



(12) 发明专利

(10) 授权公告号 CN 116521858 B

(45) 授权公告日 2024. 04. 30

(21) 申请号 202310445169.7

(22) 申请日 2023.04.20

(65) 同一申请的已公布的文献号
申请公布号 CN 116521858 A

(43) 申请公布日 2023.08.01

(73) 专利权人 浙江浙里信征信有限公司
地址 310000 浙江省杭州市西湖区文一西路83号浙江财经大学文华校区综合楼201室
专利权人 天道金科股份有限公司

(72) 发明人 马滨 任军霞 李响 唐嘉成
仇斌杰 赵建波

(74) 专利代理机构 杭州信与义专利代理有限公司 33450
专利代理师 马育妙

(51) Int. Cl.

G06F 16/34 (2019.01)

G06F 16/35 (2019.01)

(56) 对比文件

CN 103544255 A, 2014.01.29

CN 110543559 A, 2019.12.06

CN 110909153 A, 2020.03.24

CN 115470344 A, 2022.12.13

US 2018096057 A1, 2018.04.05

审查员 周亚楠

权利要求书2页 说明书10页 附图6页

(54) 发明名称

基于动态聚类和可视化的上下文语义序列比较方法

(57) 摘要

本发明公开了一种基于动态聚类和可视化的上下文语义序列比较方法,其中提供的ContextWing系统,支持对两个数据流之间不断演变的上下文序列模式进行两两比较。计算模型部分能够生成动态主题和序列模式,计算公众关注度和成对相关性的。系统中还包含一种新颖的多层双边翼隐喻设计,能够直观地展示不同上下文融合的序列模式,以揭示两个序列在时间和语义方面的异同。交互式工具则支持选择中心词及其上下文关键词,以迭代地生成模式以进行重点探索。另外,系统还支持静态和流式设置分析,支持更广泛的应用场景。



1. 一种基于动态聚类 and 可视化的上下文语义序列比较方法, 其特征在于, 对于实时流数据, 基于BERTopic和KMeans++的动态聚类方法对连续更新的推文进行动态聚类后, 再对动态流进行可视化分析, 可视化分析具体包括步骤:

S1, 根据用户选定的中心词, 通过计算推文中每个单词与所述中心词的相似度来提取所述中心词的上下文关键词; 并计算所述上下文关键词和所述中心词的公众关注度 $\mathcal{A}(c, k)$, $\mathcal{A}(c, k)$ 用于计算所述中心词与所述上下文关键词之间的距离, 以通过 $\mathcal{A}(c, k)$ 来量化针对后续选定的上下文关键词集合中的每个上下文关键词在后续构建的语义序列模式视图中的羽毛层的层次的水平位置;

S2, 计算所述上下文关键词与两个关键实体之间的关联度 C_i^t 并可视化, 与所述上下文关键词计算共现频率的实体定义为所述关键实体; 经过关联度 C_i^t 计算后, 将与所述中心词具有关联度且排名前n的所述上下文关键词形成为上下文关键词集合;

S3, 根据所述公众关注度 $\mathcal{A}(c, k)$ 、所述中心词及其上下文关键词集合, 通过迭代搜索方法, 生成语义序列模式并可视化;

步骤S1中, 计算所述中心词的所述上下文关键词的公众关注度的方法包括步骤:

S11, 计算所述公众关注度 $\mathcal{A}(c, k)$, 计算方法通过如下公式 (1) 表达:

$$\mathcal{A}(c, k) = \text{Log} \left[\frac{\sum_{i=1}^n u_i(c, k) \eta_i \cdot \sum_{i=1}^n u_i(c, k) r_i / \sum_{i=1}^n u_i(c, k)}{\sum_{i=1}^n u_i(c, -k) \eta_i \cdot \sum_{i=1}^n u_i(c, -k) r_i / \sum_{i=1}^n u_i(c, -k)} \right] \quad \text{公式(1)}$$

公式 (1) 中, k 表示用户或系统选定的所述中心词;

c 表示所述上下文关键词;

n 表示数据集中的推文总数;

$u_i(c, k)$ 是一个包含条件, 表示第 i 条推文是否包含 c 和 k, 如果是, 则 $u_i(c, k) = 1$, 否则为 0;

$u_i(c, -k)$ 表示第 i 条推文是否包含 c 但不包含 k, 如果是, 则 $u_i(c, -k) = 1$, 否则为 0;

