方法通常能够更高效且精确。但是,基于模版的事件挖掘方法需要有先验知识,并且预定义的知识模版不适宜大范围使用,缺乏泛化能力。目前主要的研究方向为基于数据的事件抽取方法。

[0046] 应理解,本发明所提供的事件生成方法可以应用于事件生成系统,请参阅图1,图1为本发明实施例中事件生成系统的一个结构示意图,如图所示,本发明所提供的事件生成装置可部署于服务器,在服务器在挖掘到事件后可存储于本地,并将事件发送至终端设备,由终端设备向用户展示事件。其中,终端设备包含但不限于笔记本电脑、平板电脑、掌上电脑、手机以及个人电脑(personal computer,PC)。

[0047] 为了便于理解,下面将结合图2对本发明中的事件生成方法进行介绍,请参阅图2,图2为本发明实施例中基于文本信息的事件生成系统一个架构示意图,如图所示,具体地:

[0048] 步骤S1中,首先事件生成装置需要获取新闻集合,通常情况下,新闻集合中包含了大量的新闻内容,新闻内容又包括了新闻标题和新闻正文等,本发明主要以新闻标题为主,对新闻内容进行划分。

[0049] 步骤S2中,对新闻标题按照实体信息进行划分,可以分为若干个堆,其中,每个堆包括不同的实体信息,比如一个堆中为“人物A相关新闻”,一个堆为“腾讯相关新闻”,一个堆为“北京新闻”。

[0050] 步骤S3中,对不同堆中的新闻标题进行聚合,比如一个堆为“人物A相关新闻”,在这一堆中包括事件A,事件A为“人物A五十大寿人物B送对联事件相关新闻”。在另一堆“人物B相关新闻”中包含事件B和事件C,其中,事件B为“人物A五十大寿人物B送对联事件相关新闻”,事件C为“人物B在新节目中点赞人物C人品事件相关新闻”。其中,每个相关新闻都包括至少一篇新闻内容。

[0051] 步骤S4中,对包含不同实体信息的事件进行融合,比如,“人物A相关新闻”中包括“人物A五十大寿人物B送对联事件相关新闻”,而在“人物B相关新闻”中也包括“人物A五十大寿人物B送对联事件相关新闻”,于是可以将这两个事件进行融合,也就认为这是同一个事件,可分配同一个标识。

[0052] 步骤S5中,将新挖掘事件与历史事件进行融合,去掉重复的事件,将未融合的事件加入至数据库中。

[0053] 下面将从事件生成装置的角度,对本发明中基于文本信息的事件生成方法进行介

绍,请参阅图3,本发明实施例中基于文本信息的事件生成方法一个实施例包括:

[0054] 101、获取待处理文本信息集合,其中,待处理文本信息集合包含N个待处理文本信息,N为大于或等于1的整数;

[0055] 本实施例中,部署于服务器上的事件生成装置首先可以获取待处理文本信息集合,这里的待处理文本信息集合中包含至少一个待处理文本信息。其中,待处理文本信息具体可以指新闻的标题。

[0056] 102、对待处理文本信息集合中的每个待处理文本信息进行分组处理,得到M个文本小组,其中,每个文本小组中的待处理文本信息对应于相同的词语类别,M为大于或等于1的整数;

[0057] 本实施例中,事件生成装置需要对待处理文本信息集合中的每个待处理文本信息进行分组,假设有十万条待处理文本信息,那么就需要对这十万条待处理文本信息进行归类划分,分在一起的待处理文本信息形成一个文本小组,从而构成M个文本小组,M可以是大于或等于1的整数。

[0058] 其中,每个文本小组对应于相同的词语类别,词语类别包含但不仅限于实体信息、通名信息、地域信息、领域信息和事件信息。实体信息是指一个具体的事物,比如“腾讯科技有限公司”以及“北京”等。通名信息是指一类具体的事务,比如“老人”、“孕妇”、“教师”以及“警察”等。地域信息是指具体的地点名称,通常不一定包含在新闻的标题中,需要通过神经网络对标题进行地域识别才能得到。领域信息是指一个领域,比如“外卖”、“楼市”以及“通信”等。事件信息是指一类事件,比如“大雪”以及“上涨”等。

[0059] 103、对目标文本小组进行分簇处理,得到文本结果簇,其中,目标文本小组为M个文本小组中的任意一个文本小组,每个文本结果簇包含至少一个的关键词;

[0060] 本实施例中,事件生成装置需要对各个文本小组进行聚类分簇,这里以M个文本小组中的其中一个文本小组为例进行介绍,该文本小组即为目标文本小组。在目标文本小组中可以包括多个待处理文本信息,即新闻标题,通过这些新闻标题挖掘其中的关键字,然后基于关键字进行聚类,从而得到至少一个文本结果簇,每个文本结果簇中由包含有至少一个关键词。

[0061] 104、根据文本结果簇生成对应的第一事件。

[0062] 本实施例中,事件生成装置基于文本结果簇,在每个文本结果簇中,当新闻标题满足事件生成条件时能够生成第一事件,因此,基于K文本结果簇最多能够生成K个事件。其中,事件生成条件为需要在包含一些基本元素,比如事件时间、地点、相关实体信息、关键词以及新闻来源等。

[0063] 本发明实施例中,提供了一种基于文本信息的事件生成方法,首先获取待处理文本信息集合,然后对待处理文本信息集合中的每个待处理文本信息进行分组处理,得到M个文本小组,每个文本小组中的待处理文本信息对应于相同的词语类别,再对目标文本小组进行分簇处理,得到文本结果簇,其中,目标文本小组为M个文本小组中的任意一个文本小组,每个文本结果簇包含至少一个的关键词,最后根据文本结果簇生成对应的第一事件。通过上述方式,在待处理文本信息集合包含大量待处理文本信息的情况下,能够根据词语类别对这些待处理文本信息先进行分组,然后进行聚类处理,避免了同时处理大量的待处理文本信息,同时提升了聚类的可信度,一方面节省了处理资源和计算资源,提升事件挖掘效

率,另一方面,对分组后的待处理文本信息进行聚类,可以有效地增强事件生成的准确度。

[0064] 可选地,在上述图3对应的实施例的基础上,本发明实施例提供的基于文本信息的事件生成方法第一个可选实施例中,对待处理文本信息集中的每个待处理文本信息进行分组处理,得到M个文本小组,可以包括:

[0065] 获取目标待处理文本信息,其中,目标待处理文本信息属于待处理文本信息集中的任意一个待处理文本信息;

[0066] 若目标待处理文本信息中包含实体信息,则将目标待处理文本信息划分至第一文本小组,其中,第一文本小组属于M个文本小组中的一个文本小组,实体信息用于表示一个固定信息。

[0067] 本实施例中,由于新闻数量巨大,为提高事件抽取的效率,本发明实施例对待处理文本信息(即新闻标题)进行了分组,输入的待处理文本信息可以为一周内的新闻标题。分组方法为依据待处理文本信息(新闻标题)中包含的实体信息,通名信息,地域信息,事件信息以及领域信息对待处理文本信息(新闻标题)进行分组,同一待处理文本信息(新闻标题)包含多个实体,可被分到多个文本小组中,即不同的文本小组可能包含相同的新闻内容,具体的文档分堆策略如下:

[0068] 以任意一个待处理文本信息的划分方式为例,假设该待处理文本信息为目标待处理文本信息,那么判断这个目标待处理文本信息中是否包含实体信息,如果包含,就直接将目标待处理文本信息划分到实体信息对应的第一文本小组中。其中,实体信息是指一个具体的事物,比如“腾讯科技有限公司”以及“北京”等,比如“A市外国语学校4名女生美术馆办画展”中包含了实体信息“A市外国语学校”,于是可以划分到第一文本小组中。

[0069] 第一文本小组中还可以包括子文本小组,每个子文本小组可以对应于具体的实体信息,比如第一个子文本小组是“A市外国语学校”相关新闻,第二个子文本小组是“腾讯”相关新闻,第三个子文本小组是“人物A”相关新闻,每个待处理文本信息通常包含至少一个实体信息,比如“人物A五十大寿人物B送对联事件相关新闻”中,既包含实体信息“人物A”,又包含实体信息“人物B”,因此,可能会被划分到两个不同的文本小组内。

[0070] 其次,本发明实施例中,在对待处理文本信息集中的每个待处理文本信息进行分组处理的过程中,首先,获取目标待处理文本信息,如果该目标待处理文本信息中包含实体信息,则将目标待处理文本信息划分至第一文本小组,第一文本小组属于M个文本小组中的一个文本小组,实体信息用于表示一个固定信息。通过上述方式,由于新闻数量巨大,为了提高事件抽取的效率,可以对待处理文本信息进行分类,从而得到多个不同的文本小组,再分别对每个文本小组中的待处理文本信息进行处理,这样可以降低数据处理量,从而提升数据处理效率。

[0071] 可选地,在上述图3对应的第一个实施例的基础上,本发明实施例提供的基于文本信息的事件生成方法第二个可选实施例中,获取目标待处理文本信息之后,还可以包括:

[0072] 若目标待处理文本信息中不包含实体信息,则判断目标待处理文本信息中是否包含第一词语组合或第二词语组合,其中,第一词语组合包含地域信息和通名信息,第二词语组合包含地域信息和领域信息,地域信息用于表示地点名称,通名信息用于表示一类固定信息,领域信息用于表示一个领域;

[0073] 若目标待处理文本信息中包含第一词语组合或第二词语组合,则将目标待处理文

本信息划分至第二文本小组,其中,第二文本小组属于M个文本小组中的一个文本小组。

[0074] 本实施例中,以任意一个待处理文本信息的划分方式为例,假设该待处理文本信息为目标待处理文本信息,那么判断这个目标待处理文本信息中是否包含实体信息,如果包含,就直接将目标待处理文本信息划分到实体信息对应的第一文本小组中。反之,如果目标待处理文本信息中不包含实体信息,那么需要判断目标待处理文本信息中是否包含第一词语组合或第二词语组合,这里的第一词语组合包含地域信息和通名信息,第二词语组合包含地域信息和领域信息,通名信息是指一类具体的事务,比如“老人”、“孕妇”、“教师”以及“警察”等。地域信息是指具体的地点名称,通常不一定包含在新闻的标题中,需要通过神经网络对标题进行地域识别才能得到。领域信息是指一个领域,比如“外卖”、“楼市”以及“通信”等。

[0075] 比如“A市高速上一小车自燃交警发布夏季防自燃攻略”中包含地域信息“A市”和通名信息“交警”,即包含第一词语组合,因此,可以划分到第二文本小组中。

[0076] 第二文本小组中还可以包括子文本小组,每个子文本小组可以对应于具体的第一词语组合或者第二词语组合,每个待处理文本信息通常包含至少一个第一词语组合或者第二词语组合,可能会被划分到两个不同的文本小组内。

