# (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 108280772 A

( 43 )申请公布日 2018.07.13

( 21 )申请号 201810067320 .7

( 22)申请日 2018 .01 .24

(71 )申请人 北京航空航天大学

地址 100191 北京市海淀区学院路37号

(72)发明人 马帅 李莹莹 蒋浩谊 刘喆

(74 )专利代理机构 北京中创阳光知识产权代理

有限责任公司 11003

代理人 尹振启

( 51 )Int .Cl .

*G06Q 50/00*( 2012 .01 ) *G06F 17/30*( 2006 .01 )

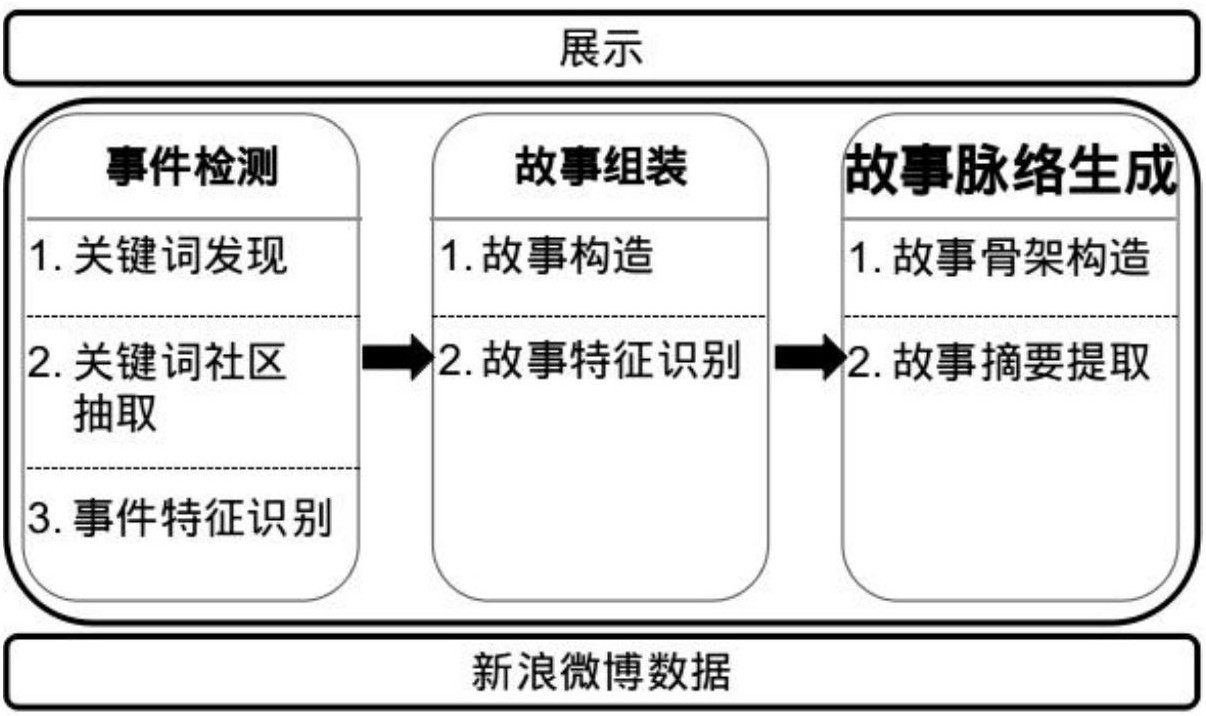


权利要求书1页 说明书4页 附图1页

( 54 )发明名称

社交网络中基于事件关联的故事脉络生成

方法

( 57 )摘要

本发明提出了一种在社交网络中基于事件关联的故事脉络生成方法。如图1所示为本发明的处理流程图，该流程主要包括三个阶段。首先进行事件检测，从微博流中检测强相关的事件；然后通过关联事件组装故事；最后生成并展示每个故事的故事脉络。

CN 108280772 A



# 权 利 要 求 书 1/1 页



CN 108280772 A

1 .一种在社交网络中基于事件关联的故事脉络生成方法，包括三个步骤，步骤1，从微博流中检测强相关的事件；步骤2，然后通过关联事件组装故事；步骤3，生成并展示每个故事的故事脉络。