η_i 表示第 i 条推文是否被转发, 如果是, 则 $\eta_i = 1$, 否则为 0;

r_i 表示第 i 条推文被转发的数量;

S12, 根据 $\mathcal{A}(c, k)$ 的值进行可视化。

2. 根据权利要求 1 所述的基于动态聚类 and 可视化的上下文语义序列比较方法, 其特征在于, 基于BERTopic和KMeans++的动态聚类方法对连续更新的推文进行动态聚类的方法包括步骤:

A1, BERTopic模型根据用户给定的所述中心词, 对连续更新的推文中的所述上下文关键词进行文本识别, 得到初始化 t 时刻待聚类的所述上下文关键词;

A2, 使用KMeans++算法初始化 t 时刻的聚类 C_t^k , 首次聚类完成后, 将聚类中心传递给 t+1 时刻的聚类 C_{t+1}^k ;

A3, 在每个聚类时刻, 判断 C_{t+1}^k 中前 m 个所述上下文关键词是否同样存在于 C_t^k 中, 若是, 则将 C_{t+1}^k 与 C_t^k 进行簇的合并, 并对合并后的簇中的所述上下文关键词按照基于类的 TF-

IDF得分进行排序,将排名前x的所述上下文关键词形成的集合作为数据更新后的 C_{t+1}^k ;

A4,采用步骤A2-A3的方法,完成对所有时刻识别到的所述上下文关键词的聚类,并将最终合并的簇中的前y个所述上下文关键词所在的推文作为待进行可视分析的对象。

3.根据权利要求1所述的基于动态聚类 and 可视化的上下文语义序列比较方法,其特征在于,步骤S1中,通过余弦相似度计算方法,对所述中心词与推文中的每个单词进行相似度计算,并将排名前n的单词作为所述上下文关键词集合。

4.根据权利要求1所述的基于动态聚类 and 可视化的上下文语义序列比较方法,其特征在于,步骤S2中, C_i^t 的计算方法通过如下公式(2)表达:

$$C_i^t = \frac{\text{Rank}(\beta_{iA}^t - \beta_{iB}^t)}{N^t}, \forall i \in W^t \text{ 公式(2)}$$

公式(2)中, β_{iA}^t 、 β_{iB}^t 分别表示所述上下文关键词i和关键实体A、关键实体B在时刻t的共现频率;

Rank表示计算了上下文关键词i的共现频率之差 $\beta_{iA}^t - \beta_{iB}^t$ 在所有 $i \in W^t$ 中的排名;

N^t 表示时刻t下中心词i的上下文关键词总数;

W^t 表示时刻t下中心词的所有上下文关键词集合。

5.根据权利要求1所述的基于动态聚类 and 可视化的上下文语义序列比较方法,其特征在于,步骤S3中,生成所述语义序列模式的方法包括步骤:

S31,形成初始序列,所述初始序列包含保留推文中出现顺序的由用户选定的所述中心词和所述上下文关键词;

S32,遍历所述关键词集合中的每个所述上下文关键词,查找在所述初始序列中新加入集合中的一个单词后使得形成的语义新序列中的词在推文中的共现频率最大的单词,然后将寻找到的所述上下文关键词加入到所述初始序列中实现序列扩充,并在所述关键词集合中过滤掉新加入到所述初始序列中的所述上下文关键词;

S33,以步骤S32扩充得到的所述语义新序列为所述初始序列并返回步骤S31,从过滤剩余的所述关键词集合中继续扩充所述初始序列,直至扩充后的序列达到预设的序列长度,将最终得到的所述语义新序列作为生成的所述语义序列模式。

基于动态聚类 and 可视化的上下文语义序列比较方法

技术领域

[0001] 本发明涉及数据分析技术领域,具体涉及一种基于动态聚类 and 可视化的上下文语义序列比较方法。

背景技术

[0002] 随着社交媒体的快速发展,许多人喜欢通过发布消息来表达自己的观点和概念,传播重大新闻,这些新闻以数据流的方式出现,包含相同关键词的推文集合形成一个社交媒体数据流。为了方便社会科学研究人员和舆论分析人员快速理解大量社交媒体数据,提供嵌入社交媒体信息的意见摘要尤为重要。这些推文的可视化摘要可以让用户快速理解这些文本数据。

