



(12) 发明专利申请

(10) 申请公布号 CN 114428859 A

(43) 申请公布日 2022. 05. 03

(21) 申请号 202210095962.4

(22) 申请日 2022.01.26

(71) 申请人 京东科技信息技术有限公司

地址 100176 北京市北京经济技术开发区
科创十一街18号院2号楼6层601

(72) 发明人 孙宏超 陈蒙

(74) 专利代理机构 北京清亦华知识产权代理事
务所(普通合伙) 11201

代理人 杜月

(51) Int. Cl.

G06F 16/35 (2019.01)

G06F 40/216 (2020.01)

G06F 40/258 (2020.01)

G06F 40/30 (2020.01)

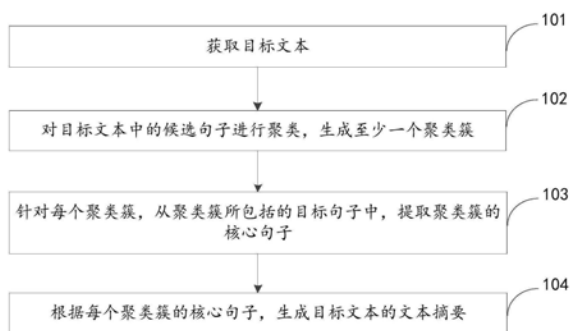
权利要求书3页 说明书11页 附图4页

(54) 发明名称

文本摘要的生成方法及其装置

(57) 摘要

本公开提出了一种文本摘要的生成方法及其装置,涉及自然语言处理领域,该文本摘要的生成方法包括:获取目标文本;对目标文本中的候选句子进行聚类,生成至少一个聚类簇;针对每个聚类簇,从聚类簇所包括的目标句子中,提取聚类簇的核心句子;根据每个聚类簇的核心句子,生成目标文本的文本摘要。由此,通过对目标文本中句子进行聚类处理的方式,可以获取目标文本的主题,并从主题中提取文本摘要,使得文本摘要更加全面、准确。



1. 一种文本摘要的生成方法,其特征在于,包括:
获取目标文本;
对所述目标文本中的候选句子进行聚类,生成至少一个聚类簇;
针对每个所述聚类簇,从所述聚类簇所包括的目标句子中,提取所述聚类簇的核心句子;
根据每个所述聚类簇的核心句子,生成所述目标文本的文本摘要。
2. 根据权利要求1所述的方法,其特征在于,所述对所述目标文本中的句子进行聚类,生成至少一个聚类簇,还包括:
对所述目标文本进行拆分,生成多个所述候选句子;
对所述候选句子进行主题聚类,生成所述至少一个聚类簇,其中,一个主题对应一个聚类簇。
3. 根据权利要求2所述的方法,其特征在于,所述对所述候选句子进行主题聚类,生成所述至少一个聚类簇,包括:
获取所述候选句子出现在所述聚类簇中的出现概率;
选取所述出现概率最大的聚类簇,确定为所述候选句子所隶属的目标聚类簇。
4. 根据权利要求3所述的方法,其特征在于,所述获取所述候选句子出现在所述聚类簇中的出现概率,包括:
获取所述聚类簇的完整性参数;
获取所述聚类簇与所述候选句子的同质性参数;
获取所述聚类簇与所包括的所述目标句子之间存在的共现词的语义权重;
根据所述完整性参数、所述同质性参数和所述语义权重,获取所述出现概率。
5. 根据权利要求4所述的方法,其特征在于,所述获取所述聚类簇的完整性参数,包括:
获取所述聚类簇中当前所包括目标句子的第一数量;
获取所述目标文本中所述候选句子的第二数量;
根据所述第一数量和所述第二数量,确定所述完整性参数。
6. 根据权利要求4所述的方法,其特征在于,所述获取所述聚类簇与所述候选句子的同质性参数,包括:
获取任一分词在所述候选句子中出现的第一词频;
获取所述任一分词在所述聚类簇中出现的第二词频和所述聚类簇的总词数;
获取所述任一分词在各个聚类簇的第一权重;
根据所述第一词频、所述第二词频、所述总词数、所述第一权重和词典大小,确定所述同质性参数。
7. 根据权利要求4所述的方法,其特征在于,所述获取所述聚类簇与所包括的所述目标候选句子之间的共现词的语义权重,包括:
获取所述共现词在每个所述聚类簇中所述目标句子中出现的第二词频,并根据每个所述共现词的第二词频,确定所述语义权重。
8. 根据权利要求7所述的方法,其特征在于,所述根据每个所述共现词的第二词频,确定所述语义权重,包括:
根据任意两个共现词的第二词频,获取所述任意两个共现词之间的子语义权重;

对所有的所述任意两个共现词之间的子语义权重求和,得到所述语义权重。

9. 根据权利要求8所述的方法,其特征在于,采用如下公式确定所述子语义权重:

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

其中,所述 w_{ij} 表示所述子语义权重;所述 x' 表示所述目标句子;所述 $n_{x'}^{w_i}$ 表示共现词 w_i 出现在所述目标句子中的第二词频;所述 $n_{x'}^{w_j}$ 表示共现词 w_j 出现在所述目标句子中的第二词频。

10. 根据权利要求2-9任一项所述的方法,其特征在于,所述对所述候选句子进行主题聚类,生成所述至少一个聚类簇,包括:

通过主题聚类模型对所述候选句子进行主题聚类分析,获取所述候选句子在所述聚类簇中出现的出现概率。

11. 根据权利要求1-9任一项所述的方法,其特征在于,所述从所述聚类簇所包括的目标句子中,提取所述聚类簇的核心句子,包括:

获取所述聚类簇中的目标句子的第二权重;

根据所述第二权重,从所述聚类簇中的目标句子中,选取所述聚类簇的核心句子。

12. 根据权利要求11所述的方法,其特征在于,所述获取所述聚类簇中的目标句子的第二权重,包括:

获取所述聚类簇中目标句子之间的有向图,其中,所述有向图中的节点表示所述目标句子,所述节点之间的边表示所述目标句子之间的贡献关系;

根据所述有向图和文本排名TextRank算法,获取所述聚类簇中每个所述目标句子的第二权重。

13. 一种文本摘要的生成装置,其特征在于,包括:

获取模块,用于获取目标文本;

聚类模块,用于对所述目标文本中的候选句子进行聚类,生成至少一个聚类簇;

提取模块,用于针对每个所述聚类簇,从所述聚类簇所包括的目标句子中,提取所述聚类簇的核心句子;

生成模块,用于根据每个所述聚类簇的核心句子,生成所述目标文本的文本摘要。

14. 根据权利要求13所述的装置,所述聚类模块,还用于:

对所述目标文本进行拆分,生成多个所述候选句子;

对所述候选句子进行主题聚类,生成所述至少一个聚类簇,其中,一个主题对应一个聚类簇。

15. 根据权利要求14所述的装置,所述聚类模块,还用于:

获取所述候选句子出现在所述聚类簇中的出现概率;

选取所述出现概率最大的聚类簇,确定为所述候选句子所隶属的目标聚类簇。

16. 根据权利要求15所述的装置,所述聚类模块,还用于:

获取所述聚类簇的完整性参数;

获取所述聚类簇与所述候选句子的同质性参数；

获取所述聚类簇与所包括的所述目标句子之间存在的共现词的语义权重；

根据所述完整性参数、所述同质性参数和所述语义权重，获取所述出现概率。

17. 根据权利要求16所述的装置，所述聚类模块，还用于：

获取所述聚类簇中当前所包括目标句子的第一数量；

获取所述目标文本中所述候选句子的第二数量；

根据所述第一数量和所述第二数量，确定所述完整性参数。

18. 根据权利要求16所述的装置，所述聚类模块，还用于：

获取任一分词在所述候选句子中出现的第二词频；

获取所述任一分词在所述聚类簇中出现的第二词频和所述聚类簇的总词数；

获取所述任一分词在各个聚类簇的第一权重；

根据所述第一词频、所述第二词频、所述总词数、所述第一权重和词典大小，确定所述同质性参数。

19. 根据权利要求16所述的装置，所述聚类模块，还用于：

获取所述共现词在每个所述聚类簇中所述目标句子中出现的第二词频，并根据每个所述共现词的第二词频，确定所述语义权重。

20. 根据权利要求19所述的装置，所述聚类模块，还用于：

根据任意两个共现词的第二词频，获取所述任意两个共现词之间的子语义权重；

对所有的所述任意两个共现词之间的子语义权重求和，得到所述语义权重。

21. 根据权利要求20所述的装置，其特征在于，还采用如下公式确定所述子语义权重：

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

其中，所述 w_{ij} 表示所述子语义权重；所述 x' 表示所述目标句子；所述 $n_{x'}^{w_i}$ 表示共现词 w_i 出现在所述目标句子中的第二词频；所述 $n_{x'}^{w_j}$ 表示共现词 w_j 出现在所述目标句子中的第二词频。

22. 一种电子设备，其特征在于，包括存储器、处理器；

其中，所述处理器通过读取所述存储器中存储的可执行程序代码来运行与所述可执行程序代码对应的程序，以用于实现如权利要求1-12中任一项所述的方法。

23. 一种计算机可读存储介质，其上存储有计算机程序，其特征在于，该程序被处理器执行时实现如权利要求1-12中任一项所述的方法。

24. 一种计算机程序产品，包括计算机程序，所述计算机程序在被处理器执行时实现根据权利要求1-12中任一项所述的方法。

文本摘要的生成方法及其装置

技术领域

[0001] 本公开涉及自然语言处理领域,尤其涉及一种文本摘要的生成方法及其装置。

背景技术

[0002] 文本摘要是一种从给定的文本中提取有价值的信息,并在保留源文本文档的相关内容的前提下,将其以简单、简短和浓缩的形式呈现给用户的方法。一个很好的文本摘要算法需要摘要结果相关性高,冗余少,压缩比合适和覆盖率高。目前技术中,文本摘要方法主要是通过抽取式文本摘要,生成式文本摘要。

发明内容

[0003] 本公开旨在至少在一定程度上解决相关技术中的技术问题之一。

[0004] 为此,本公开的一个目的在于提出一种文本摘要的生成方法。

[0005] 本公开的第二个目的在于提出一种文本摘要的生成装置。

[0006] 本公开的第三个目的在于提出一种电子设备。

[0007] 本公开的第四个目的在于提出一种非瞬时计算机可读存储介质。

[0008] 本公开的第五个目的在于提出一种计算机程序产品。

[0009] 为达上述目的,本公开第一方面实施方式提出了一种文本摘要的生成方法,包括:获取目标文本;对目标文本中的候选句子进行聚类,生成至少一个聚类簇;针对每个聚类簇,从聚类簇所包括的目标句子中,提取聚类簇的核心句子;根据每个聚类簇的核心句子,生成目标文本的文本摘要。

[0010] 本公开通过对文本进行聚类,提取聚类文本中的核心句子,生成目标文本的文本摘要。由此,可以提高文本摘要的关联性,且连贯性好,可读性强。

[0011] 根据本公开的一个实施方式,所述文本摘要的生成方法还包括:对目标文本进行拆分,生成多个候选句子;对候选句子进行主题聚类,生成至少一个聚类簇,其中,一个主题对应一个聚类簇。

[0012] 根据本公开的一个实施方式,所述文本摘要的生成方法还包括:获取候选句子出现在聚类簇中的出现概率;选取出现概率最大的聚类簇,确定为候选句子所隶属的目标聚类簇。

[0013] 根据本公开的一个实施方式,所述文本摘要的生成方法还包括:获取聚类簇的完整性参数;获取聚类簇与候选句子的同质性参数;获取聚类簇与所包括的目标句子之间存在的共现词的语义权重;根据完整性参数、同质性参数和语义权重,获取出现概率。

[0014] 根据本公开的一个实施方式,所述文本摘要的生成方法还包括:获取聚类簇中当前所包括目标句子的第一数量;获取目标文本中候选句子的第二数量;根据第一数量和第二数量,确定完整性参数。

[0015] 根据本公开的一个实施方式,所述文本摘要的生成方法还包括:获取任一分词在候选句子中出现的词频;获取任一分词在聚类簇中出现的第二词频和聚类簇的总词

数;获取任一分词在各个聚类簇的第一权重;根据第一词频、第二词频、总词数、第一权重和词典大小,确定同质性参数。

[0016] 根据本公开的一个实施方式,所述文本摘要的生成方法还包括:获取共现词在每个聚类簇中目标句子中出现的第二词频,并根据每个共现词的第二词频,确定语义权重。

[0017] 根据本公开的一个实施方式,所述文本摘要的生成方法还包括:根据任意两个共现词的第二词频,获取任意两个共现词之间的子语义权重;对所有的任意两个共现词之间的子语义权重求和,得到语义权重。

[0018] 根据本公开的一个实施方式,所述文本摘要的生成方法还包括:

$$[0019] \quad cw_{ij} = \frac{\sum_{x' \subseteq z} n_{x'}^{w_i}}{\sum_{x' \subseteq z} n_{x'}^{w_i} + \sum_{x' \subseteq z} n_{x'}^{w_j}}, \quad (w_i, w_j) \in x'$$