[0077] 再次,本发明实施例中,若目标待处理文本信息中不包含实体信息,则判断目标待处理文本信息中是否包含第一词语组合或第二词语组合,若目标待处理文本信息中包含第一词语组合或第二词语组合,则将目标待处理文本信息划分至第二文本小组,其中,第二文本小组属于M个文本小组中的一个文本小组。通过上述方式,经过一系列的实验可以得到,在新闻标题中不包含实体信息的时候,优先以第一词语组合所包含的地域信息和通名信息,或者第二词语组合所包含的地域信息和领域信息为主,这样可以使得每个文本小组分到相对均衡的新闻标题,提升事件挖掘效率。

[0078] 可选地,在上述图3对应的第二个实施例的基础上,本发明实施例提供的基于文本信息的事件生成方法第三个可选实施例中,判断目标待处理文本信息中是否包含第一词语组合或第二词语组合之后,还可以包括:

[0079] 若目标待处理文本信息中不包含第一词语组合或第二词语组合,则判断目标待处理文本信息中是否包含地域信息;

[0080] 若目标待处理文本信息中包含地域信息,则将目标待处理文本信息划分至第三文本小组,其中,第三文本小组属于M个文本小组中的一个文本小组。

[0081] 本实施例中,以任意一个待处理文本信息的划分方式为例,假设该待处理文本信息为目标待处理文本信息,那么判断这个目标待处理文本信息中是否包含实体信息,如果不包含实体信息,则判断目标待处理文本信息中是否包含第一词语组合或第二词语组合,如果也不包含第一词语组合或第二词语组合,则继续判断是否包含地域信息,如果包含地域信息,那么就将目标待处理文本信息加入到第三文本小组。

[0082] 但是一些情况下地域信息是指具体的地点名称,通常不一定包含在待处理文本信息中,需要通过神经网络对待处理文本信息进行地域识别才能得到。比如“荔枝公园寻梅狂拍季开始啦”中包含地域信息“荔枝公园”,因此,可以划分到第三文本小组中。

[0083] 第三文本小组中还可以包括子文本小组,每个子文本小组可以对应于具体的地域信息,每个待处理文本信息通常包含至少一个地域信息,可能会被划分到不同的文本小组

内。

[0084] 进一步地,本发明实施例中,如果目标待处理文本信息中不包含第一词语组合或第二词语组合,则判断目标待处理文本信息中是否包含地域信息,如果目标待处理文本信息中包含地域信息,则将目标待处理文本信息划分至第三文本小组。通过上述方式,经过一系列的实验可以得到,在新闻标题中不包含实体信息、第一词语组合或者第二词语组合的时候,优先以地域信息为主,这样可以使得每个文本小组分到相对均衡的新闻标题,提升事件挖掘效率。

[0085] 可选地,在上述图3对应的第三个实施例的基础上,本发明实施例提供的基于文本信息的事件生成方法第四个可选实施例中,判断目标待处理文本信息中是否包含地域信息之后,还可以包括:

[0086] 若目标待处理文本信息中不包含地域信息,则判断目标待处理文本信息中是否包含第三词语组合或第四词语组合,其中,第三词语组合包含通名信息和事件信息,第四词语组合包含领域信息和事件信息,事件信息用于表示一个事件;

[0087] 若目标待处理文本信息中包含第三词语组合或第四词语组合,则将目标待处理文本信息划分至第四文本小组,其中,第四文本小组属于M个文本小组中的一个文本小组;

[0088] 若目标待处理文本信息中不包含第三词语组合或第四词语组合,则从待处理文本信息集合中剔除目标待处理文本信息。

[0089] 本实施例中,以任意一个待处理文本信息的划分方式为例,假设该待处理文本信息为目标待处理文本信息,那么判断这个目标待处理文本信息中是否包含实体信息,如果不包含实体信息,则判断目标待处理文本信息中是否包含第一词语组合或第二词语组合,如果也不包含第一词语组合或第二词语组合,则继续判断是否包含地域信息,如果不包含地域信息,那么再判断是否包含第三词语组合或第四词语组合,如果包含第三词语组合或第四词语组合,那么可以将目标待处理文本信息划分至第四文本小组,如果不包含第三词语组合或第四词语组合,则从待处理文本信息集合中剔除目标待处理文本信息。

[0090] 其中,第三词语组合包含通名信息和事件信息,第四词语组合包含领域信息和事件信息。比如“股市持续下跌偶然上涨各界褒贬不一”中包含领域信息“股市”和事件信息“下跌”以及“上涨”,即包含第四词语组合,因此,可以划分到第四文本小组中。

[0091] 第四文本小组中还可以包括子文本小组,每个子文本小组可以对应于具体的第三词语组合或者第四词语组合,每个待处理文本信息通常包含至少一个第三词语组合或者至少一个第四词语组合,可能会被划分到不同的文本小组内。

[0092] 经过分组后,每组的待处理文本信息个数在100至2500左右不等,大大提高了事件挖掘的效率,同时,相同类型但不同实体信息的待处理文本信息也被分到了不同的文本小组中,而包含同一个实体信息的不同的待处理文本信息会被分到同一个文本小组中。

[0093] 更进一步地,本发明实施例中,如果目标待处理文本信息中包含第三词语组合或第四词语组合,则将目标待处理文本信息划分至第四文本小组,如果目标待处理文本信息中不包含第三词语组合或第四词语组合,则从待处理文本信息集合中剔除目标待处理文本信息。通过上述方式,经过一系列的实验可以得到,在新闻标题中不包含实体信息、第一词语组合、第二词语组合或者地域信息的时候,优先以第三词语组合所包含的通名信息和事件信息,或者第四词语组合所包含的领域信息和事件信息为主,这样可以使得每个文本小

组分到相对均衡的新闻标题,提升事件挖掘效率。

[0094] 可选地,在上述图3对应的实施例的基础上,本发明实施例提供的基于文本信息的事件生成方法第五个可选实施例中,对目标文本小组进行分簇处理,得到文本结果簇,可以包括:

[0095] 从目标文本小组中获取待选关键词集合,其中,待选关键词集合中包含至少一个待选关键词;

[0096] 根据目标文本小组中的待处理文本信息数量,计算得到关键词个数;

[0097] 按照关键词个数从待选择关键词集合中选择P个目标关键词,其中,P为大于或等于1的整数,且P表示关键词个数;

[0098] 根据P个目标关键词生成至少一个文本结果簇。

[0099] 本实施例中,将以M个文本小组中的任意一个文本小组为例进行分簇介绍,假设该文本小组为目标文本小组。在目标文本小组包含多个待处理文本信息,即多个新闻标题。首先需要从目标文本小组中获取待选关键词集合,可以采用术语频率逆文档频率(term frequency inverse document frequency,TF-IDF)和/或文本排序(text rank)两种方法对目标文本小组中待处理文本信息的单词计算权重,从而得到待选关键词集合。

[0100] 可以理解的是,TF-IDF是一种统计方法,用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加,但同时会随着它在语料库中出现的频率成反比下降。TF-IDF的主要思想是,如果某个词或短语在一篇文章中出现的频率高,并且在其他文章中很少出现,则认为此词或者短语具有很好的类别区分能力,适合用来分类。

[0101] textrank是一种用来做关键词提取的算法,也可以用于提取短语和自动摘要。若待处理文本信息中存在若干个目标关键词相邻的情况,那么这些目标关键词可以构成一个关键短语。例如,在一篇介绍“支持向量机”的文章中,可以找到三个关键词,分别为支持、向量、机,通过关键短语提取,可以得到支持向量机。将每个待处理文本信息看成一个节点,若两个待处理文本信息之间有相似性,认为对应的两个节点之间有一个无向有权边,权值是相似度,从而计算得到的重要性最高的若干待处理文本信息可以当作摘要。

[0102] 权重计算方法经常会和余弦相似度共同使用于向量空间模型中,用以判断两份文件之间的相似性。

[0103] 但通常情况下待选关键词集合所包含的待选关键词数量非常多,所以还需要从待选关键词集合中再选择出权重在前P位的目标关键词。假设待选关键词集合有200个,然后根据目标文本小组中待处理文本信息的数量,计算后续需要获取的关键词个数。对关键词个数的计算,可以通过对待处理文本信息的数量和事件数量的分布数据分析得到。假设事件数量和待处理文本信息的数量满足幂率分布:

[0104] $Y \propto \beta X^{-r}$;

[0105] Y表示一个事件包括的关键词个数, β 表示一个系数,X表示待处理文本信息的数量,r表示表示经验值。假设每个事件平均由6个关键词描述(考虑到同义词),则r取3,得到关键词个数的经验取值为:

[0106] $P \propto 7N^{-3}$;

[0107] 其中,P表示关键词个数,N表示待处理文本信息的数量。

[0108] 假设P为5,则从待选择关键词集合中选择权重排序在前5位的关键词作为目标关键词。然后,对每个目标关键词构建一个特征向量,这个特征向量具有一个性质,两个向量的余弦相似度描述了两个目标关键词的共现程度。余弦相似度是通过计算两个向量的夹角余弦值来评估它们之间的相似度,余弦相似度将向量根据坐标值绘制到向量空间中,如最常见的二维空间。

[0109] 最后根据这5个目标关键词的聚类情况生成至少一个文本结果簇,每个文本结果簇对应至少一个目标关键词。

[0110] 需要说明的是,针对一个文本小组,目标关键词来自于待处理文本信息。针对目标关键词有以下两个假设,目标关键词列表能唯一确定这个事件,且目标关键词的特点是相互之间共现比较强,不同事件关键词共现较弱,频次高的单词不一定是目标关键词。

[0111] 其次,本发明实施例中,介绍了一种对目标文本小组进行分簇处理,得到文本结果簇的方法,首先从目标文本小组中获取待选关键词集合,然后根据目标文本小组中的待处理文本信息数量,计算得到关键词个数,再按照关键词个数从待选择关键词集合中选择P个目标关键词,最后根据P个目标关键词生成至少一个文本结果簇。通过上述方式,对关键词聚类并提取事件,且聚类关键词的特征向量依据待处理文本信息生成,从而提升了关键词聚类的可靠性。

[0112] 可选地,在上述图3对应的第五个实施例的基础上,本发明实施例提供的基于文本信息的事件生成方法第六个可选实施例中,根据P个目标关键词生成至少一个文本结果簇,可以包括:

[0113] 根据P个目标关键词生成P个特征向量,其中,每个目标关键词对应于一个特征向量,每个特征向量中包含特征元素,特征元素用于表示目标关键词在待处理文本信息中的出现情况;