[0003] 词云是为文本数据提供可视化摘要的常用方法。然而,词云提供的上下文信息有限,不能提供关键词之间的联系来传达句子的意思。因此,我们提取在句子中按顺序出现的关键词序列作为推文的摘要。同时,由于许多推文包含相同的序列,我们将这种序列定义为“模式”。例如,“选举辩论定于周四晚上9点开始”、“选举辩论将于周四开始”等。人们有不同的表达方式,但他们都提到了相同的关键词和顺序:“选举-辩论-周四开始”,这样频繁出现的语义序列即为一个模式。模式是非常多样的,需要比较它们之间的异同来了解民意。此外,由于这些模式属于不同的时间段,还需要从时间层面对模式进行比较。此外,为了帮助分析公众态度,需要比较模式和不同数据流之间的关系。为了处理这些复杂的分析,可以使用可视化技术来支持比较。

[0004] 文本的视觉比较是一个广泛的研究课题。但是,目前缺乏支持同时比较序列的时变特征和语义特征,以及在不同数据流中的分析方法。首先,在序列分析中很难将语义比较和动态比较结合起来。一些学者使用树形结构解决了序列比较的挑战,帮助人们快速理解基本概念和想法,然而,这种方法仅限于静态文本序列数据,不支持时间比较。支持多个标签云之间的时间趋势比较的工作又无法支持序列比较,因为关键词之间缺乏连接。因此,很难将序列的时间和语义比较同时可视化。其次,比较不同数据流中的语义和动态具有挑战性。一些工作解决了两个数据流之间多项目的成对可视化比较的挑战,但仍然不能应用于序列来显示更多的上下文和连接。第三,除了历史的社交媒体数据,实时分析对现实世界的流动数据来说更具挑战性,但也更重要,难点在于它需要快速的建模方法和动态可视化来揭示短时间内的特征。总的来说,缺乏一种可视化技术来支持同时在两个数据流中对时间和语义序列模式进行两两比较,也缺乏支持实时模式的分析

发明内容

[0005] 本发明以实现对文本序列的时间和语义比较同时可视化,并实现不同数据流间的语义和动态比较为目的,提供了一种基于动态聚类 and 可视化的上下文语义序列比较方法。

[0006] 为达此目的,本发明采取以下技术方案:

[0007] 提供一种基于动态聚类 and 可视化的上下文语义序列比较方法,对于实时流数据,

基于BERTopic和KMeans++的动态聚类方法对连续更新的推文进行动态聚类后,再对动态流进行可视化分析,可视化分析具体包括步骤:

[0008] S1,根据用户选定的中心词,通过计算推文中每个单词与所述中心词的相似度来提取所述中心词的上下文关键词;并计算所述上下文关键词和所述中心词的公众关注度 $\mathcal{A}(c, k)$;

[0009] S2,计算所述上下文关键词与这两个关键实体之间的关联度 C_t^t 并可可视化;

[0010] S3,根据所述中心词及其上下文关键词集合,通过迭代搜索方法,生成语义序列模式并可可视化。

[0011] 作为优选,基于BERTopic和KMeans++的动态聚类方法对连续更新的推文进行动态聚类的方法包括步骤:

[0012] A1, BERTopic模型根据用户给定的所述中心词,对连续更新的推文中的所述上下文关键词进行文本识别,得到初始化t时刻待聚类的所述上下文关键词;

[0013] A2,使用KMeans++算法初始化t时刻的聚类 C_t^k ,首次聚类完成后,将聚类中心传递给t+1时刻的聚类 C_{t+1}^k ;

[0014] A3,在每个聚类时刻,判断 C_{t+1}^k 中前m个所述上下文关键词是否同样存在于 C_t^k 中,若是,则将 C_{t+1}^k 与 C_t^k 进行簇的合并,并对合并后的簇中的所述上下文关键词按照基于类的TF-IDF得分进行排序,将排名前x的所述上下文关键词形成的集合作为数据更新后的 C_{t+1}^k ;

[0015] A4,采用步骤A2-A3的方法,完成对所有时刻识别到的所述上下文关键词的聚类,并将最终合并的簇中的前y个所述上下文关键词所在的推文作为待进行可视分析的对象。

[0016] 作为优选,步骤S1中,通过余弦相似度计算方法,对所述中心词与推文中的每个单词进行相似度计算,并将排名前n的单词作为所述上下文关键词集合。

[0017] 作为优选,步骤S1中,计算所述中心词的所述上下文关键词的公众关注度的方法包括步骤:

[0018] S11,计算所述公众关注度 $\mathcal{A}(c, k)$,计算方法通过如下公式(1)表达:

$$[0019] \quad \mathcal{A}(c, k) = \text{Log} \left[\frac{\sum_{i=1}^n u_i(c, k) \eta_i \cdot \sum_{i=1}^n u_i(c, k) r_i / \sum_{i=1}^n u_i(c, k)}{\sum_{i=1}^n u_i(c, -k) \eta_i \cdot \sum_{i=1}^n u_i(c, -k) r_i / \sum_{i=1}^n u_i(c, -k)} \right] \quad \text{公式(1)}$$

[0020] 公式(1)中,k表示用户或系统选定的所述中心词;

[0021] c表示所述上下文关键词;

[0022] n表示数据集中的推文总数;

[0023] $u_i(c, k)$ 是一个包含条件,表示第i条推文是否包含c和k,如果是,则 $u_i(c, k) = 1$,否则为0;

[0024] $u_i(c, -k)$ 表示第i条推文是否包含c但不包含k,如果是,则 $u_i(c, -k) = 1$,否则为0;

[0025] η_i 表示第i条推文是否被转发,如果是,则 $\eta_i = 1$,否则为0;

[0026] r_i 表示第i条推文被转发的数量;

[0027] S12,根据 $\mathcal{A}(c, k)$ 的值进行可视化。

[0028] 作为优选,步骤S2中, C_i^t 的计算方法通过如下公式(2)表达:

$$[0029] \quad C_i^t = \frac{\text{Rank}(\beta_{iA}^t - \beta_{iB}^t)}{N^t}, \forall i \in W^t \text{ 公式(2)}$$

[0030] 公式(2)中, β_{iA}^t 、 β_{iB}^t 分别表示所述上下文关键词i和关键实体A、关键实体B在时刻t的共现频率;

[0031] Rank表示计算了上下文关键词i的共现频率之差 $\beta_{iA}^t - \beta_{iB}^t$ 在所有 $i \in W^t$ 中的排名;

[0032] N^t 表示时刻t下中心词i的上下文关键词总数;

[0033] W^t 表示时刻t下中心词的所有上下文关键词集合。

[0034] 作为优选,步骤S3中,生成所述语义序列模式的方法包括步骤:

[0035] S31,形成初始序列,所述初始序列包含保留推文中出现顺序的由用户选定的所述中心词和所述上下文关键词;

[0036] S32,遍历所述关键词集合中的每个所述上下文关键词,查找在所述初始序列中新加入集合中的一个单词后使得形成的语义新序列中的词在推文中的共现频率最大的单词,然后将寻找到的所述上下文关键词加入到所述初始序列中实现序列扩充,并在所述关键词集合中过滤掉新加入到所述初始序列中的所述上下文关键词;

[0037] S33,以步骤S32扩充得到的所述语义新序列为所述初始序列并返回步骤S31,从过滤剩余的所述关键词集合中继续扩充所述初始序列,直至扩充后的序列达到预设的序列长度,将最终得到的所述语义新序列作为生成的所述语义序列模式。

[0038] 本发明提供的ContextWing系统,支持对两个数据流之间不断演变的上下文序列模式进行两两比较。计算模型部分能够生成动态主题和序列模式,计算公众关注度和成对相关。系统中还包含一种新颖的多层双边翼隐喻设计,能够直观地展示不同上下文融合的序列模式,以揭示两个序列在时间和语义方面的异同。交互式工具则支持选择中心词及其上下文关键词,以迭代地生成模式以进行重点探索。另外,系统还支持静态和流式设置分析,支持更广泛的应用场景。

附图说明

[0039] 为了更清楚地说明本发明实施例的技术方案,下面将对本发明实施例中所需要使用的附图作简单地介绍。显而易见地,下面所描述的附图仅仅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。

[0040] 图1是本发明实施例提供的社交媒体上下文文本可视化的系统界面图;

[0041] 图2是图1中A区域显示的主题视图的界面放大图;

[0042] 图3是图1中B区域显示的控制视图的界面放大图;

[0043] 图4是图1中C区域显示的模式视图的局部放大图;

[0044] 图5是图2中a1区域显示的推文的数量直方图;

- [0045] 图6是图2中a2区域显示的动态的词云的界面示意图；
- [0046] 图7是图1中D区域显示的原始推文的细节视图的界面放大图；
- [0047] 图8是本发明实施例提供的可视化隐喻设计原理示意图；
- [0048] 图9是可视化隐喻的语义合并方法示意图；
- [0049] 图10是可视分析界面的系统架构流程图；
- [0050] 图11是主题视图的示例图。