[0020] 其中, w_{ij} 表示子语义权重; x' 表示目标句子; $n_{x'}^{w_i}$ 表示共现词 w_i 出现在目标句子中的第二词频; $n_{x'}^{w_j}$ 表示共现词 w_j 出现在目标句子中的第二词频。

[0021] 根据本公开的一个实施方式,所述文本摘要的生成方法还包括:通过主题聚类模型对候选句子进行主题聚类分析,获取候选句子在聚类簇中出现的出现概率。

[0022] 根据本公开的一个实施方式,所述文本摘要的生成方法还包括:获取聚类簇中的目标句子的第二权重;根据第二权重,从聚类簇中的目标句子中,选取聚类簇的核心句子。

[0023] 根据本公开的一个实施方式,所述文本摘要的生成方法还包括:获取聚类簇中目标句子之间的有向图,其中,有向图中的节点表示目标句子,节点之间的边表示目标句子之间的贡献关系;根据有向图和文本排名TextRank算法,获取聚类簇中每个目标句子的第二权重。

[0024] 为达上述目的,本公开第二方面实施例提出了一种文本摘要的生成装置,包括:获取模块,用于获取目标文本;聚类模块,用于对目标文本中的候选句子进行聚类,生成至少一个聚类簇;提取模块,用于针对每个聚类簇,从聚类簇所包括的目标句子中,提取聚类簇的核心句子;生成模块,用于根据每个聚类簇的核心句子,生成目标文本的文本摘要。

[0025] 根据本公开的一个实施方式,所述聚类模块,还用于:对目标文本进行拆分,生成多个候选句子;对候选句子进行主题聚类,生成至少一个聚类簇,其中,一个主题对应一个聚类簇。

[0026] 根据本公开的一个实施方式,所述聚类模块,还用于:获取候选句子出现在聚类簇中的出现概率;选取出现概率最大的聚类簇,确定为候选句子所隶属的目标聚类簇。

[0027] 根据本公开的一个实施方式,所述聚类模块,还用于:获取聚类簇的完整性参数;获取聚类簇与候选句子的同质性参数;获取聚类簇与所包括的目标句子之间存在的共现词的语义权重;根据完整性参数、同质性参数和语义权重,获取出现概率。

[0028] 根据本公开的一个实施方式,所述聚类模块,还用于:获取聚类簇中当前所包括目标句子的第一数量;获取目标文本中候选句子的第二数量;根据第一数量和第二数量,确定完整性参数。

[0029] 根据本公开的一个实施方式,所述聚类模块,还用于:获取任一分词在候选句子中

出现的第一词频;获取任一分词在聚类簇中出现的第二词频和聚类簇的总词数;获取任一分词在各个聚类簇的第一权重;根据第一词频、第二词频、总词数、第一权重和词典大小,确定同质性参数。

[0030] 根据本公开的一个实施方式,所述聚类模块,还用于:获取共现词在每个聚类簇中目标句子中出现的第二词频,并根据每个共现词的第二词频,确定语义权重。

[0031] 根据本公开的一个实施方式,所述聚类模块,还用于:根据任意两个共现词的第二词频,获取任意两个共现词之间的子语义权重;对所有的任意两个共现词之间的子语义权重求和,得到语义权重。

[0032] 根据本公开的一个实施方式,所述文本摘要的生成装置,还包括:

$$[0033] \quad cw_{ij} = \frac{\sum_{x' \subseteq z} n_{x'}^{w_i}}{\sum_{x' \subseteq z} n_{x'}^{w_i} + \sum_{x' \subseteq z} n_{x'}^{w_j}}, (w_i, w_j) \in x'$$

[0034] 其中, w_{ij} 表示子语义权重; x' 表示目标句子; $n_{x'}^{w_i}$ 表示共现词 w_i 出现在目标句子中的第二词频; $n_{x'}^{w_j}$ 表示共现词 w_j 出现在目标句子中的第二词频。

[0035] 为达上述目的,本公开第三方面实施例提出了一种电子设备,包括:至少一个处理器;以及与所述至少一个处理器通信连接的存储器;其中,所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个处理器执行,以实现如本公开第一方面实施例所述的文本摘要的生成方法。

[0036] 为达上述目的,本公开第四方面实施例提出了一种存储有计算机指令的非瞬时计算机可读存储介质,其中,所述计算机指令用于实现如本公开第一方面实施例所述的文本摘要的生成方法。

[0037] 为达上述目的,本公开第五方面实施例提出了一种计算机程序产品,包括计算机程序,所述计算机程序在被处理器执行时用于实现如本公开第一方面实施例所述的文本摘要的生成方法。

附图说明

[0038] 图1是本公开一个实施方式的一种文本摘要的生成方法的示意图;

[0039] 图2是本公开另一个实施方式的一种文本摘要的生成方法的示意图;

[0040] 图3是本公开提供的狄利克雷过程混合模型的构造过程的示意图;

[0041] 图4是本公开提供的中国餐馆过程的构造过程的示意图;

[0042] 图5是本公开另一个实施方式的一种文本摘要的生成方法的示意图;

[0043] 图6是本公开另一个实施方式的一种文本摘要的生成方法的示意图;

[0044] 图7是本公开另一个实施方式的一种文本摘要的目标节点的结构示意图;

[0045] 图8是本公开一个实施方式的一种文本摘要的生成装置的示意图;

[0046] 图9是本公开一个实施方式的一种电子设备的示意图。

具体实施方式

[0047] 下面详细描述本公开的实施例,所述实施例的示例在附图中示出,其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的,旨在用于解释本公开,而不能理解为对本公开的限制。

[0048] 图1为本公开提出的一种文本摘要的生成方法的一种示例性实施方式的示意图,如图1所示,该文本摘要的生成方法包括以下步骤:

[0049] S101,获取目标文本。

[0050] 在本公开实施例中,获取目标文本的方法可为多种。

[0051] 可选地,通过外设输入装置将目标文本输入至电子设备中。应该说明的是,该外设输入装置可为外设键盘、外设扫描仪等。

[0052] 可选地,电子设备还可以通过连接文本库,从文本库中获取目标文本。应该说明的是,该目标文本可为提前准备好的,并预存在文本库中,该文本库可存储在电子设备的存储空间中,也可位于服务器中。

[0053] 可选地,还可以通过访问网页,从网页上下载目标文本文件。

[0054] S102,对目标文本中的候选句子进行聚类,生成至少一个聚类簇。

[0055] 可以理解的是,目标文本是由一个或句子组成,不同的句子可能有不同的语义,对应地,不同的语义可代表不同的主题。在对目标文本进行处理生成候选句子后,可对候选句子进行主题聚类,生成至少一个聚类簇。需要说明的是,该聚类簇中的句子有共同的主题,这种主题与其他聚类簇的主题互不相同。举例来说,目标文本包括N个候选句子,其中句子A、B、C主题相同,与句子D主题不同,则句子A、B、C可包括在同一个聚类簇中,句子D位于其他聚类簇中。