1. 如权利要求1所述的方法，其特征在于，在步骤1中，事件检测由关键词发现，关键词社区抽取，事件特征识别三个步骤组成；在步骤2中，包括故事组装和故事特征识别步骤，根据主题对事件分组，并把事件组装成故事，首先，基于事件的隐式语义特征聚类，并把一个类作为一个故事；然后，根据故事中的事件集合标识其它的故事特征，所述标识其它的故事特征包括时间、地点、参与者和关键词；在步骤3中，包括故事骨架构造和故事摘要提取步骤，为每个故事生成故事脉络，用事件有向无环图和故事摘要可视化故事脉络，用弱联通分量和最大生成树构造一个有向无环图，然后提取故事的摘要。
2. 如权利要求2所述的方法，其特征在于，在步骤1中，在所述关键词发现步骤中，采用异常词检测方法检测包含不寻常模式的关键词并获得关键词图；在关键词社区抽取步骤中，采用重叠的社区检测算法过滤关键词社区并把一个社区作为一个事件的关键词集合；在事件特征识别步骤中，根据事件的关键词识别事件的其它特征，所述识别事件的其它特征包括时间、地点、参与者、描述和微博集合，所述时间表示该事件被检测的时间，微博集合中的微博包含事件的所有关键词，所述地点为命名实体中最频繁出现的地点，所述参与者为命名实体中所有人物和组织。
3. 如权利要求2所述的方法，其特征在于，在步骤2的故事构造步骤中，LDA捕捉事件的隐式语义特征，观察发现相同主题下的事件倾向于有更多的共有的词，首先进行预聚类，然后根据预聚类结果初始化LDA并进行精细聚类，在所述精细聚类中，根据预聚类结果初始化 LDA ,相同预聚类中的事件中的词被分配到相同的主题中，然后对LDA进行抽样，并根据事件的主题向量把事件分配到故事中；在所述故事特征识别步骤中，根据该故事的事件集合识别故事特征，首先将故事的开始时间和结束时间分别被赋值为事件集合中的最早时间和最晚时间，然后将地点、参与者和关键词集合作为相应的事件特征的并集。
4. 如权利要求2所述的方法，其特征在于，在步骤3的故事骨架构造中，将故事骨架表示为一个事件有向无环图，首先计算任意两事件间的有向边并生成一个事件图，然后构造事件图的弱联通分量集合；最后为每个弱联通分量构造最大生成树；然后进行故事摘要提取，从故事内事件的描述中提取一些句子用于表示故事摘要。首先为每个最大生成树生成摘要，然后将所述摘要合并起来表示故事摘要。

说 明 书 1/4 页



CN 108280772 A

# 社交网络中基于事件关联的故事脉络生成方法

技术领域

[0001] 本发明涉及一种社交网络，尤其涉及一种社交网络中基于事件关联的故事脉络生成方法。

背景技术

[0002] 社交网络已经被政府、公司和总统(例：奥巴马和特朗普)等广泛的用于发布新闻和报道事件。社交网络中新闻的实时性和快速传播性使得社交网络成为一个浏览新闻的重要媒介，通过短文本传递关键信息也有助于我们理解事件。然而，微博的快速积累，以及社交网络中的短文本包含大量含口语化的表达和错别字等使得监控事件的发展过程(即故事脉络)难以完全获取，另外，阅读所有相关的独立微博也将耗费大量的时间。因此，检测事件、高效的组装故事和用合适的方式可视化故事脉络成为亟待解决的问题。

[0003] 现有技术中，基于用户查询的故事脉络生成方法用与用户查询相关的文本生成故事脉络。一是提取与查询相关的新闻文本，根据基于图的词聚类方法发现事件，并根据事件间关联关系生成故事脉络。二是根据主题相关的对象构造一个带权重的图，并使用最小权重的连通支配集选择对象生成故事脉络。三是检测相关的tweets，构造带权重的图，并使用最小权重的连通支配集选择tweets生成故事脉络。

[0004] 自动的故事脉络生成方法不需要用户指定与故事相关的关键词，自动的根据文本生成故事脉络。大致可分为两类：逐步方法和统一方法。逐步方法把故事脉络生成形式化为多个独立的组件：事件检测、故事组装和脉络生成。统一方法用一个模型生成故事脉络。 [0005] 逐步方法把故事脉络生成形式化为多个独立的组件：事件检测、故事组装和脉络生成。现有技术进行事件检测的方法，一是把社交流建模成动态的微博网络并把网络中的一个类作为一个事件；二是对新闻文本流聚类并把一个类作一个事件。现有技术进行故事组装中，基于语义距离和给定的阈值把事件分配给已经存在的故事。现有技术进行脉络生成中，或者用事件间的Jaccard系数和给定的阈值连接事件并生成故事脉络；或者用自定义的函数，即连接强度，生成故事脉络。