具体实施方式

[0051] 下面结合附图并通过具体实施方式来进一步说明本发明的技术方案。

[0052] 其中,附图仅用于示例性说明,表示的仅是示意图,而非实物图,不能理解为对本专利的限制;为了更好地说明本发明的实施例,附图某些部件会有省略、放大或缩小,并不代表实际产品的尺寸;对本领域技术人员来说,附图中某些公知结构及其说明可能省略是可以理解的。

[0053] 本发明实施例的附图中相同或相似的标号对应相同或相似的部件;在本发明的描述中,需要理解的是,若出现术语“上”、“下”、“左”、“右”、“内”、“外”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发明和简化描述,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,因此附图中描述位置关系的用语仅用于示例性说明,不能理解为对本专利的限制,对于本领域的普通技术人员而言,可以根据具体情况理解上述术语的具体含义。

[0054] 在本发明的描述中,除非另有明确的规定和限定,若出现术语“连接”等指示部件之间的连接关系,该术语应做广义理解,例如,可以是固定连接,也可以是可拆卸连接,或成一体;可以是机械连接,也可以是电连接;可以是直接相连,也可以通过中间媒介间接相连,可以是两个部件内部的连通或两个部件的相互作用关系。对于本领域的普通技术人员而言,可以根据具体情况理解上述术语在本发明中的具体含义。

[0055] 本发明实施例提供一种基于动态聚类 and 可视化的上下文语义序列比较方法,其分析过程如下:

[0056] 本发明提出了一种如图1所示的名为ContextWi ng的社交媒体上下文文本可视化的系统,它集合了一个集成的计算模型、一个新颖的视觉设计以及一个对称的翅膀结构,连接了具有相同中心词(例如图1中的“chi na”的序列),并按照相同上下文关键词(如图1中的“ukrai ne”、“trademark”等)对序列进行合并,合并后用颜色和成对的层级加以区分,以更清楚地显示其语义的差异和相似之处。序列从上到下垂直排列,对应于不同的时间段。模式中的关键词按在原文中出现的顺序从左到右连接。同时,不同层级的位置和颜色编码了语义信息和关键实体(比如社交媒体事件中的“人物A”和“人物B”)的相关性,比较语义和成对实体的关系可以了解人们对关键实体的倾向性。因此,该视觉设计使用户可以同时上下文的时间特征和语义特征进行成对的视觉比较,克服了词云和词树的局限性。

[0057] 图1提供的系统界面包括A、B、C、D四个区域分别对应显示的如图2-4、图7所示的主题视图、控制视图、模式视图和原始推文的细节视图4个视图部分,其中主题视图部分又包括了如图5中所述的原始推文的数量直方图以及如图6中所示的显示动态词云的界面图。用户可以在图2所示的A区域显示的主题视图界面中选择不同的中心词和上下文关键词来生

成语义序列模式,可以通过图3中所示的B区域显示的控制视图观察用户所选的中心词和上下文关键词,并可通过嵌入在控制视图中的重置或返回功能,对用户所选的中心词和/或上下文关键词进行重置或返回到图2中所示的主题视图重新选择中心词和上下文关键词。图4所示的C区域显示的模式视图(即可视化的翅膀隐喻设计)用于可视化生成的语义序列模式。

[0058] 如何生成语义序列模式以及如何对不同数据流进行实时的模式分析,以及如何对分析结果进行可视化是本发明的关键技术内容,以下分三大块内容对实现上述关键技术的原理进行具体说明:

[0059] 一、搭建计算模型及数据流模式分析

[0060] 本发明所搭建的计算模型主要承担如下计算功能:关键词分类计算、成对相关性计算、公众关注度计算、上下文语义序列模式生成以及根据生成的模式,对不同数据流进行数据分析。

[0061] 1、关键词分类计算

[0062] 在静态设置下(表示数据是历史数据,不是实时更新的),使用Word2Vec(Word2Vec是将单词转换成向量形式的神经网络模型。通过转换,可以把对文本内容的处理简化为向量空间中的向量运算,计算出向量空间上的相似度,来表示文本语义上的相似度),得到原始推文中每个单词的向量,并计算向量的余弦相似度寻找到与用户给定的中心词相似的关键词,相似度越高表示两个词向量之间具有越高的语义相关性。由于给定的原始推文为历史数据,本发明可以通过获得先验知识来指定中心词,是中心词的聚类效果更符合专家预期。由于每个集群通常有大量的单词,本发明保留频率较高的前n个单词作为可视化的关键词。同样地,基于向量的余弦相似度,可以对每个中心词进行其上下文的关键词的提取,获得相似度较高的前n个词。考虑到可视化时需要确保主题视图清晰,一般选排名前20-30个上下文关键词。