[0056] 生成聚类簇的方法可以为多种,可选地,可通过主题聚类分析算法方法对候选句子进行聚类,以生成至少一个聚类簇。

[0057] 可选地,还可以通过调用主题聚类模型,对候选句子进行处理,以生成至少一个聚类簇。需要说明的是,该主题聚类模型可为提前训练好的,并预存在电子设备的存储空间中,以在需要时调取使用。

[0058] S103,针对每个聚类簇,从聚类簇所包括的目标句子中,提取聚类簇的核心句子。

[0059] 目标句子为对应聚类簇中包含的所有句子,通过对聚类簇中的目标句子进行处理,通过对聚类簇进行提取,可以提取出聚类簇的核心句子。可以理解的是,该核心句子中包含聚类簇的主题信息。

[0060] 可选地,可将聚类簇中的目标句子与句子库中的句子进行匹配,以生成聚类簇与句子库中的句子的匹配度。如果匹配度越高,该句子库中的句子为聚类簇核心句子的可能性越大,对应的可以认为匹配度最高的句子为聚类簇的核心句子。需要说明的是,上述句子库为存储有所有主题句子的数据库。

[0061] 可选地,还可将聚类簇中的目标句子通过提取算法进行处理,以生成目标句子的语义权重,并根据目标句子的语义权重生成聚类簇的核心句子。

[0062] S104,根据每个聚类簇的核心句子,生成目标文本的文本摘要。

[0063] 在对目标文本完成主题聚类后,可得到至少一个聚类簇和至少一个聚类簇对应的核心句子,对核心句子进行处理,以生成目标文本的文本摘要。

[0064] 可选地,当生成的聚类簇为单个时,则可认为该聚类簇对应的核心句子即为该目标文本的文本摘要。

[0065] 可选地,当生成的聚类簇为多个时,则可选取包含句子数量最多的前k个主题的前m个核心句子作为最终的文本摘要。需要说明的是,该k和m为提前设定好的,可根据实际情况进行修改。

[0066] 可选地,可将多个聚类簇的核心句子按照语义进行拼接,以生成目标文本的文本摘要。

[0067] 可选地,可进一步地对多个聚类簇的核心句子进行权重提取,并根据核心句子的权重来生成目标文本的文本摘要。

[0068] 在本公开实施例中,首先获取目标文本,然后目标文本中的候选句子进行聚类,生成至少一个聚类簇,而后针对每个聚类簇,从聚类簇所包括的目标句子中,提取聚类簇的核心句子,最后根据每个聚类簇的核心句子,生成目标文本的文本摘要。由此,由此,通过对目标文本中句子进行聚类处理的方式,可以更好地挖掘目标文本的意图,提高获取文本摘要的效率和准确性,并且能够获取更全面的文本摘要。

[0069] 上述实施例中,对目标文本中的句子进行聚类,生成至少一个聚类簇,可结合图2进一步解释,图2为本公开提出的一种文本摘要的生成方法的另一种示例性实施方式的示意图,如图所示,该方法包括:

[0070] S201,对目标文本进行拆分,生成多个候选句子。

[0071] 在获取到目标文本后,可根据拆分条件对目标文本进行拆分,以获取多个候选句子。其中,该拆分条件可为一个完整的句子、一段完整的语义等,具体可根据实际需要进行设定。

[0072] 可选地,在对目标文本进行拆分前,可先对目标文本进行数据清洗,通过数据清洗,可将目标文本中错误的数据和无用的数据进行过滤,以便后续拆分更加容易,并且得到的候选句子更加准确。

[0073] 可选地,该数据清洗过程可包括:(1)去除目标文本中的特殊字符和停用词,存入词典;(2)将目标文本转换成由候选句子组成的数组;(3)提取每一个候选句子产生的主题。应该说明的是,该主题可为双词文法(bi-gram)或三词文法(tri-gram)。其中,词典可理解为目标文本单词的集合,该词典可存储在电子设备的存储空间中。

[0074] 需要说明的是,停用词是指在处理自然语言数据(或文本)之前或之后会自动过滤掉某些字或词,本实施例中的停用词可根据实际需要进行设定。

[0075] S202,对候选句子进行主题聚类,生成至少一个聚类簇,其中,一个主题对应一个聚类簇。

[0076] 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聚类簇,每个主题对应一个聚类簇。

[0077] 在本公开实施例中,该主题聚类模型可为隐狄利克雷分布(Latent Dirichlet Allocation),应该说明的是,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。

[0078] 在本公开实施例中,可以通过对聚类簇主题的提取,实现以聚类主题引入目标文

本的语义信息,可以更方便对目标文本进行摘要提取。

[0079] 需要说明的是,该狄利克雷过程混合模型可以理解作为一种聚类方法,但是不需要指定类别数量,它可以从数据中腿短聚类簇的数量。

[0080] 在狄利克雷过程混合模型中,包括以下定义内容:

[0081] $G \sim DP(\alpha_0, G_0)$,

[0082] $\theta_i | G \sim G$,

[0083] $x_i | \theta_i \sim F(\theta_i)$ 。

[0084] 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 $x_1, x_2, \dots, x_n$,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。

[0085] 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 θ_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。

[0086] 如图3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 。

[0087] $F(\theta_i)$ 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 。

[0088] 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。

[0089] 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 $A_1, \dots, A_r$,其中, r 为自然数,如果存在 $(G(A_1), \dots, G(A_r)) \sim \text{Dir}(\alpha_0 G_0(A_1), \dots, \alpha_0 G_0(A_r))$,则认为 G 服从由基分布 G_0 和参数 α_0 构成的狄利克雷过程,可以表示为 $G \sim DP(\alpha_0, G_0)$ 。

[0090] 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 $\theta_1, \dots, \theta_n \sim \text{Mult}$,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 $(G(A_1), \dots, G(A_r)) | \theta_1, \dots, \theta_n \sim \text{Dir}(\alpha_0 G_0(A_1) + n_1, \dots, \alpha_0 G_0(A_r) + n_r)$,其中,公式中的 $n_1, \dots, n_r$ 表示的是观测值属于 A_k 的数目。

[0091] 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 $(1, 2, 3 \dots, k)$ 。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。

[0092] 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。

[0093] 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 $k \leq n$ 。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。

[0094] 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:

[0095] 如图4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 $\varphi_1, \dots, \varphi_k$, 分配桌子给每一个顾客 $z_1, \dots, z_n$, 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。