[0114] 根据P个特征向量确定聚类个数;

[0115] 根据聚类个数生成所对应的文本结果簇。

[0116] 本实施例中,介绍了一种文本结果簇的方法,首先,根据P个目标关键词生成P个特征向量,假设P为5,即存在5个目标关键词,且对应于5组关键词的特征向量。每个特征向量中包含至少一个特征元素。

[0117] 为了便于介绍,请参阅表1,表1为目标文本小组中待处理文本信息的一个示意。

[0118] 表1

[0119]	序号	待处理文本信息
	1	北京打掉六个“号贩子”团伙
[0120]	2	北京警方打掉6个医院“号贩子”团伙已发拘留54人
	3	北京警方打掉一号贩子团伙 号贩子两月抢700多专家号
	4	北京警察“四管齐下”教育逆行女司机
	5	这个北京警察火了!教育逆行女司机这口才
	6	北京交警教育逆行女司机为何会火

[0121] 基于表1所示的目标文本小组中的待处理文本信息,假设目标关键词有5个,请参

阅表2,表2是构建的关键词特征向量示意。

[0122] 表2

[0123]

目标关键词	1号文本	2号文本	3号文本	4号文本	5号文本	6号文本
贩子	1	1	1	0	0	0
打掉	1	1	1	0	0	0
警察	1	1	1	1	1	1
女司机	0	0	0	1	1	1
逆行	0	0	0	1	1	1

[0124] 可以理解的是,上述目标文本小组中待处理文本信息的数量仅为一个示意,在实际应用中,一个文本小组中待处理文本信息的数量可以在100至2000。而目标关键词的数量也可以是其他数值,此处仅为一个示意。

[0125] 每个目标关键词生成一个特征向量,比如“贩子”的特征向量为“111000”,“女司机”的特征向量为“000111”。特征向量的维度为待处理文本信息的数量,每个特征向量的维度对应一个待处理文本信息,当目标关键词在待处理文本信息中出现时,特性向量在该维度的值为1,反之为0。“1”和“0”均为特征元素。

[0126] 最后基于目标关键词的特征向量,对P个目标关键词进行无监督聚类,从而生成至少一个文本结果簇。其中,无监督聚类的方法可以是硬聚类K-means算法,它是数据点到原型的某种距离作为优化的目标函数,利用函数求极值的方法得到迭代运算的调整规则。K-means算法以欧式距离作为相似度测度,它是求对应某一初始聚类中心向量最优分类,使得评价指标最小,算法采用误差平方和准则函数作为聚类准则函数。

[0127] 再次,本发明实施例中,提供了一种获取文本结果簇的方法,首先根据P个目标关键词生成P个特征向量,然后根据P个特征向量确定聚类个数,最后根据聚类个数生成所对应的文本结果簇。通过上述方式,能够根据关键词的特征向量确定聚类个数,从而提升方案的可行性。

[0128] 可选地,在上述图3对应的第六个实施例的基础上,本发明实施例提供的基于文本信息的事件生成方法第七个可选实施例中,根据P个特征向量确定聚类个数,可以包括:

[0129] 获取预设聚类个数;

[0130] 根据预设聚类个数以及P个特征向量,计算得到轮廓系数;

[0131] 根据预设聚类个数计算各个预设聚类簇之间的平均相似度;

[0132] 根据预设聚类个数计算每个预设聚类簇内待处理文本信息的相似度;

[0133] 根据各个预设聚类簇之间的平均相似度,以及每个预设聚类簇内待处理文本信息的相似度,计算得到离散系数;

[0134] 根据轮廓系数以及离散系数,计算簇数得分;

[0135] 若簇数得分满足簇数选取条件,则将预设聚类个数确定为聚类个数。

[0136] 本实施例中,介绍了一种确定聚类个数的具体方式,基于目标关键词的特征向量,对P个关键词进行无监督聚类,无监督聚类的目的在于把相似的东西聚在一起,一个聚类算法通常只需要知道如何计算相似度就可以开始工作了。聚类算法一般有五种方法,最主要的是划分方法和层次方法两种。划分聚类算法通过优化评价函数把数据集分割为A个部分,需要A作为输入参数。典型的分割聚类算法有K-means聚类算法,K-medoids聚类算法以及基

于随机选择的聚类 (CLARANS) 算法。层次聚类由不同层次的分割聚类组成, 层次之间的分割具有嵌套的关系。它不需要输入参数, 典型的分层聚类算法有平衡迭代削减聚类 (BIRCH) 算法、基于密度的聚类 (DBSCAN) 算法和使用代表点的聚类 (CURE) 算法等。

[0137] 下面以K-means聚类算法为例进行说明。聚类过程中尝试多个预设聚类个数的结果, 对于任意的预设聚类个数, 聚类结果的每个预设聚类簇中包含一个或多个目标关键词。对每个待处理文本信息而言, 通过选取与待处理文本信息待处理文本信息相似度最高的簇作为预设聚类簇, 待处理文本信息与预设聚类簇的相似度与待处理文本信息包含预设聚类簇中目标关键词的数目成正比。通过计算每个预设聚类簇中目标关键词在类中和类内的分布, 以及待处理文本信息的情况, 为每个预设聚类个数打分, 然后选取簇数得分最高的预设聚类个数, 从而动态的选取最优的预设聚类个数作为最终的聚类个数。

[0138] 对于设定的预设聚类个数, 首先需要计算该预设聚类个数所对应的轮廓系数, 此外, 还需要计算预设聚类个数所对应的离散系数, 结合轮廓系数和离散系数可以得到簇数得分, 簇数得分越高, 表明所设定的预设聚类个数越贴近最优聚类个数。

[0139] 进一步地, 本发明实施例中, 在确定聚类个数的时候需要采用假设法, 也就是先设定预设聚类个数, 然后计算预设聚类个数所对应的轮廓系数和离散系数, 最后根据轮廓系数和离散系数计算得到簇数得分。如果簇数得分满足簇数选取条件, 则将预设聚类个数确定为聚类个数, 反之, 如果不满足簇数选取条件, 则可以对下一个预设聚类个数进行判断。通过上述方式, 为确定聚类个数提供了具体的实现方式, 考虑到不同预设聚类个数的簇数得分情况, 能够选择更加合理的聚类个数, 从而提升方案的可靠性。

[0140] 可选地, 在上述图3对应的第七个实施例的基础上, 本发明实施例提供的基于文本信息的事件生成方法第八个可选实施例中, 根据预设聚类个数以及P个特征向量, 计算得到轮廓系数, 可以包括:

[0141] 采用如下方式计算轮廓系数:

$$[0142] \quad score_k^{sc} = \sum_{i=0}^P \frac{b(k, i) - a(k, i)}{\max(b(k, i), a(k, i))};$$

[0143] 其中, $score_k^{sc}$ 表示在预设聚类个数下得到的轮廓系数, P表示目标关键词的数目, k表示预设聚类个数, i表示第i个特征向量, a(k, i)表示在预设聚类个数为k时第i个特征向量到同一簇内其他特征向量的平均距离, b(k, i)表示在预设聚类个数为k时第i个特征向量到其他簇所有特征向量的平均距离;

[0144] 根据预设聚类个数计算各个预设聚类簇之间的平均相似度, 包括:

[0145] 采用如下方式计算各个预设聚类簇之间的平均相似度:

$$[0146] \quad \tilde{s}_i = \frac{\sum_{j=0}^{P_i} Len(\sum_{l=0}^{l=P, l \neq i} set(keyword_l)) \cap set(sentence_j)}{P_i};$$

[0147] 其中, $\tilde{s}_i$ 表示第i个预设聚类簇与其他预设聚类簇之间的平均相似度, P_i 表示第i个预设聚类簇中待处理文本信息的数目, $set(keyword_l)$ 表示第i个预设聚类簇的关键词列表, $set(sentence_j)$ 表示第i个预设聚类簇的第j个待处理文本信息的词语列表, $\cap$ 表示词语列表的交集, $Len()$ 表示词语列表中的词语个数;

[0148] 根据预设聚类个数计算每个预设聚类簇内待处理文本信息的相似度,可以包括:

[0149] 采用如下方式计算预设聚类簇内待处理文本信息的相似度:

$$[0150] \quad s_i = \frac{\sum_{j=0}^{n_i} \text{Len}(\text{set}(\text{keyword}_i) \cap \text{set}(\text{sentence}_j))}{n_i};$$

[0151] 其中, s_i 表示第 i 个预设聚类簇待处理文本信息的相似度;

[0152] 根据各个预设聚类簇之间的平均相似度,以及每个预设聚类簇内待处理文本信息的相似度,计算得到离散系数,包括:

[0153] 采用如下方式计算离散系数:

$$[0154] \quad \text{score}_k^{\text{doc}} = \frac{\sum_{i=0}^{i=k} s_i}{\sum_{i=0}^{i=k} \tilde{s}_i};$$

[0155] 其中, $\text{score}_k^{\text{doc}}$ 表示在预设聚类个数下得到的离散系数;

[0156] 根据轮廓系数以及离散系数,计算簇数得分,可以包括:

[0157] 采用如下方式计算簇数得分:

$$[0158] \quad \text{score}_k = \text{score}_k^{\text{sc}} + \text{score}_k^{\text{doc}};$$

[0159] 其中, score_k 表示在预设聚类个数下得到的簇数得分。

[0160] 本实施例中,通过下面介绍当方法动态选取最优的预设聚类个数作为最终的聚类个数。

[0161] 具体地,可以计算目标关键词聚集和离散程度的轮廓系数。即采用如下方式计算轮廓系数:

$$[0162] \quad \text{score}_k^{\text{sc}} = \sum_{i=0}^P \frac{b(k,i) - a(k,i)}{\max(b(k,i), a(k,i))};$$

[0163] 其中, $\text{score}_k^{\text{sc}}$ 表示在预设聚类个数下得到的轮廓系数, P 表示目标关键词的数目, k 表示预设聚类个数, i 表示第 i 个特征向量, $a(k,i)$ 表示在预设聚类个数为 k 时第 i 个特征向量到同一簇内其他特征向量的平均距离, $b(k,i)$ 表示在预设聚类个数为 k 时第 i 个特征向量到其他簇所有特征向量的平均距离;

[0164] 可以计算待处理文本信息质量的匹配分数,即先采用如下方式计算各个预设聚类簇之间的平均相似度:

$$[0165] \quad \tilde{s}_i = \frac{\sum_{j=0}^{P_i} \text{Len}(\sum_{l=0}^{l=P, l \neq i} \text{set}(\text{keyword}_l)) \cap \text{set}(\text{sentence}_j))}{P_i};$$