[0006] 现有技术中统一方法用一个模型生成故事脉络。一是假设一个新闻文章报道故事

的一个进展，通过把文章连接到最相关的文章生成故事脉络。二是用无监督的隐变量模型在新闻文本流中建模故事在连续时间的发展过程。三是构造一个基于非负矩阵分解的主题模型并用该模型从微博流中追踪故事在连续时间的发展过程。

[0007] 然而，现有技术基于用户查询的故事脉络生成方法强依赖于用户指定的查询，这使得该类方法的使用受到一定的局限。而自动的故事脉络生成方法逐步方法中，若两个事件间的相似度大于给定的阈值，连接两个事件，然后设定一个合理的阈值很困难，并且这种相似度计算方式不客观，难以获得真实的结果。社交网络数据较新闻文本有独特的特征，基于新闻文本的故事脉络生成方法在社交网络数据中不能得到承诺的效果。而统一方法中单个微博不包含事件的所有关键元素(例地点和参与者) ，因此统一方法不能直接应用到社交网络中，以及不能连接有较长时间跨度的事件。

说 明 书 2/4 页



CN 108280772 A

发明内容

[0008] 为解决上述问题，本发明提出了一种在社交网络中基于事件关联的故事脉络生成方法。该方法主要包括三个步骤。首先，从微博流中检测强相关的事件；然后，通过关联事件组装故事；最后，生成并展示每个故事的故事脉络。本发明的社交网络中基于事件关联的故事脉络生成方法中事件脉络更清晰，可视化效果更好，可以更优的应用到社交网络中帮助用户理解事件的发展过程。

附图说明

[0009] 图1为本发明的处理流程图；

[0010] 图2为本发明一实施例的处理流程步骤。

具体实施方式

[0011] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。此外，下面所描述的本发明各个实施方式中所涉及到的技术特征只要彼此之间未构成冲突就可以相互组合。

[0012] 本发明提出了一种在社交网络中基于事件关联的故事脉络生成方法。如图1所示为本发明的处理流程图，该流程主要包括三个阶段。首先进行事件检测，从微博流中检测强相关的事件；然后通过关联事件组装故事；最后生成并展示每个故事的故事脉络。

|  |  |
| --- | --- |
| [0013] | 如图2所示为一实施例的处理流程步骤： |
| [0014] | 步骤1，每十分钟检测一次事件； |
| [0015] | 步骤2，用DBSCAN对数据集预聚类，得到预聚类结果P； |
| [0016] | 步骤3，根据预聚类结果初始化LDA ,并进行精细聚类.一个类对应一个故事。 |
| [0017] | 步骤4，为每个故事生成一个故事脉络。首先构造故事骨架；再提取故事摘要。 |
| [0018] | 需要设置的参数包括DBSCAN需要的参数半径radius和最小点数minpts。LDA需要 |
| 的超参数alpha、beta和故事(主题)数storynumber。 | |
| [0019] | 参数的默认值如下所示： |
| [0020] | DBSCAN：0 .65<radius≤0 .75，2≤minpts≤4 |
| [0021] | LDA:alpha＝0 .1，beta＝0 .03，50≤storynumber<450 |
| [0022] | 本发明首先使用一种适合短文本且能检测强相关事件的事件检测算法检测事件； |

然后通过用隐式语义信息关联事件来组装故事；最后为每个故事生成一个故事脉络。故事脉络即可以帮助用户捕捉一个故事的大意，也可以浏览故事中详细的发展过程。

[0023] 事件检测阶段中，根据事件中的微博集合必须强相关的假设从微博流中检测事件。首先从微博中获得一个关键词图用于保留微博的关键信息；然后使用关键词图中的k- clique过滤事件的关键词集合；最后根据事件的关键词识别其它的事件特征，即时间、地点、参与者、描述和微博集合。

[0024] 本发明用基于图的方法从微博流中检测事件，从Ring系统下载微博数据，采用IK Analyzer分词并去除停用词。事件检测由关键词发现，关键词社区抽取，事件特征识别三个