[0063] 量化中心词与其上下文关键词之间关系最直接的方法是计算其在文本中的共现频率。然而,我们再实际应用中发现,简单地基于共现频率呈现的信息有一定的局限性,为了提升后续序列成对比较的效果,本发明还创新性地提出用公众关注度来表征中心词和上下文关键词之间的密切关系。我们将提出的公众关注度表征为 $\mathcal{A}(c, k)$,其用于计算中心词与上下文关键词之间的距离,这个距离能够较为准确的反映推文被转发的受欢迎程度。 $\mathcal{A}(c, k)$ 的计算方法通过如下公式(2)表达:

$$[0064] \quad \mathcal{A}(c, k) = \text{Log} \left[\frac{\sum_{i=1}^n u_i(c, k) \eta_i \cdot \sum_{i=1}^n u_i(c, k) r_i / \sum_{i=1}^n u_i(c, k)}{\sum_{i=1}^n u_i(c, -k) \eta_i \cdot \sum_{i=1}^n u_i(c, -k) r_i / \sum_{i=1}^n u_i(c, -k)} \right] \quad \text{公式(2)}$$

[0065] 公式(2)中,k表示用户或系统选定的中心词;

[0066] c表示上下文关键词;

[0067] n表示数据集中的推文总数;

[0068] $u_i(c, k)$ 是一个包含条件,表示第i条推文是否包含c和k,如果是,则 $u_i(c, k) = 1$,否则为0;

[0069] $u_i(c, -k)$ 表示第i条推文是否包含c但不包含k,如果是,则 $u_i(c, -k) = 1$,否则为0;

[0070] η_i 表示第i条推文是否被转发,如果是,则 $\eta_i = 1$,否则为0;

[0071] r_i 表示第*i*条推文被转发的数量。

[0072] 公式(2)中,分子和分母都反映了包含条件下转发数的经验估计。这种方法可以帮助描述中心词与每个上下文关键词之间的距离,如果 $\mathcal{A}(c, k)$ 为正数,说明*c*和*k*的关系越密切,公众关注度越高,如果是负数,就意味着它们的关系不密切,公众的关注度越低。

[0073] 2、成对相关性的计算

[0074] 每一个事件都必然有两个关键的主体,是讨论的焦点,对舆论走向产生很大影响。本发明将两个数据流与关键词的相关性根据它们的共现频率进行量化(即“关联度”),记为 C_i^t ,并且 $C_i^t \in [0,1]$ 。

[0075] 本发明创新提出了 C_i^t 的计算方法, C_i^t 通过如下公式(1)计算而得:

$$[0076] \quad C_i^t = \frac{\text{Rank}(\beta_{iA}^t - \beta_{iB}^t)}{N^t}, \forall i \in W^t \text{ 公式(1)}$$

[0077] 公式(1)中, β_{iA}^t 、 β_{iB}^t 分别表示上下文关键词*i*和关键实体A、关键实体B在时刻*t*的共现频率;

[0078] Rank表示计算了中心词*i*的共现频率差 $\beta_{iA}^t - \beta_{iB}^t$ 在所有 $i \in W^t$ 中的排名;

[0079] N^t 表示时刻*t*下中心词的上下文关键词总数;

[0080] W^t 表示时刻*t*下中心词的上下文关键词集合。

[0081] 如果 C_i^t 接近于1,则上下文关键词*i*在时刻*t*与关键实体A或关键实体A所在的数据流更相关。

[0082] $\text{Rank}(\beta_{iA}^t - \beta_{iB}^t)/N^t$ 的计算方式举例如下:

[0083] 比如上下文关键词*i*为“apple”,该关键词与关键实体A(如人名A)的共现频率 β_{iA}^t 为10,与关键实体B(如人名B)的共现频率 β_{iB}^t 是5,则 $\beta_{iA}^t - \beta_{iB}^t = 5$ 。假设,除了“apple”还有4个单词有这样的共现频率差,且根据共现频率差的值由大到小排列,“apple”的共现频率差的值排名第二,则“apple”与关键实体A的关联度 $C_i^t = 2/5$ 。