[0096] 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。

[0097] 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。

[0098] 上述实施例中,对候选句子进行主题聚类,生成至少一个聚类簇,可结合图5进一步解释,图5为本公开提出的一种文本摘要的生成方法的另一种示例性实施方式的示意图,如图所示,该方法包括:

[0099] S501, 获取候选句子出现在聚类簇中的出现概率。

[0100] 在本公开实施例中,首先获取聚类簇的完整性参数,获取聚类簇与候选句子的同质性参数,获取聚类簇与所包括的目标句子之间存在的共现词的语义权重,然后根据完整性参数、同质性参数和语义权重,获取出现概率。

[0101] 其中,获取聚类簇的完整性参数,可通过以下步骤获取:获取聚类簇中当前所包括目标句子的第一数量,获取目标文本中候选句子的第二数量,根据第一数量和第二数量,确定完整性参数。具体地,可通过如下公式计算完整性参数:

$$[0102] \quad \left(\frac{m_z}{D-1+\alpha_0 D} \right)$$

[0103] 其中 α_0 为设定的参数, m_z 表示聚类簇中包含的候选句子的第一数量, D 表示所有聚类簇中候选句子的第二数量。

[0104] 进一步地,获取聚类簇与候选句子的同质性参数,可包括以下步骤:获取任一分词在候选句子中出现的词频,获取任一分词在聚类簇中出现的第二词频和聚类簇的总词数,获取任一分词在各个聚类簇的第一权重,根据第一词频、第二词频、总词数、第一权重和

词典大小,确定同质性参数。具体地,同质性参数的计算公式可为:

$$[0105] \quad \left(\frac{\prod_{w \in x} \prod_{j=1}^{N_x^w} (n_z^w \cdot ICF_w) + \beta + j - 1}{\prod_{i=1}^{N_x} n_z + V\beta + i - 1} \right)$$

[0106] 其中,i和j均为自然数, N_x^w 表示在候选句子x中单词w的第一词频, n_z^w 表示在聚类簇z中单词w的第二词频, N_x 表示候选句子x的总词数, 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。具体地,ICF($w \in x$)可通过如下公式进行计算:

$$[0107] \quad ICF(w \in x) = \log \left(\frac{|Z|}{|w \in Z|} \right)$$

[0108] 进一步地,获取聚类簇与所包括的目标候选句子之间的共现词的语义权重,可包括以下步骤:获取共现词在每个聚类簇中目标句子中出现的第二词频,并根据每个共现词的第二词频,确定语义权重。

[0109] 进一步地,根据每个共现词的第二词频,确定语义权重,可以包括以下步骤:根据任意两个共现词的第二词频,获取任意两个共现词之间的子语义权重,对所有的任意两个共现词之间的子语义权重求和,得到语义权重。

[0110] 其中, $n_x^{w_j}$ 表示共现词 w_j 出现在目标句子中的第二词频, x' 表示目标句子; $n_x^{w_i}$ 表示共现词 w_i 出现在目标句子中的第二词频。

[0111] 进一步地,该子语义权重 cw_{ij} 的公式可表示为:

$$[0112] \quad cw_{ij} = \frac{\sum_{x' \subseteq Z} n_{x'}^{w_i}}{\sum_{x' \subseteq Z} n_{x'}^{w_i} + \sum_{x' \subseteq Z} n_{x'}^{w_j}} (w_i, w_j) \in x'$$

[0113] 根据上述实施例中所描述的内容,聚类簇中的出现概率可用以下公式进行表示:

$$[0114] \quad p(z_x = z | \vec{z}, \vec{x}, \alpha_0, \beta) = \left(\frac{m_z}{D-1+\alpha_0 D} \right) * \left(\frac{\prod_{w \in x} \prod_{j=1}^{N_x^w} (n_z^w \cdot ICF_w) + \beta + j - 1}{\prod_{i=1}^{N_x} n_z + V\beta + i - 1} \right) * \left(1 + \sum_{w_i \in x \wedge w_j \in x} cw_{ij} \right) \quad (1)$$

[0115] 可以看出,聚类簇中包含的句子数量越多,等完整性参数的数值越高,该聚类簇的完整性越高。可以理解的是,该完整性参数越大,该候选句子对应的主题就越能代表目标文档的语义。

[0116] 在本公开实施例中同质性函数其数值可以用来衡量聚类簇和候选之间的同质性。等号右侧第二项的数值越高,聚类簇和候选句子之间的同质性越高。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。

[0117] 可以理解的是,聚类簇和候选句子之间的共现词的语义权重越高,说明该候选句子,该候选句子为目标句子的可能性越大。

[0118] S502,选取出现概率最大的聚类簇,确定为候选句子所隶属的目标聚类簇。

[0119] 在本公开实施例中,候选句子与不同的聚类簇之间的出现概率各不相同,出现概率越大,说明该候选句子隶属该聚类簇的可能性越大。

[0120] 举例来说,目标文本包括三个聚类簇A、B、C候选句子对聚类簇A的出现概率为0.9,对聚类簇B的出现概率为0.8,对聚类簇C的出现概率为0.7,则我们可以认为聚类簇A为候选句子所隶属的目标聚类簇。

[0121] 由此,通过对候选句子与聚类簇进行匹配,可以通过匹配概率来判断句子的语义,从而可以简单且高效的获取句子语义。

[0122] 应该说明的是,上述实施例中所涉及到的数据均可对目标文本进行处理以获得。

[0123] 上述实施例中,从聚类簇所包括的目标句子中,提取聚类簇的核心句子还可包括以下步骤,如图6所示:

[0124] S601,获取聚类簇中的目标句子的第二权重。

[0125] 在本公开实施例中,聚类簇可通过有向图进行表示,其中,有向图中的节点表示目标句子,节点之间的边表示句子之间的贡献关系。

[0126] 实现中,可通过TextRank算法对目标句子进行处理,以生成目标句子的第二权重。需要说明的是,TextRank算法是一种基于图的用于关键词抽取和文档摘要的排序算法,它利用一篇文档内部的词语间的共现信息(语义)便可以抽取关键词,它能够从一个给定的文本中抽取出该文本的关键词、关键词组,并使用抽取式的自动文摘方法抽取出该文本的关键句。

[0127] 该TextRank算法计算公式如下:

$$[0128] \quad WS(V_i) = (1 - d) + d * \sum_{V_j \in \text{In}(V_i)} \frac{W_{ji}}{\sum_{V_k \in \text{Out}(V_j)} W_{jk}} WS(V_j)$$