说 明 书 3/4 页



CN 108280772 A

步骤组成。

[0025] 在关键词发现中，本发明用异常词检测方法检测包含不寻常模式的关键词并获得关键词图。

[0026] 在关键词社区抽取中，单个关键词不足以包含事件的全部内容，共现且强相关的关键词集合会增加事件的内容覆盖面。因此本发明用重叠的社区检测算法过滤关键词社区并把一个社区作为一个事件的关键词集合。

[0027] 在事件特征识别中，根据事件的关键词识别事件的其它特征，即时间、地点、参与者、描述和微博集合。时间表示该事件被检测的时间，通常设置为10分钟。微博集合中的微博必须包含事件的所有关键词。描述是从该事件的微博集合中提取的一句包含关键词最多的句子。对该事件的微博分词并得到所有的命名实体，即地点，人物和组织。地点指命名实体中最频繁出现的地点。参与者指命名实体中所有人物和组织。

[0028] 故事组装阶段中，根据主题对事件分组，并把事件组装成故事。首先，基于事件的隐式语义特征聚类，并把一个类作为一个故事；然后，根据故事中的事件集合标识其它的故事特征，即时间、地点、参与者和关键词。

[0029] 本发明使用基于隐式语义特征的方法组装故事。首先进行故事构造，使用一个改良的LDA对事件分组，所述LDA(Latent Dirichlet Allocation)是一种文档主题生成模型，也称为主题模型。并把一个分组作为一个故事；然后进行故事特征识别，根据故事中的事件集合标识其它的故事特征，即时间、地点、参与者和关键词。

[0030] 故事构造中，LDA可以捕捉事件的隐式语义特征。观察发现相同主题下的事件倾向于有更多的共有的词。这个先验知识可以极大的减少LDA的搜索空间。本发明首先使用 DBSCAN进行预聚类，然后根据预聚类结果初始化LDA并进行精细聚类。

[0031] 在预聚类中，使用DBSCAN进行预聚类。首先为每个事件e构造词向量we。we ,k＝1，当第k个词在事件e中时；we ,k＝0，在其它任何条件下。然后把词向量做为输入，用DBSCAN把事件聚类到预聚类结果P中，其中P＝{p1 ,p2 ,...,pI}并且pi是一个包含一个事件集合的预聚类。DBSCAN中使用的聚类函数如下所示：

[0032] dis(ei ,ej)＝1-cosine(wei ,wej)

[0033] 在精细聚类中，根据DBSCAN的预聚类结果P初始化LDA ,相同预聚类中的事件中的词被分配到相同的主题中。然后对LDA进行Gibbs抽样，并根据事件的主题向量把事件分配到故事中。本发明认为一个事件只能属于一个故事，因此事件被分配到有最高概率的故事中。

[0034] 在故事特征识别中，为方便用户查询故事，本发明根据该故事的事件集合识别故事特征。首先，故事的开始时间和结束时间分别被赋值为事件集合中的最早时间和最晚时间；然后地点、参与者和关键词集合是相应的事件特征的并集。

[0035] 故事脉络生成阶段中，为每个故事生成故事脉络。用事件有向无环图和故事摘要可视化故事脉络。首先，用弱联通分量和最大生成树构造一个有向无环图；然后提取故事的摘要。故事脉络即可以帮助用户捕捉故事的大意，也可以浏览故事中详细的发展过程。 [0036] 故事骨架由一个事件有向无环图表示。首先，计算任意两事件间的有向边(公式2)并生成一个事件图；然后构造事件图的弱联通分量集合；最后为每个弱联通分量构造最大生成树。故事的骨架由多个最大生成树表示。

说 明 书 4/4 页



CN 108280772 A

[0037] w(ei ,ej)＝I(Tei ,Tej)·siml(ei ,ej)·(cp·simp(ei ,ej)+cc·simc(ei ,ej)) [0038] (2)其中，ei和ej表示两个事件；I(Tei ,Tej)表示两事件的时间关系；siml、simp和 simc表示两个事件间地点、参与者和关键词间的相似度。cp和cc是这些相似度的权重系数，且满足cp+cc＝1。

[0039] I(Tei ,Tej)＝1，当Tei<Tej时；I(Tei ,Tej)＝0，在其它情况下。siml (ei ,ej)＝1，当两个事件的地点Lei等于Lej时；siml(ei ,ej)＝0 .5，当地点Lei地理上属于Lej时；siml(ei ,ej)＝ 0，在其它情况下。simp(ei ,ej)表示两个事件的参与者Pei和Pej的Jaccard系数。simc(ei ,ej)表示两个事件的关键词Cei和Cej的Jaccard系数。