[0166] 其中, $\tilde{s}_i$ 表示第 i 个预设聚类簇与其他预设聚类簇之间的平均相似度, P_i 表示第 i 个预设聚类簇中待处理文本信息的数目, $\text{set}(\text{keyword}_i)$ 表示第 i 个预设聚类簇的关键词列表, $\text{set}(\text{sentence}_j)$ 表示第 i 个预设聚类簇的第 j 个待处理文本信息的词语列表, $\cap$ 表示词语列表的交集, $\text{Len}()$ 表示词语列表中的词语个数;

[0167] 然后采用如下方式计算预设聚类簇内待处理文本信息的相似度：

$$[0168] \quad s_i = \frac{\sum_{j=0}^{n_i} \text{Len}(\text{set}(\text{keyword}_i) \cap \text{set}(\text{sentence}_j))}{n_i};$$

[0169] 其中, s_i 表示第 i 个预设聚类簇待处理文本信息的相似度；

[0170] 最后, 结合各个预设聚类簇之间的平均相似度, 以及每个预设聚类簇内待处理文本信息的相似度, 可以计算离散系数, 即采用如下方式计算离散系数：

$$[0171] \quad \text{score}_k^{\text{doc}} = \frac{\sum_{i=0}^{i=k} s_i}{\sum_{i=0}^{i=k} \tilde{s}_i};$$

[0172] 其中, $\text{score}_k^{\text{doc}}$ 表示在预设聚类个数下得到的离散系数。

[0173] 将轮廓系数以及离散系数加和, 从而得到簇数得分, 即采用如下方式计算簇数得分：

$$[0174] \quad \text{score}_k = \text{score}_k^{\text{sc}} + \text{score}_k^{\text{doc}};$$

[0175] 其中, score_k 表示在预设聚类个数下得到的簇数得分。

[0176] 更进一步地, 本发明实施例中, 提供了计算轮廓系数的具体方式, 计算离散系数的具体方式, 以及计算簇数得分的具体方式。通过上述方式, 能够为方案的实现提供具体依据, 从而保证方案的可行性和可操作性。

[0177] 可选地, 在上述图3对应的第六个实施例的基础上, 本发明实施例提供的基于文本信息的事件生成方法第九个可选实施例中, 根据 P 个特征向量确定聚类个数, 可以包括：

[0178] 根据 P 个特征向量生成 Q 个待选择文本簇, 其中, 每个待选择文本簇中包含至少一个待处理文本信息, Q 为大于或等于 1 的整数；

[0179] 1) 获取 Q 个待选择文本簇中一个待选择文本簇的待处理文本信息数量；

[0180] 2) 若待处理文本信息数量大于或等于数量阈值, 则在上一个聚类计数值上增加 1, 得到本次聚类计数值；

[0181] 重复执行步骤 1) 和步骤 2), 直至 Q 个待选择文本簇中的待处理文本信息数量均与数量阈值比对完成, 根据本次聚类计数值确定聚类个数。

[0182] 本实施例中, 介绍了另一种确定聚类个数的具体方式。除了可以采用 K-means 聚类算法, 还可以采用其他距离方式, 比如基于密度的聚类 (DBSCAN) 算法。使用密度聚类时将直接得到聚类结果, 不需要选择预设聚类个数, 使用 DBSCAN 算法时直接得到聚类结果, 即 Q 个待选择文本簇。假设得到 3 个待选择文本簇, 且每个待选择文本簇中包含至少一个待处理文本信息, 假设第一个待选择文本簇中待处理文本信息的数量为 35 个, 第二个待选择文本簇中待处理文本信息的数量为 50 个, 第三个待选择文本簇中待处理文本信息的数量为 8 个, 且假设数量阈值为 10 个。于是依次对待选择文本簇进行选择, 先选择第一个待选择文本簇, 35 个待处理文本信息的数量大于数量阈值, 因此, 第一个待选择文本簇满足条件, 那么聚类计数值由初始值 0 增加 1, 得到本次聚类计数值为 1。然后, 选择下一个待选择文本簇, 如第二个待选择文本簇, 显然, 50 个待处理文本信息的数量也大于数量阈值, 因此, 第二个待选择文本簇也满足条件, 那么聚类计数值在 1 的基础上再增加 1, 得到本次聚类计数值为 2。最后, 选

择第三个待选择文本簇,8个待处理文本信息的数量小于数量阈值,因此,第三个待选择文本簇不满足条件,那么则不对本次聚类计数值做任何处理。最后得到的本次聚类计数值为2,也就是说,聚类个数可以设定为2。

[0183] 进一步地,本发明实施例中,介绍了另一种确定聚类个数的方式,除了假设法设定预设聚类个数以外,还可以生成若干个待选择文本簇,然后依次对每个待选择文本簇进行筛选,如果待处理文本信息数量大于或等于数量阈值,那认为是合格的待选择文本簇。通过上述方式,为聚类个数的确定提供了另一个可行的方式,从而提升了方案的可行性和可操作性。且该方式也具有较高的可信度,由此增加了方案的灵活性和实用性。

[0184] 可选地,在上述图3以及图3对应的第一个至第九个实施例中任一项的基础上,本发明实施例提供的基于文本信息的事件生成方法第十个可选实施例中,根据文本结果簇生成对应的第一事件之后,还可以包括:

[0185] 获取第二事件;

[0186] 根据第一事件和第二事件计算事件相似度,其中,第一事件对应于至少一个第一文本信息,第二事件对应于至少一个第二文本信息;

[0187] 若相似度大于或等于预设门限,则确定第二事件与第一事件为相似事件,并将第二事件与第一事件进行合并;

[0188] 若相似度小于预设门限,则为第二事件分配标识,其中,不同的标识用于表示不同的事件。

[0189] 本实施例中,在事件生成装置生成了第一事件之后,继续生成第二事件,因此,这里的第一事件即为历史事件,第二事件为新加入的事件,每个事件对应于多个文本信息,比如第一事件为“国庆节全国人民观看阅兵仪式的相关新闻”,第一事件中包括200条相关文本信息,即第一文本信息。第二事件为“国庆节天安门前万人观看升旗的相关新闻”,第二事件中包括300条相关文本信息,即第二文本信息。为了便于理解,请参阅图4,图4为本发明实施例中增量更新事件的一个实施例示意图,如图所示,由于事件是根据动态新闻流生成的,因此,前后两批数据之间可能会有部分文本信息相同,即第一文本信息和第二文本信息有部分重合,基动态增量新闻流生成的新事件可能已经在历史事件中存在,需要对新生成的事件进行融合并去重。

[0190] 请参阅图5,图5为本发明实施例中事件融合的一个实施例示意图,如图所示,第一事件与第二事件中会存在一部分重叠的文本信息,即图5中的阴影部分,这部分重叠区域即可以表示为第一事件和第二事件计算事件相似度,如果该相似度大于或等于预设门限,那么可以认为第一事件和第二事件属于相似事件,也可以将这两个事件进行融合,从而分配同一个事件的标识,比如标识为10265。如果相似度小于预设门限,则对新加入的第二事件分配新的标识,假设历史事件已经有标识1至10265,那么第二事件的标识不同于1至10265中任意一个标识,比如为10266。

[0191] 再进一步地,本发明实施例中,事件生成装置根据文本结果簇生成对应的第一事件之后,还可以获取第二事件,然后根据第一事件和第二事件计算事件相似度,若相似度大于或等于预设门限,则确定第二事件与第一事件为相似事件,并将第二事件与第一事件进行合并,若相似度小于预设门限,则为第二事件分配标识,其中,不同的标识用于表示不同的事件。通过上述方式,能够根据文本信息的重合度对事件进行融合,从而避免相似度较大

的事件被重复存储,一方面有利于提升存储资源的利用率,另一方面,在展示事件的时候可以更准确的定位到与该事件相关的新闻内容。

[0192] 可选地,在上述图3对应的第十个实施例的基础上,本发明实施例提供的基于文本信息的事件生成方法第十一个可选实施例中,根据第一事件和第二事件计算事件相似度,可以包括:

[0193] 采用如下方式计算事件相似度:

$$[0194] \quad \alpha = \frac{|A \cap B|}{\min(|A|, |B|)};$$

[0195] 其中, α 表示事件相似度,A表示第一事件对应于第一文本信息的个数,B表示第二事件对应于第二文本信息的个数。

[0196] 本实施例中,介绍了一种计算两个事件相似度的方法,当两个事件的相似度大于或等于预设门限时,认为两个事件为同一事件,并可以将新的事件融合更新到历史的事件中,事件之间的相似度可以通过两个事件共有的文本信息个数和分别包含的文本信息个数计算得到,假设相似度的预设门限为60%,第一事件对应于第一文本信息的个数为200,第二事件对应于第二文本信息的个数为300,第一文本信息和第二文本信息重合的个数为100,则利用如下公式:

$$[0197] \quad \alpha = \frac{|A \cap B|}{\min(|A|, |B|)};$$

$$[0198] \quad \alpha = \frac{|100|}{\min(200, 300)}$$

$$[0199] \quad = 50\%$$

[0200] 50%小于预设门限60%,因此,认为第二事件与第一事件是不同事件,需要给第二事件分配一个新的标识。

[0201] 再进一步地,本发明实施例中,提供了一种计算事件相似度的方式,通过两个事件共有的文本信息个数和分别包含的文本信息个数,可以计算得到事件相似度。通过上述方式,为事件相似度的计算提供了一种可行的方法,从而提升了方案的可操作性和可行性。

[0202] 下面对本发明中的事件生成装置进行详细描述,请参阅图6,图6为本发明实施例中事件生成装置一个实施例示意图,事件生成装置20包括:

[0203] 获取模块201,用于获取待处理文本信息集合,其中,所述待处理文本信息集合包含N个待处理文本信息,所述N为大于或等于1的整数;

[0204] 处理模块202,用于对所述获取模块201获取的所述待处理文本信息集合中的每个待处理文本信息进行分组处理,得到M个文本小组,其中,每个文本小组中的待处理文本信息对应于相同的词语类别,所述M为大于或等于1的整数;

[0205] 所述处理模块202,还用于对目标文本小组进行分簇处理,得到文本结果簇,其中,所述目标文本小组为所述M个文本小组中的任意一个文本小组,每个文本结果簇包含至少一个的关键词;

[0206] 生成模块203,用于根据所述处理模块202处理得到的所述文本结果簇生成对应的第一事件。

[0207] 本实施例中,获取模块201获取待处理文本信息集合,其中,所述待处理文本信息集合包含N个待处理文本信息,所述N为大于或等于1的整数,处理模块202对所述获取模块201获取的所述待处理文本信息集合中的每个待处理文本信息进行分组处理,得到M个文本小组,其中,每个文本小组中的待处理文本信息对应于相同的词语类别,所述M为大于或等于1的整数,所述处理模块202,还用于对目标文本小组进行分簇处理,得到文本结果簇,其中,所述目标文本小组为所述M个文本小组中的任意一个文本小组,每个文本结果簇包含至少一个的关键词,生成模块203根据所述处理模块202处理得到的所述文本结果簇生成对应的第一事件。