[0084] 3、生成上下文语义序列模式

[0085] 为更加简要的总结原始推文的信息,本发明设定语义序列由动词、名词和形容词组成,序列长度可以为4(4个单词)或进一步调整。重复出现的序列为一个序列模式,模式的生成过程是一个搜索过程,搜索过程具体如下:

[0086] 假定,用户选定的中心词和上下文关键词分别记为centralkeyword和*w*,经过上述关联度 C_i^t 的计算后,与该中心词具有关联度排名前*n*的上下文关键词形成关键词集合。首先,形成初始序列,初始序列包括一个中心词centralkeyword和一个上下文关键词*w*,形成一个二元组。初始序列中的中心词centralkeyword和上下文关键词*w*的排序顺序与原本在推文中的出现顺序一致,为*w*-central keyword或central keyword-*w*。

[0087] 然后,遍历关键词集合中的每个上下文关键词,查找在初始序列中新加入集合中的一个单词后使得形成的语义新序列中的词在推文中的共现频率最大的单词,将这个单词

确定为最终从关键词集合中取出并新加入到初始序列中的上下文关键词,新加入上下文关键词后,初始序列的二元组形成变更为三元组形式,实现了对初始序列的扩充。为更灵活地设置语义序列对推文文本的覆盖度,本发明还添加了一个跳过值,即允许新加入的上下文关键词和当前元组中关键词的位置关系在跳过值范围内波动。进一步地,跳过值根据序列长度 l 设置为 $\frac{l}{2} + 1$ 。举例而言,假设序列长度 l 为20,跳过值则为11,对当前元组而言上一个新加入的单词的位置是5,则允许新加入的上下文关键词和当前元组中的关键词的位置关系的波动范围为当前元组中1到16位。通过跳过值的设置允许序列根据文本的长度调整关键词的相对距离,因此更灵活地设置了语义序列对推文文本的覆盖度。

[0088] 4、基于生成的语义序列模式对不同数据流进行数据分析

[0089] 与静态设置相比,流数据分析面临许多困难,一方面是计算效率要更快且更精准,另一方面需要灵活地可视化支持。但由于事件的主题存在不断变化、继承、消逝等特点,因此聚类会更加复杂。为了解决这个问题,本发明采用基于BERTopic和KMeans++的动态聚类方法,以即时的处理连续更新的推文,处理方法具体如下:

[0090] 首先,应用BERTopic模型(BERTopic是一种主题建模技术,利用Transformer和c-TF-IDF创建密集的集群,允许解释主题,同时在主题描述中保留重要的单词)来生成文档在高维空间中的语义向量,通过UMAP(Uniform Manifold Approximation and Projection,是一种新的降维流形学习技术。UMAP是基于黎曼几何和代数拓扑的理论框架构建的,它假设可用数据样本均匀分布在拓扑空间中,可以从这些有限数据样本中近似并映射到低维空间)进一步降维便于后续计算。由于BERTopic模型不支持流数据集中的动态聚类,因此本发明将其与KMeans++算法相结合,KMeans++是适用于流数据的最快的聚类算法之一。使用KMeans++算法获取各事件的主题簇后,使用基于类的TF-IDF向量生成主题表示。

[0091] 在流数据模式下不使用基于Word2Vec的方法的原因包括两个方面。首先,词向量的生成依赖于每分钟的语料库,但同一词在每分钟数据中的向量会发生变化。因此,聚类中心无法传递到下一代,除非设置一个大的滑动窗口,并将整个窗口视为一个词袋。但是,这种方法会带来与真实时间的时差。因此,本发明使用基于Transformer的预训练模型,每分钟可以产生相同的词向量。因此,聚类中心可以传递到下一分钟,实现实时聚类,获得连贯的主题。其次,基于Word2Vec的方法需要初始关键词来提取相似度高的词,需要事先知道事件的主题。因此需要一种自动聚类方法来帮助用户了解即将到来的主题。因此本发明采用BERTopic+KMeans++的方法来对连续更新的推文进行动态聚类。

[0092] 以下将详细介绍BERTopic+KMeans++的方法对连续更新的推文进行动态聚类的过程:

[0093] 动态KMeans算法的基本原理是用最后一次聚类结果初始化聚类中心,当数据在一分钟内到达时,首先使用KMeans++来初始化聚类 C_t^k 。在第一次聚类完成后,将聚类中心传递给下一分钟的聚类 C_{t+1}^k ,以维护上一步得到的信息,提高聚类效率。考虑到用户对实时变化信息的限制,设定每次聚类最多生成6个主题,每个主题下最多生成20个上下文关键词。