[0040] 故事摘要提取中，从故事内事件的描述中提取一些句子用于表示故事摘要。首先，为每个最大生成树生成摘要；然后这些摘要合并起来表示故事摘要。

[0041] 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。

# 说 明 书 附 图 1/1 页



CN 108280772 A

图1

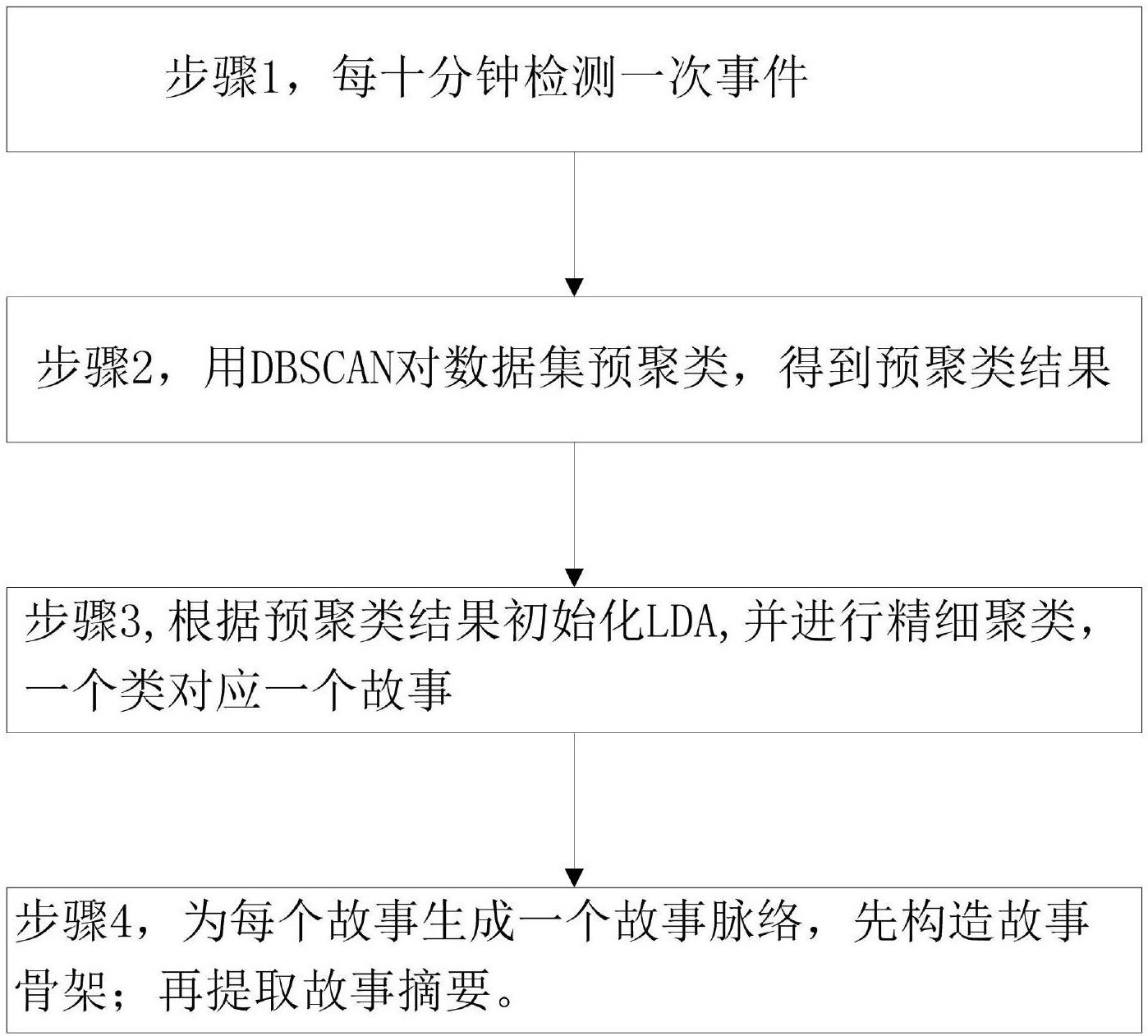


图2