[0208] 本发明实施例中,提供了一种事件生成装置,首先获取待处理文本信息集合,然后对待处理文本信息集合中的每个待处理文本信息进行分组处理,得到M个文本小组,每个文本小组中的待处理文本信息对应于相同的词语类别,再对目标文本小组进行分簇处理,得到文本结果簇,其中,目标文本小组为M个文本小组中的任意一个文本小组,每个文本结果簇包含至少一个的关键词,最后根据文本结果簇生成对应的第一事件。通过上述方式,在待处理文本信息集合包含大量待处理文本信息的情况下,能够根据词语类别对这些待处理文本信息先进行分组,然后进行聚类处理,避免了同时处理大量的待处理文本信息,同时提升了聚类的可信度,一方面节省了处理资源和计算资源,提升事件挖掘效率,另一方面,对分组后的待处理文本信息进行聚类,可以有效地增强事件生成的准确度。

[0209] 可选地,在上述图6所对应的实施例的基础上,本发明实施例提供的事件生成装置20的另一实施例中,

[0210] 所述处理模块202,具体用于获取目标待处理文本信息,其中,所述目标待处理文本信息属于所述待处理文本信息集合中的任意一个待处理文本信息;

[0211] 若所述目标待处理文本信息中包含实体信息,则将所述目标待处理文本信息划分至第一文本小组,其中,所述第一文本小组属于所述M个文本小组中的一个文本小组,所述实体信息用于表示一个固定信息。

[0212] 其次,本发明实施例中,在对待处理文本信息集合中的每个待处理文本信息进行分组处理的过程中,首先,获取目标待处理文本信息,如果该目标待处理文本信息中包含实体信息,则将目标待处理文本信息划分至第一文本小组,第一文本小组属于M个文本小组中的一个文本小组,实体信息用于表示一个固定信息。通过上述方式,由于新闻数量巨大,为了提高事件抽取的效率,可以对待处理文本信息进行分类,从而得到多个不同的文本小组,再分别对每个文本小组中的待处理文本信息进行处理,这样可以降低数据处理量,从而提升数据处理效率。

[0213] 可选地,在上述图6所对应的实施例的基础上,本发明实施例提供的事件生成装置20的另一实施例中,

[0214] 所述处理模块202,具体用于获取目标待处理文本信息之后,若所述目标待处理文本信息中不包含所述实体信息,则判断所述目标待处理文本信息中是否包含第一词语组合或第二词语组合,其中,所述第一词语组合包含地域信息和通名信息,所述第二词语组合包含地域信息和领域信息,所述地域信息用于表示地点名称,所述通名信息用于表示一类固

定信息,所述领域信息用于表示一个领域;

[0215] 若所述目标待处理文本信息中包含所述第一词语组合或所述第二词语组合,则将所述目标待处理文本信息划分至第二文本小组,其中,所述第二文本小组属于所述M个文本小组中的一个文本小组。

[0216] 再次,本发明实施例中,若目标待处理文本信息中不包含实体信息,则判断目标待处理文本信息中是否包含第一词语组合或第二词语组合,若目标待处理文本信息中包含第一词语组合或第二词语组合,则将目标待处理文本信息划分至第二文本小组,其中,第二文本小组属于M个文本小组中的一个文本小组。通过上述方式,经过一系列的实验可以得到,在新闻标题中不包含实体信息的时候,优先以第一词语组合所包含的地域信息和通名信息,或者第二词语组合所包含的地域信息和领域信息为主,这样可以使得每个文本小组分到相对均衡的新闻标题,提升事件挖掘效率。

[0217] 可选地,在上述图6所对应的实施例的基础上,本发明实施例提供的事件生成装置20的另一实施例中,

[0218] 所述处理模块202,具体用于判断所述目标待处理文本信息中是否包含第一词语组合或第二词语组合之后,若所述目标待处理文本信息中不包含所述第一词语组合或所述第二词语组合,则判断所述目标待处理文本信息中是否包含所述地域信息;

[0219] 若所述目标待处理文本信息中包含所述地域信息,则将所述目标待处理文本信息划分至第三文本小组,其中,所述第三文本小组属于所述M个文本小组中的一个文本小组。

[0220] 进一步地,本发明实施例中,如果目标待处理文本信息中不包含第一词语组合或第二词语组合,则判断目标待处理文本信息中是否包含地域信息,如果目标待处理文本信息中包含地域信息,则将目标待处理文本信息划分至第三文本小组。通过上述方式,经过一系列的实验可以得到,在新闻标题中不包含实体信息、第一词语组合或者第二词语组合的时候,优先以地域信息为主,这样可以使得每个文本小组分到相对均衡的新闻标题,提升事件挖掘效率。

[0221] 可选地,在上述图6所对应的实施例的基础上,本发明实施例提供的事件生成装置20的另一实施例中,

[0222] 所述处理模块202,具体用于判断所述目标待处理文本信息中是否包含所述地域信息之后,若所述目标待处理文本信息中不包含所述地域信息,则判断所述目标待处理文本信息中是否包含第三词语组合或第四词语组合,其中,所述第三词语组合包含所述通名信息和事件信息,所述第四词语组合包含所述领域信息和所述事件信息,所述所述事件信息用于表示一个事件;

[0223] 若所述目标待处理文本信息中包含所述第三词语组合或所述第四词语组合,则将所述目标待处理文本信息划分至第四文本小组,其中,所述第四文本小组属于所述M个文本小组中的一个文本小组;

[0224] 若所述目标待处理文本信息中不包含所述第三词语组合或所述第四词语组合,则从所述待处理文本信息集合中剔除所述目标待处理文本信息。

[0225] 更进一步地,本发明实施例中,如果目标待处理文本信息中包含第三词语组合或第四词语组合,则将目标待处理文本信息划分至第四文本小组,如果目标待处理文本信息中不包含第三词语组合或第四词语组合,则从待处理文本信息集合中剔除目标待处理文本

信息。通过上述方式,经过一系列的实验可以得到,在新闻标题中不包含实体信息、第一词语组合、第二词语组合或者地域信息的时候,优先以第三词语组合所包含的通名信息和事件信息,或者第四词语组合所包含的领域信息和事件信息为主,这样可以使得每个文本小组分到相对均衡的新闻标题,提升事件挖掘效率。

[0226] 可选地,在上述图6所对应的实施例的基础上,本发明实施例提供的事件生成装置20的另一实施例中,

[0227] 所述处理模块202,具体用于从所述目标文本小组中获取待选关键词集合,其中,所述待选关键词集合中包含至少一个待选关键词;

[0228] 根据所述目标文本小组中的待处理文本信息数量,计算得到关键词个数;

[0229] 按照所述关键词个数从所述待选择关键词集合中选择P个目标关键词,其中,所述P为大于或等于1的整数,且所述P表示所述关键词个数;

[0230] 根据所述P个目标关键词生成至少一个所述文本结果簇。

[0231] 其次,本发明实施例中,介绍了一种对目标文本小组进行分簇处理,得到文本结果簇的方法,首先从目标文本小组中获取待选关键词集合,然后根据目标文本小组中的待处理文本信息数量,计算得到关键词个数,再按照关键词个数从待选择关键词集合中选择P个目标关键词,最后根据P个目标关键词生成至少一个文本结果簇。通过上述方式,对关键词聚类并提取事件,且聚类关键词的特征向量依据待处理文本信息生成,从而提升了关键词聚类的可靠性。

[0232] 可选地,在上述图6所对应的实施例的基础上,本发明实施例提供的事件生成装置20的另一实施例中,

[0233] 所述处理模块202,具体用于根据所述P个目标关键词生成P个特征向量,其中,每个目标关键词对应于一个特征向量,每个特征向量中包含特征元素,所述特征元素用于表示目标关键词在待处理文本信息中的出现情况;

[0234] 根据所述P个特征向量确定聚类个数;

[0235] 根据所述聚类个数生成所对应的所述文本结果簇。

[0236] 再次,本发明实施例中,提供了一种获取文本结果簇的方法,首先根据P个目标关键词生成P个特征向量,然后根据P个特征向量确定聚类个数,最后根据聚类个数生成所对应的文本结果簇。通过上述方式,能够根据关键词的特征向量确定聚类个数,从而提升方案的可行性。

[0237] 可选地,在上述图6所对应的实施例的基础上,本发明实施例提供的事件生成装置20的另一实施例中,

[0238] 所述处理模块202,具体用于获取预设聚类个数;

[0239] 根据所述预设聚类个数以及所述P个特征向量,计算得到轮廓系数;

[0240] 根据所述预设聚类个数计算各个预设聚类簇之间的平均相似度;

[0241] 根据所述预设聚类个数计算每个所述预设聚类簇内待处理文本信息的相似度;

[0242] 根据所述各个预设聚类簇之间的平均相似度,以及每个预设聚类簇内待处理文本信息的相似度,计算得到离散系数;

[0243] 根据所述轮廓系数以及所述离散系数,计算簇数得分;

[0244] 若所述簇数得分满足簇数选取条件,则将所述预设聚类个数确定为所述聚类个

数。

[0245] 进一步地,本发明实施例中,在确定聚类个数的时候需要采用假设法,也就是先设定预设聚类个数,然后计算预设聚类个数所对应的轮廓系数和离散系数,最后根据轮廓系数和离散系数计算得到簇数得分。如果簇数得分满足簇数选取条件,则将预设聚类个数确定为聚类个数,反之,如果不满足簇数选取条件,则可以对下一个预设聚类个数进行判断。通过上述方式,为确定聚类个数提供了具体的实现方式,考虑到不同预设聚类个数的簇数得分情况,能够选择更加合理的聚类个数,从而提升方案的可靠性。

[0246] 可选地,在上述图6所对应的实施例的基础上,本发明实施例提供的事件生成装置20的另一实施例中,

[0247] 所述处理模块202,具体用于采用如下方式计算所述轮廓系数:

$$[0248] \quad score_k^{sc} = \sum_{i=0}^P \frac{b(k,i) - a(k,i)}{\max(b(k,i), a(k,i))};$$

[0249] 其中,所述 $score_k^{sc}$ 表示在所述预设聚类个数下得到的所述轮廓系数,所述P表示所述目标关键词的数目,所述k表示所述预设聚类个数,所述i表示第i个特征向量,所述a(k,i)表示在所述预设聚类个数为k时第i个特征向量到同一簇内其他特征向量的平均距离,所述b(k,i)表示在所述预设聚类个数为k时第i个特征向量到其他簇所有特征向量的平均距离;