[0129] 其中, $WS(V_i)$ 表示目标句子i的第二权重, $WS(V_j)$ 表示上次迭代出的句子j的第二权重,d表示阻尼系数, $\text{In}(V_i)$ 表示流入i的节点集合, $\text{Out}(V_j)$ 表示流出j节点的集合, W_{ji} 表示两个句子的相似度。可以看出,目标节点的权重大小与目标节点流入i的节点和流出i的节点权重有关。如图7所示,目标节点i与流入i的节点和流出i的节点有共同边,具体地,除了孤立顶点外,每一个节点都至少与一条边相关联,可由边集完全描述。

[0130] 需要说明的是,本实施例中所描述的阻尼系数可根据实际模型进行设定,通常该系数大小为0.85。

[0131] S602,根据第二权重,从聚类簇中的目标句子中,选取聚类簇的核心句子。

[0132] 可以理解的是,目标句子的第二权重越大,该目标句子所占的语义就越接近聚类簇的主题。由此,可根据第二权重所占的比例来确定聚类簇的核心句子。

[0133] 可选地,可以根据第二权重值大小对聚类簇中的目标句子进行排序,选取前n个目标句子生成聚类簇的核心句子。其中,n为大于或等于的整数。

[0134] 可选地,还可以通过核心句子生成算法对第二权重进行处理,以生成聚类簇的核心句子。

[0135] 在本公开实施例中,首先获取聚类簇中的目标句子的第二权重,然后根据第二权重,从聚类簇中的目标句子中,选取聚类簇的核心句子。由此,通过对聚类簇中的句子语义

进行数据化,可以更加方便的对目标句子进行分析,从而获取准确的核心句子。

[0136] 图8为本公开提出的一种文本摘要的生成装置的示意图,如图8所示,该文本摘要的生成装置800,包括:获取模块801、聚类模块802、提取模块803和生成模块804,其中:

[0137] 获取模块801,用于获取目标文本;

[0138] 聚类模块802,用于对目标文本中的候选句子进行聚类,生成至少一个聚类簇;

[0139] 提取模块803,用于针对每个聚类簇,从聚类簇所包括的目标句子中,提取聚类簇的核心句子;

[0140] 生成模块804,用于根据每个聚类簇的核心句子,生成目标文本的文本摘要。

[0141] 进一步地,聚类模块802,还用于:对目标文本进行拆分,生成多个候选句子;对候选句子进行主题聚类,生成至少一个聚类簇,其中,一个主题对应一个聚类簇。

[0142] 进一步地,聚类模块802,还用于:获取候选句子出现在聚类簇中的出现概率;选取出现概率最大的聚类簇,确定为候选句子所隶属的目标聚类簇。

[0143] 进一步地,聚类模块802,还用于:获取聚类簇的完整性参数;获取聚类簇与候选句子的同质性参数;获取聚类簇与所包括的目标句子之间存在的共现词的语义权重;根据完整性参数、同质性参数和语义权重,获取出现概率。

[0144] 进一步地,聚类模块802,还用于:获取聚类簇中当前所包括目标句子的第一数量;获取目标文本中候选句子的第二数量;根据第一数量和第二数量,确定完整性参数。

[0145] 进一步地,聚类模块802,还用于:获取任一分词在候选句子中出现的第一词频;获取任一分词在聚类簇中出现的第二词频和聚类簇的总词数;获取任一分词在各个聚类簇的第一权重;根据第一词频、第二词频、总词数、第一权重和词典大小,确定同质性参数。

[0146] 进一步地,聚类模块802,还用于:获取共现词在每个聚类簇中目标句子中出现的第二词频,并根据每个共现词的第二词频,确定语义权重。

[0147] 进一步地,聚类模块802,还用于:根据任意两个共现词的第二词频,获取任意两个共现词之间的子语义权重;对所有的任意两个共现词之间的子语义权重求和,得到语义权重。

[0148] 进一步地,聚类模块802,采用如下公式确定子语义权重:

$$[0149] \quad cw_{ij} = \frac{\sum_{x' \subseteq z} n_{x'}^{w_i}}{\sum_{x' \subseteq z} n_{x'}^{w_i} + \sum_{x' \subseteq z} n_{x'}^{w_j}}, (w_i, w_j) \in x'$$

[0150] 其中, w_{ij} 表示子语义权重; x' 表示目标句子; $n_{x'}^{w_i}$ 表示共现词 w_i 出现在目标句子中的第二词频; $n_{x'}^{w_j}$ 表示共现词 w_j 出现在目标句子中的第二词频。

[0151] 为了实现上述实施例,本公开实施例还提出一种电子设备900,如图9所示,该电子设备900包括:处理器901和处理器通信连接的存储器902,存储器902存储有可被至少一个处理器执行的指令,指令被至少一个处理器901执行,以实现如本公开第一方面实施例的文本摘要的生成方法。

[0152] 为了实现上述实施例,本公开实施例还提出一种存储有计算机指令的非瞬时计算机可读存储介质,其中,计算机指令用于使计算机实现如本公开第一方面实施例的文本摘

要的生成方法。

[0153] 为了实现上述实施例,本公开实施例还提出一种计算机程序产品,包括计算机程序,计算机程序在被处理器执行时实现如本公开第一方面实施例的文本摘要的生成方法。

[0154] 在本公开的描述中,需要理解的是,术语“中心”、“纵向”、“横向”、“长度”、“宽度”、“厚度”、“上”、“下”、“前”、“后”、“左”、“右”、“竖直”、“水平”、“顶”、“底”“内”、“外”、“顺时针”、“逆时针”、“轴向”、“径向”、“周向”等指示的方位或位置关系为基于附图所示的方位或位置关系,仅是为了便于描述本公开和简化描述,而不是指示或暗示所指的装置或元件必须具有特定的方位、以特定的方位构造和操作,因此不能理解为对本公开的限制。

[0155] 此外,术语“第一”、“第二”仅用于描述目的,而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此,限定有“第一”、“第二”的特征可以明示或者隐含地包括一个或者更多个该特征。在本公开的描述中,“多个”的含义是两个或两个以上,除非另有明确具体的限定。

[0156] 在本说明书的描述中,参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本公开的至少一个实施例或示例中。在本说明书中,对上述术语的示意性表述不必针对的是相同的实施例或示例。而且,描述的具体特征、结构、材料或者特点可以在任一个或多个实施例或示例中以合适的方式结合。此外,在不相互矛盾的情况下,本领域的技术人员可以将本说明书中描述的不同实施例或示例以及不同实施例或示例的特征进行结合和组合。

[0157] 尽管上面已经示出和描述了本公开的实施例,可以理解的是,上述实施例是示例性的,不能理解为对本公开的限制,本领域的普通技术人员在本公开的范围内可以对上述实施例进行变化、修改、替换和变型。