[0250] 采用如下方式计算所述各个预设聚类簇之间的平均相似度:

$$[0251] \quad \tilde{s}_i = \frac{\sum_{j=0}^{P_i} Len(\sum_{l=0}^{l=P, l \neq i} set(keyword_l)) \cap set(sentence_j))}{P_i};$$

[0252] 其中,所述 $\tilde{s}_i$ 表示第i个预设聚类簇与其他预设聚类簇之间的平均相似度,所述 P_i 表示所述第i个预设聚类簇中待处理文本信息的数目,所述set(keyword_i)表示所述第i个预设聚类簇的关键词列表,所述set(sentence_j)表示所述第i个预设聚类簇的第j个待处理文本信息的词语列表,所述 $\cap$ 表示词语列表的交集,所述Len()表示所述词语列表中的词语个数;

[0253] 采用如下方式计算所述预设聚类簇内待处理文本信息的相似度:

$$[0254] \quad s_i = \frac{\sum_{j=0}^{n_i} Len(set(keyword_i) \cap set(sentence_j))}{n_i};$$

[0255] 其中,所述 s_i 表示所述第i个预设聚类簇待处理文本信息的相似度;

[0256] 采用如下方式计算所述离散系数:

$$[0257] \quad score_k^{doc} = \frac{\sum_{i=0}^{i=k} s_i}{\sum_{i=0}^{i=k} \tilde{s}_i};$$

[0258] 其中,所述 $score_k^{doc}$ 表示在所述预设聚类个数下得到的所述离散系数;

[0259] 所述根据所述轮廓系数以及所述离散系数,计算簇数得分,包括:

[0260] 采用如下方式计算所述簇数得分：

$$[0261] \quad score_k = score_k^{sc} + score_k^{doc};$$

[0262] 其中，所述 $score_k$ 表示表示在所述预设聚类个数下得到的所述簇数得分。

[0263] 更进一步地，本发明实施例中，提供了计算轮廓系数的具体方式，计算离散系数的具体方式，以及计算簇数得分的具体方式。通过上述方式，能够为方案的实现提供具体依据，从而保证方案的可行性和可操作性。

[0264] 可选地，在上述图6所对应的实施例的基础上，本发明实施例提供的事件生成装置20的另一实施例中，

[0265] 所述处理模块202，具体用于根据所述P个特征向量生成Q个待选择文本簇，其中，每个待选择文本簇中包含至少一个待处理文本信息，所述Q为大于或等于1的整数；

[0266] 1) 获取Q个待选择文本簇中一个待选择文本簇的待处理文本信息数量；

[0267] 2) 若所述待处理文本信息数量大于或等于数量阈值，则在上一个聚类计数值上增加1，得到本次聚类计数值；

[0268] 重复执行步骤1) 和步骤2)，直至所述Q个待选择文本簇中的待处理文本信息数量均与所述数量阈值比对完成，根据所述本次聚类计数值确定所述聚类个数。

[0269] 进一步地，本发明实施例中，介绍了另一种确定聚类个数的方式，除了假设法设定预设聚类个数以外，还可以生成若干个待选择文本簇，然后依次对每个待选择文本簇进行筛选，如果待处理文本信息数量大于或等于数量阈值，那认为是合格的待选择文本簇。通过上述方式，为聚类个数的确定提供了另一个可行的方式，从而提升了方案的可行性和可操作性。且该方式也具有较高的可信度，由此增加了方案的灵活性和实用性。

[0270] 可选地，在上述图6所对应的实施例的基础上，请参阅图7，本发明实施例提供的事件生成装置20的另一实施例中，所述事件生成装置20还包括计算模块204、合并模块205以及分配模块206；

[0271] 所述获取模块201，还用于所述生成模块203根据所述文本结果簇生成对应的第一事件之后，获取第二事件；

[0272] 所述计算模块204，用于根据所述获取模块201获取到的所述第一事件和所述第二事件计算事件相似度，其中，所述第一事件对应于至少一个第一文本信息，所述第二事件对应于至少一个第二文本信息；

[0273] 所述合并模块205，用于若所述计算模块204计算得到的所述相似度大于或等于预设门限，则确定所述第二事件与所述第一事件为相似事件，并将所述第二事件与所述第一事件进行合并；

[0274] 所述分配模块206，用于若所述计算模块204计算得到的所述相似度小于所述预设门限，则为所述第二事件分配标识，其中，不同的标识用于表示不同的事件。

[0275] 再进一步地，本发明实施例中，事件生成装置根据文本结果簇生成对应的第一事件之后，还可以获取第二事件，然后根据第一事件和第二事件计算事件相似度，若相似度大于或等于预设门限，则确定第二事件与第一事件为相似事件，并将第二事件与第一事件进行合并，若相似度小于预设门限，则为第二事件分配标识，其中，不同的标识用于表示不同的事件。通过上述方式，能够根据文本信息的重合度对事件进行融合，从而避免相似度较大

的事件被重复存储,一方面有利于提升存储资源的利用率,另一方面,在展示事件的时候可以更准确的定位到与该事件相关的新闻内容。

[0276] 可选地,在上述图7所对应的实施例的基础上,本发明实施例提供的事件生成装置20的另一实施例中,

[0277] 所述计算模块204,用于采用如下方式计算所述事件相似度:

$$[0278] \quad \alpha = \frac{|A \cap B|}{\min(|A|, |B|)};$$

[0279] 其中,所述 α 表示所述事件相似度,所述A表示所述第一事件对应于所述第一文本信息的个数,所述B表示所述第二事件对应于所述第二文本信息的个数。

[0280] 再进一步地,本发明实施例中,提供了一种计算事件相似度的方式,通过两个事件共有的文本信息个数和分别包含的文本信息个数,可以计算得到事件相似度。通过上述方式,为事件相似度的计算提供了一种可行的方法,从而提升了方案的可操作性和可行性。

[0281] 图8是本发明实施例提供的一种服务器结构示意图,该事件生成装置300可因配置或性能不同而产生比较大的差异,可以包括一个或一个以上中央处理器(central processing units,CPU)322(例如,一个或一个以上处理器)和存储器332,一个或一个以上存储应用程序342或数据344的存储介质330(例如一个或一个以上海量存储设备)。其中,存储器332和存储介质330可以是短暂存储或持久存储。存储在存储介质330的程序可以包括一个或一个以上模块(图示没标出),每个模块可以包括对服务器中的一系列指令操作。更进一步地,中央处理器322可以设置为与存储介质330通信,在事件生成装置300上执行存储介质330中的一系列指令操作。

[0282] 事件生成装置300还可以包括一个或一个以上电源326,一个或一个以上有线或无线网络接口350,一个或一个以上输入输出接口358,和/或,一个或一个以上操作系统341,例如Windows Server™,Mac OS X™,Unix™,Linux™,FreeBSD™等等。

[0283] 上述实施例中由服务器所执行的步骤可以基于该图8所示的服务器结构。

[0284] 本发明实施例中,CPU 322用于执行如下步骤:

[0285] 获取待处理文本信息集合,其中,所述待处理文本信息集合包含N个待处理文本信息,所述N为大于或等于1的整数;

[0286] 对所述待处理文本信息集合中的每个待处理文本信息进行分组处理,得到M个文本小组,其中,每个文本小组中的待处理文本信息对应于相同的词语类别,所述M为大于或等于1的整数;

[0287] 对目标文本小组进行分簇处理,得到文本结果簇,其中,所述目标文本小组为所述M个文本小组中的任意一个文本小组,每个文本结果簇包含至少一个的关键词;

[0288] 根据所述文本结果簇生成对应的第一事件。

[0289] 可选地,本发明实施例中CPU 322具体用于执行如下步骤:

[0290] 获取目标待处理文本信息,其中,所述目标待处理文本信息属于所述待处理文本信息集合中的任意一个待处理文本信息;

[0291] 若所述目标待处理文本信息中包含实体信息,则将所述目标待处理文本信息划分至第一文本小组,其中,所述第一文本小组属于所述M个文本小组中的一个文本小组,所述

实体信息用于表示一个固定信息。

[0292] 可选地,本发明实施例中CPU 322具体用于执行如下步骤:

[0293] 若所述目标待处理文本信息中不包含所述实体信息,则判断所述目标待处理文本信息中是否包含第一词语组合或第二词语组合,其中,所述第一词语组合包含地域信息和通名信息,所述第二词语组合包含地域信息和领域信息,所述地域信息用于表示地点名称,所述通名信息用于表示一类固定信息,所述领域信息用于表示一个领域;

[0294] 若所述目标待处理文本信息中包含所述第一词语组合或所述第二词语组合,则将所述目标待处理文本信息划分至第二文本小组,其中,所述第二文本小组属于所述M个文本小组中的一个文本小组。

[0295] 可选地,本发明实施例中CPU 322具体用于执行如下步骤:

[0296] 若所述目标待处理文本信息中不包含所述第一词语组合或所述第二词语组合,则判断所述目标待处理文本信息中是否包含所述地域信息;

[0297] 若所述目标待处理文本信息中包含所述地域信息,则将所述目标待处理文本信息划分至第三文本小组,其中,所述第三文本小组属于所述M个文本小组中的一个文本小组。

[0298] 可选地,本发明实施例中CPU 322具体用于执行如下步骤:

[0299] 若所述目标待处理文本信息中不包含所述地域信息,则判断所述目标待处理文本信息中是否包含第三词语组合或第四词语组合,其中,所述第三词语组合包含所述通名信息和事件信息,所述第四词语组合包含所述领域信息和所述事件信息,所述所述事件信息用于表示一个事件;

[0300] 若所述目标待处理文本信息中包含所述第三词语组合或所述第四词语组合,则将所述目标待处理文本信息划分至第四文本小组,其中,所述第四文本小组属于所述M个文本小组中的一个文本小组;

[0301] 若所述目标待处理文本信息中不包含所述第三词语组合或所述第四词语组合,则从所述待处理文本信息集合中剔除所述目标待处理文本信息。

[0302] 可选地,本发明实施例中CPU 322具体用于执行如下步骤:

[0303] 从所述目标文本小组中获取待选关键词集合,其中,所述待选关键词集合中包含至少一个待选关键词;

[0304] 根据所述目标文本小组中的待处理文本信息数量,计算得到关键词个数;

[0305] 按照所述关键词个数从所述待选择关键词集合中选择P个目标关键词,其中,所述P为大于或等于1的整数,且所述P表示所述关键词个数;

[0306] 根据所述P个目标关键词生成至少一个所述文本结果簇。

[0307] 可选地,本发明实施例中CPU 322具体用于执行如下步骤:

[0308] 根据所述P个目标关键词生成P个特征向量,其中,每个目标关键词对应于一个特征向量,每个特征向量中包含特征元素,所述特征元素用于表示目标关键词在待处理文本信息中的出现情况;

[0309] 根据所述P个特征向量确定聚类个数;

[0310] 根据所述聚类个数生成所对应的所述文本结果簇。

[0311] 可选地,本发明实施例中CPU 322具体用于执行如下步骤:

[0312] 获取预设聚类个数;

[0313] 根据所述预设聚类个数以及所述P个特征向量,计算得到轮廓系数;

[0314] 根据所述预设聚类个数计算各个预设聚类簇之间的平均相似度;

[0315] 根据所述预设聚类个数计算每个所述预设聚类簇内待处理文本信息的相似度;

[0316] 根据所述各个预设聚类簇之间的平均相似度,以及每个预设聚类簇内待处理文本信息的相似度,计算得到离散系数;

[0317] 根据所述轮廓系数以及所述离散系数,计算簇数得分;

[0318] 若所述簇数得分满足簇数选取条件,则将所述预设聚类个数确定为所述聚类个数。

[0319] 可选地,本发明实施例中CPU 322具体用于执行如下步骤:

[0320] 采用如下方式计算所述轮廓系数:

$$[0321] \quad score_k^{sc} = \sum_{i=0}^P \frac{b(k,i) - a(k,i)}{\max(b(k,i), a(k,i))};$$

[0322] 其中,所述 $score_k^{sc}$ 表示在所述预设聚类个数下得到的所述轮廓系数,所述P表示所述目标关键词的数目,所述k表示所述预设聚类个数,所述i表示第i个特征向量,所述a(k,i)表示在所述预设聚类个数为k时第i个特征向量到同一簇内其他特征向量的平均距离,所述b(k,i)表示在所述预设聚类个数为k时第i个特征向量到其他簇所有特征向量的平均距离;

[0323] 采用如下方式计算所述各个预设聚类簇之间的平均相似度:

$$[0324] \quad \tilde{s}_i = \frac{\sum_{j=0}^{P_i} Len(\sum_{l=0}^{l=P, l \neq i} set(keyword_l)) \cap set(sentence_j))}{P_i};$$

[0325] 其中,所述 $\tilde{s}_i$ 表示第i个预设聚类簇与其他预设聚类簇之间的平均相似度,所述 P_i 表示所述第i个预设聚类簇中待处理文本信息的数目,所述set(keyword_i)表示所述第i个预设聚类簇的关键词列表,所述set(sentence_j)表示所述第i个预设聚类簇的第j个待处理文本信息的词语列表,所述 $\cap$ 表示词语列表的交集,所述Len()表示所述词语列表中的词语个数;

[0326] 采用如下方式计算所述预设聚类簇内待处理文本信息的相似度:

$$[0327] \quad s_i = \frac{\sum_{j=0}^{n_i} Len(set(keyword_i) \cap set(sentence_j))}{n_i};$$

[0328] 其中,所述 s_i 表示所述第i个预设聚类簇待处理文本信息的相似度;

[0329] 采用如下方式计算所述离散系数:

$$[0330] \quad score_k^{doc} = \frac{\sum_{i=0}^{i=k} s_i}{\sum_{i=0}^{i=k} \tilde{s}_i};$$

[0331] 其中,所述 $score_k^{doc}$ 表示在所述预设聚类个数下得到的所述离散系数;

[0332] 所述根据所述轮廓系数以及所述离散系数,计算簇数得分,包括:

[0333] 采用如下方式计算所述簇数得分：

$$[0334] \quad score_k = score_k^{sc} + score_k^{doc};$$

[0335] 其中，所述 $score_k$ 表示表示在所述预设聚类个数下得到的所述簇数得分。

[0336] 可选地，本发明实施例中CPU 322具体用于执行如下步骤：

[0337] 根据所述P个特征向量生成Q个待选择文本簇，其中，每个待选择文本簇中包含至少一个待处理文本信息，所述Q为大于或等于1的整数；

[0338] 1) 获取Q个待选择文本簇中一个待选择文本簇的待处理文本信息数量；

[0339] 2) 若所述待处理文本信息数量大于或等于数量阈值，则在上一个聚类计数值上增加1，得到本次聚类计数值；

[0340] 重复执行步骤1) 和步骤2)，直至所述Q个待选择文本簇中的待处理文本信息数量均与所述数量阈值比对完成，根据所述本次聚类计数值确定所述聚类个数。

[0341] 可选地，本发明实施例中CPU 322还用于执行如下步骤：

[0342] 获取第二事件；

[0343] 根据所述第一事件和所述第二事件计算事件相似度，其中，所述第一事件对应于至少一个第一文本信息，所述第二事件对应于至少一个第二文本信息；

[0344] 若所述相似度大于或等于预设门限，则确定所述第二事件与所述第一事件为相似事件，并将所述第二事件与所述第一事件进行合并；

[0345] 若所述相似度小于所述预设门限，则为所述第二事件分配标识，其中，不同的标识用于表示不同的事件。

[0346] 可选地，本发明实施例中CPU 322具体用于执行如下步骤：

[0347] 采用如下方式计算所述事件相似度：

$$[0348] \quad \alpha = \frac{|A \cap B|}{\min(|A|, |B|)};$$

[0349] 其中，所述 α 表示所述事件相似度，所述A表示所述第一事件对应于所述第一文本信息的个数，所述B表示所述第二事件对应于所述第二文本信息的个数。

[0350] 所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统，装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

[0351] 在本申请所提供的几个实施例中，应该理解到，所揭露的系统，装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

[0352] 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单元来实现本实施例方案的目的。

[0353] 另外,在本发明各个实施例中的各功能单元可以集成在一个处理单元中,也可以是各个单元单独物理存在,也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现,也可以采用软件功能单元的形式实现。

[0354] 所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用,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全部或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。

[0355] 以上所述,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。

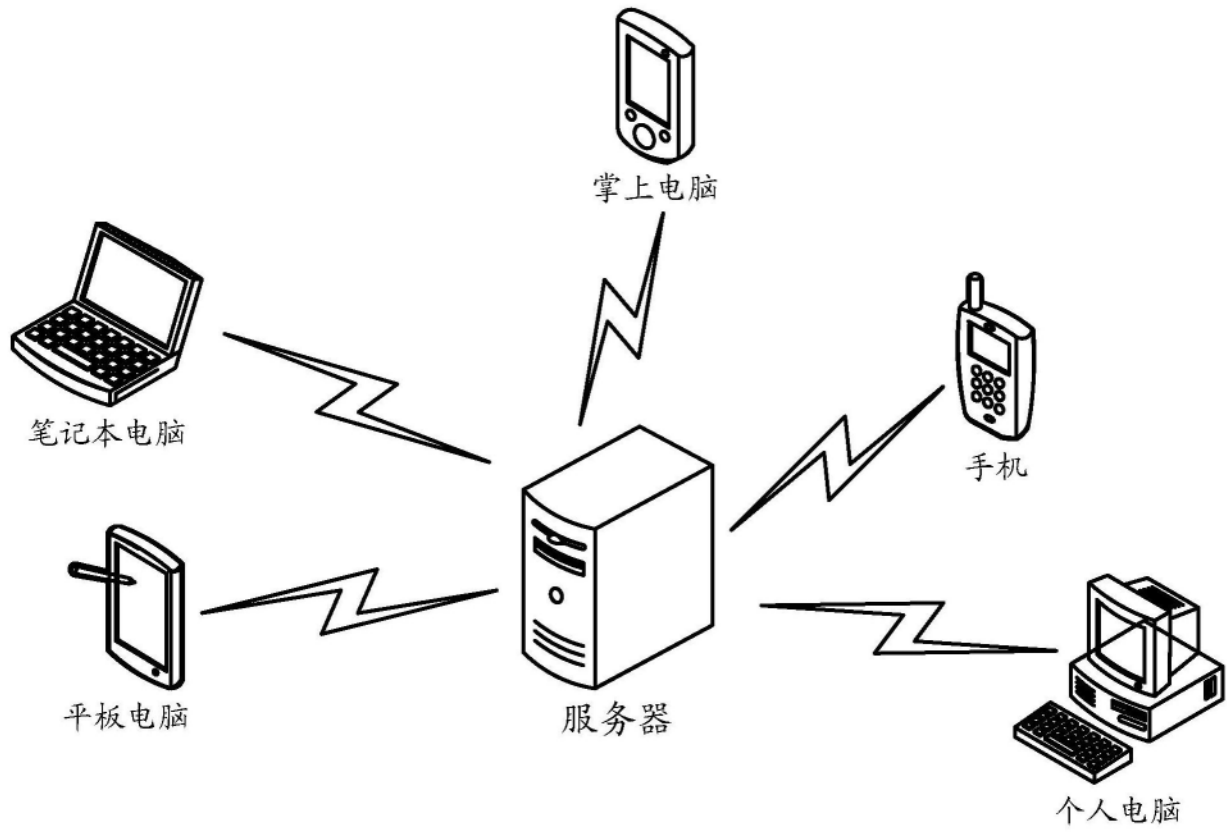


图1

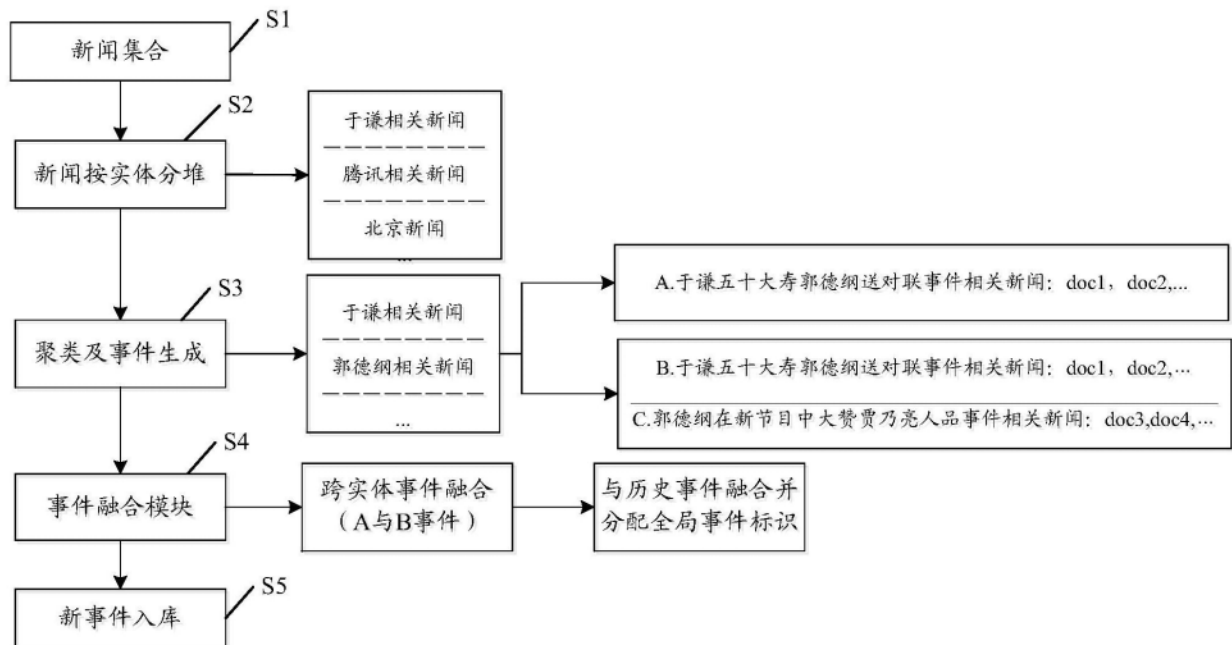


图2

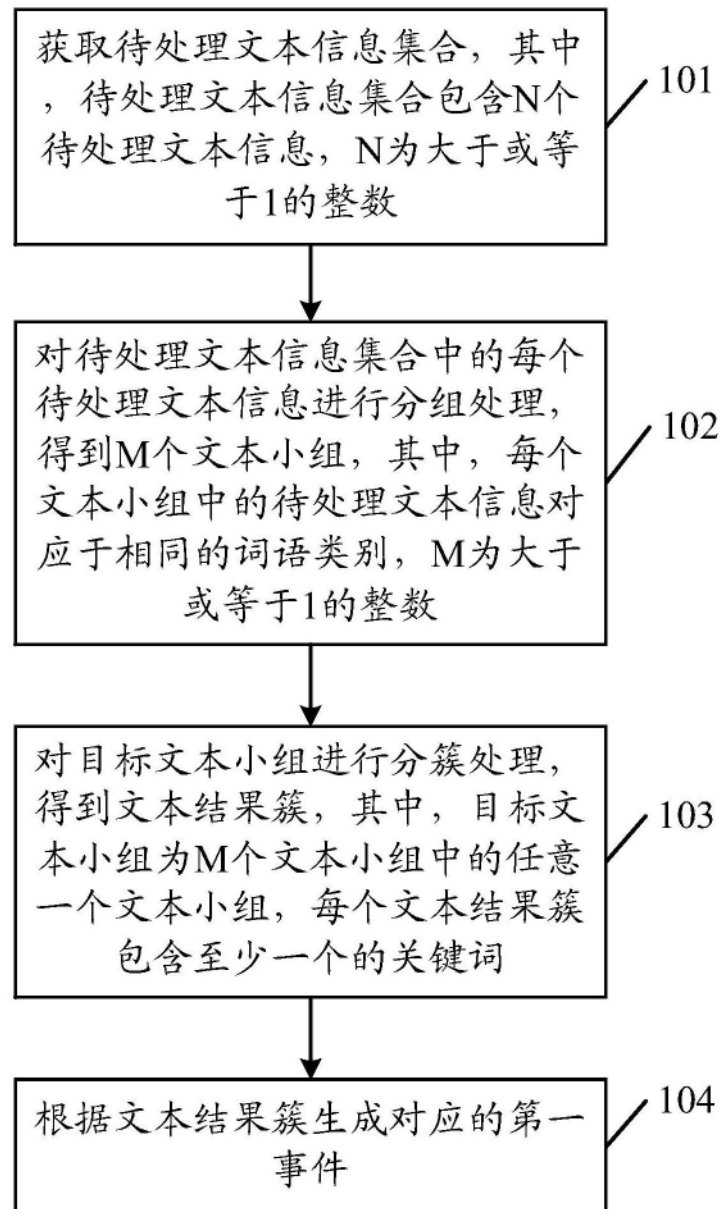


图3

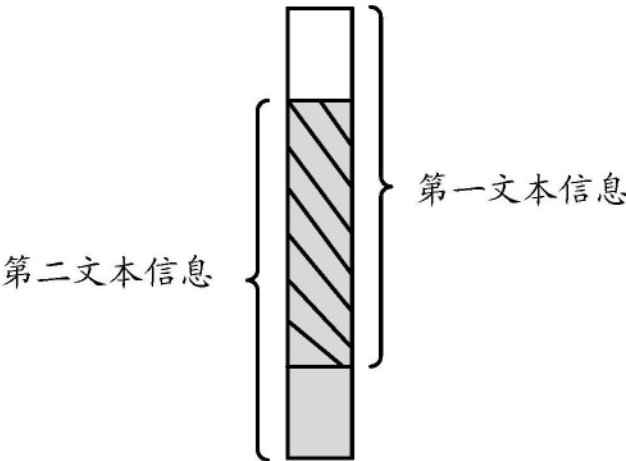


图4

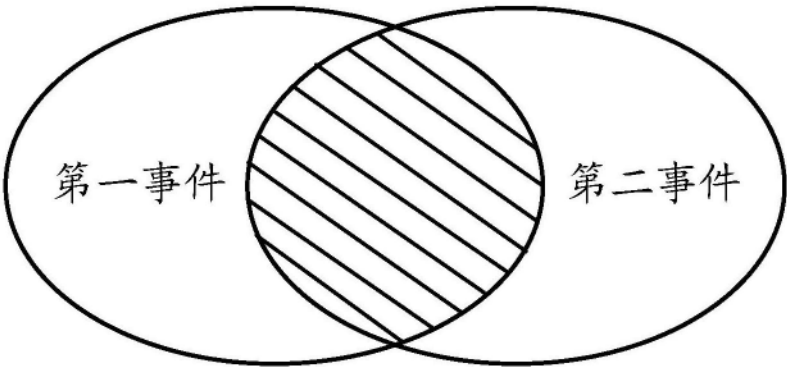


图5

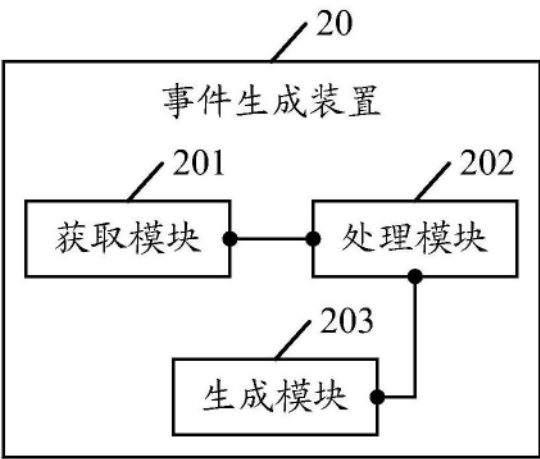


图6

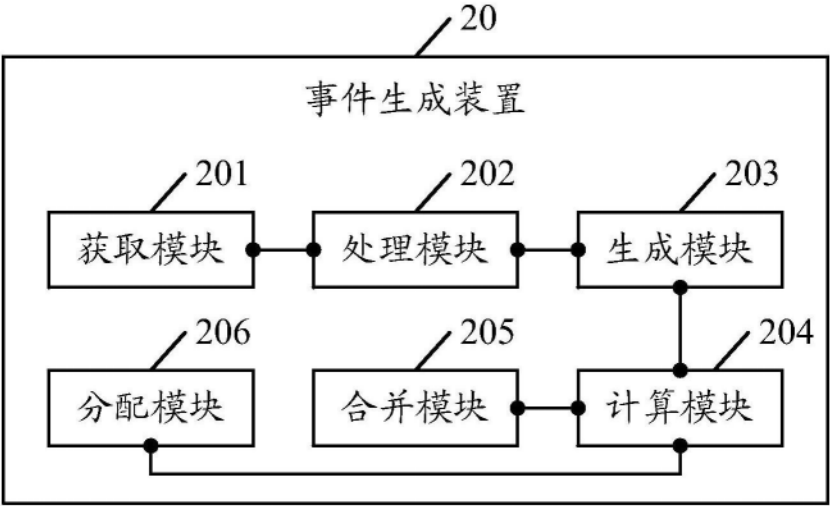


图7

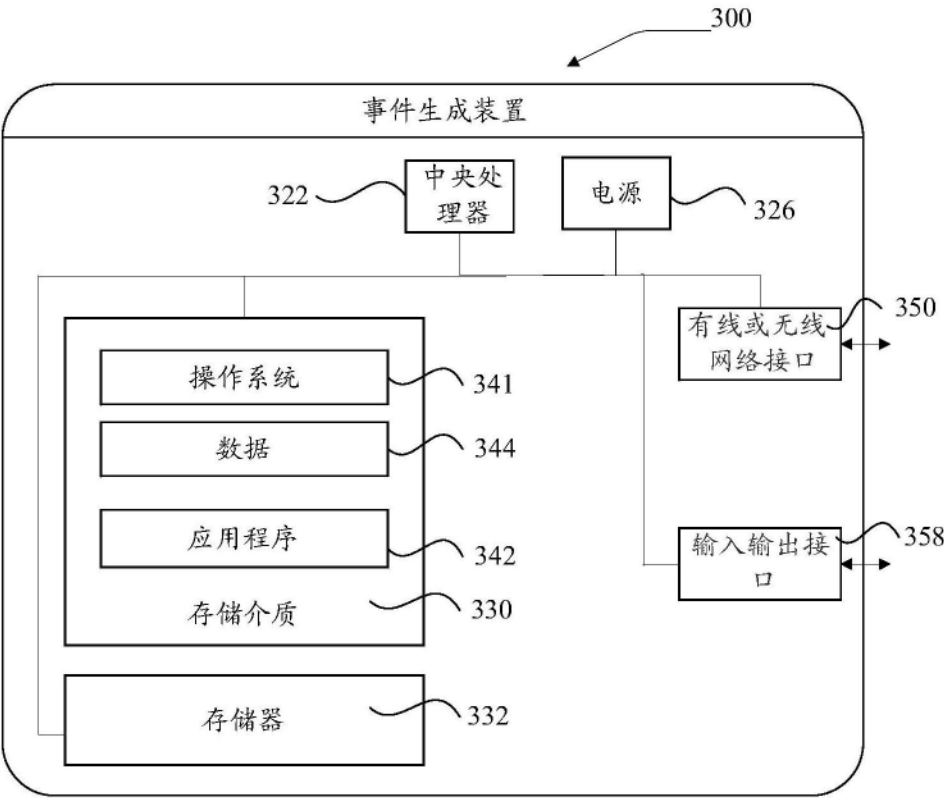


图8