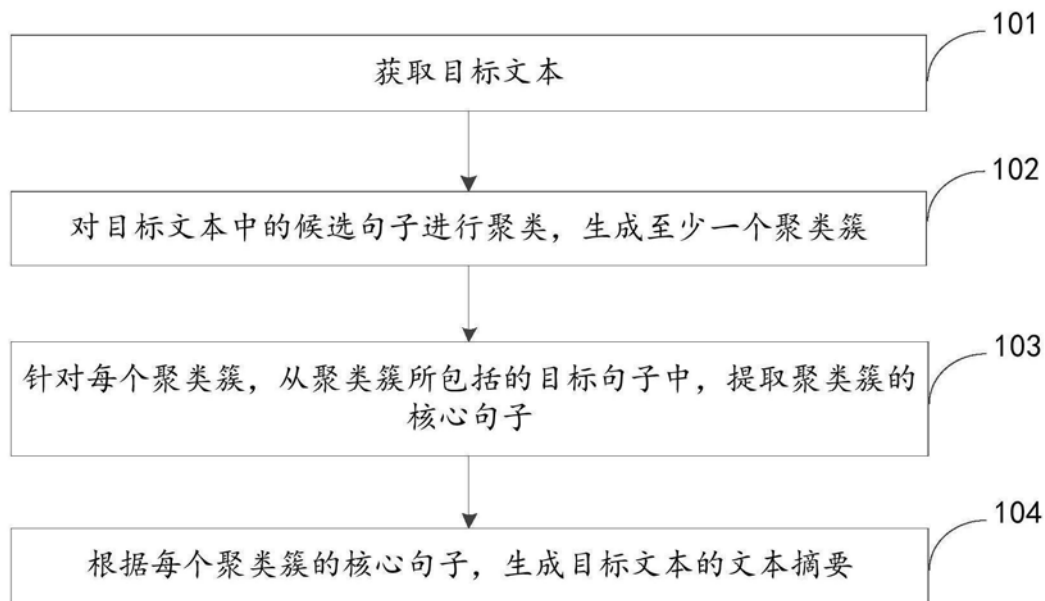


图1

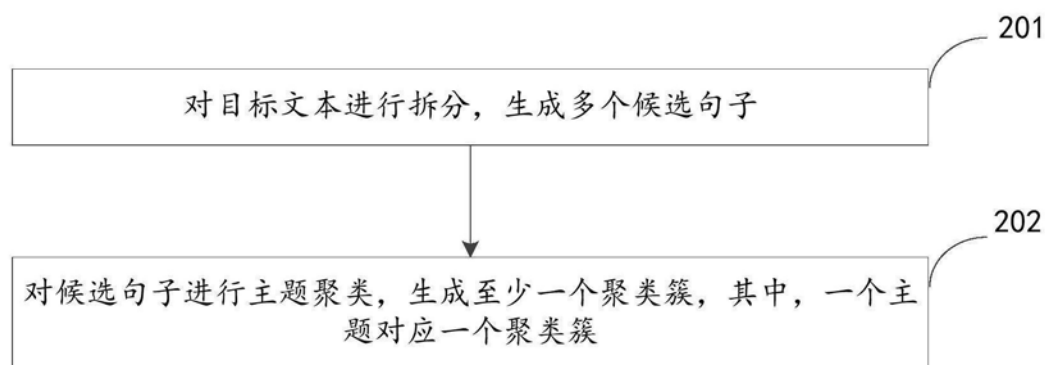


图2

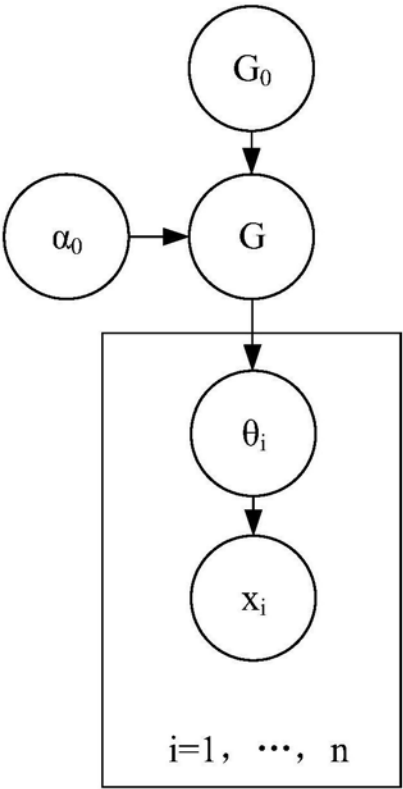


图3

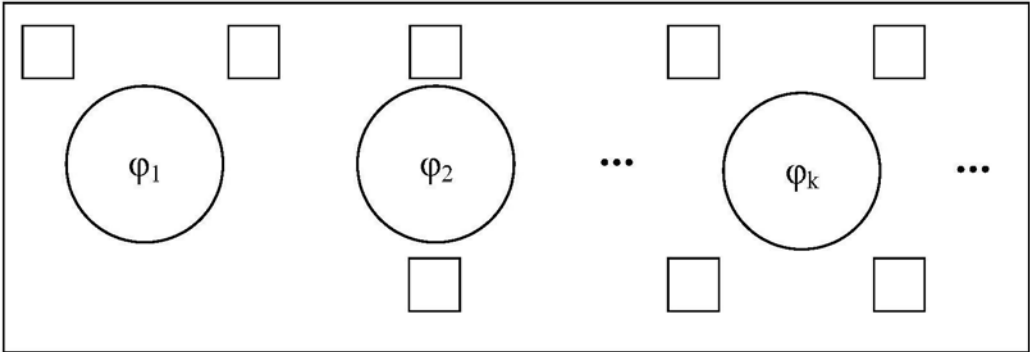


图4

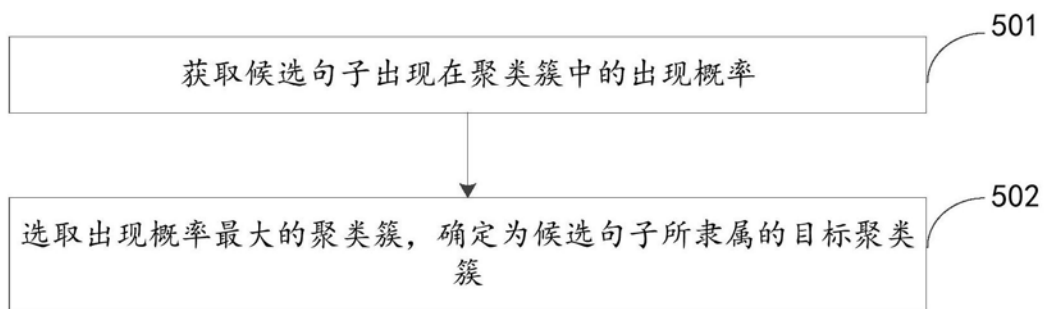


图5

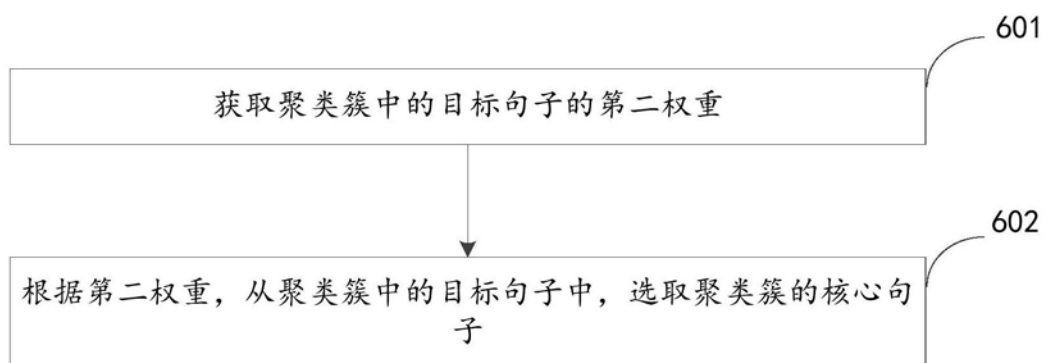


图6

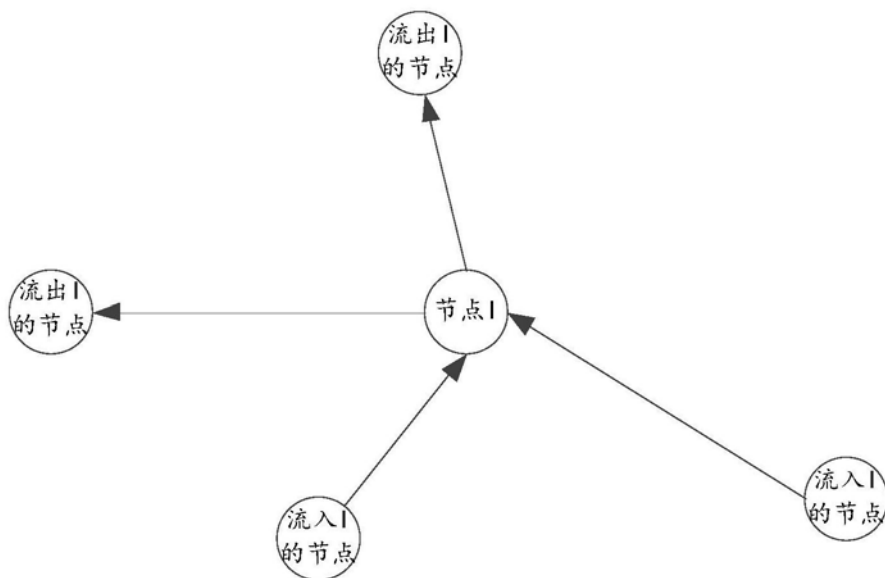


图7

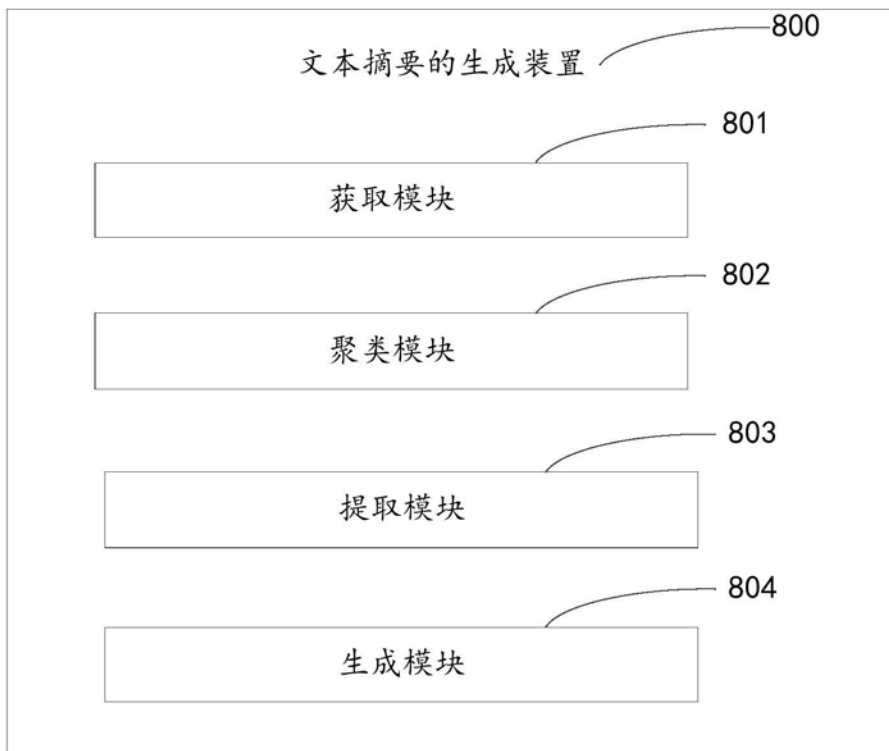


图8

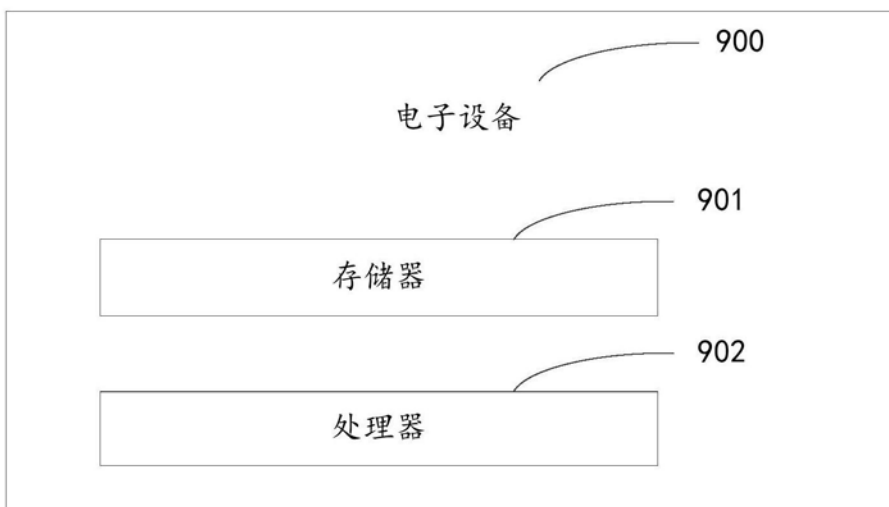


图9