[0094] 为了获得连贯的主题,在每分钟聚类后,如果 C_{t+1}^k 中前25%的关键词同样存在于 C_t^k 中,则将当前时刻的每个簇 C_{t+1}^k 与前一时刻的簇 C_t^k 合并,并对合并后的簇中的关键词

按照基于类的TF-IDF得分进行排序,将排名前x的上下文关键词形成的集合作为数据更新后的 C_{t+1}^k 。如果 C_{t+1}^k 中前25%的关键词不存在于 C_t^k 中,则不对 C_{t+1}^k 进行数据更新。

[0095] 采用上述的方法,完成对BERTopic模型在所有时刻识别到的上下文关键词的聚类,并将最终合并的簇中的前y个上下文关键词所在的推文作为待进行可视分析的对象。

[0096] 对于采用的上述BERTopic模型+KMeans++的聚类方法的性能,本发明在标准数据集(20Newsgroups数据集:是大约20,000个新闻组文档的集合,平均(几乎)划分在20个不同的新闻组中,已经成为机器学习技术的文本应用实验的流行数据集,例如文本分类和文本聚类)评估了BERTopic+KMeans++和BoW+KMeans++(词袋模型(Bag of words),将所有词语装进一个袋子里,不考虑其词法和语序的问题,即每个词语都是独立的。BERTopic是词向量模型,是考虑词语位置关系的一种神经网络模型。通过大量语料的训练,将每一个词语映射到高维度(几千、几万维以上)的向量当中)的方案,该数据集包含18846条数据。本发明测试了NMI(归一化互信息,NMI为互信息是相同数据的两个标签之间相似性的度量),

$$[0097] \quad MI(U, V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} \frac{|U_i \cap V_j|}{N} \log \frac{N|U_i \cap V_j|}{|U_i||V_j|}$$

[0098] 其中 $|U_i|$ 为簇中样本个数, $|V_j|$ 为簇中样本个数,簇中U与V的互信息如下:归一化互信息(NMI)是互信息(MI)评分的归一化,将结果在0(无互信息)和1(完全相关)之间缩放。

[0099] 根据类标签来判断结果聚类的质量。每种方法我们运行5次来计算中值,发现BERTopic+KMeans++的中值NMI为0.61[0.60-0.62],BoW+KMeans++的中值NMI为0.42[0.39-0.43],这表明在聚类结果方面BERTopic+KMeans++优于BoW方法。至于计算效率的问题,在案例研究的数据集上测试了BERTopic+KMeans++方法,平均每分钟包含80条推文。发现BERTopic+KMeans++可以在6-7秒内处理1分钟的数据。因此,对于许多每分钟推文数在800条左右的社交媒体事件数据集,BERTopic+KMeans++的方法在聚类效果和时间效率上都是可行的。

[0100] 二、可视化分析结果

[0101] 以下将介绍语义序列模式视图中的翅膀隐喻的设计原理、视觉编码和具体构建过程(如图4所示):

[0102] 本发明提出了一个新的设计,可以用于可视化上下文的不断变化的顺序模式,并允许交互来比较它们。在ContextWing中,主要的隐喻是翅膀和羽毛,如图8所示。

[0103] 翅膀隐喻:翅膀可视化了中心词的序列模式之间的连接。在水平方向上,翅膀被分为左右对称的结构。左边翅膀上的词语表示在原文中出现在中心词前面的词语,反之亦然。

[0104] 羽毛隐喻:翅膀中每一对水平对称的羽毛(也称为每一层)展示了根据相同的上下文关键词合并的序列模式。羽毛的颜色和垂直位置代表了每个上下文关键词同两个关键实体之间的相关性。水平位置代表公众关注度。

[0105] 接下来将介绍如何构建语义序列模式视图。

[0106] 1. 为选定的上下文关键词构建上述的羽毛层。本发明将每个选择的上下文关键词分配到一个具有相同长度的层,并且宽度可以根据所选单词的数量自动微调。层的垂直位置和颜色编码表示成对的比较。使用层的颜色和垂直位置来编码计算模型生成的成对相关

性。为了便于表达成对相对性,如图8所示,位置越靠下的层和关键实体A相关,反之越接近关键实体B。层次的水平位置是基于公众关注度的量化结果,它表明公众对中心词及其选择的上下文关键词的关注。如果该层水平方向上更接近中心关键词,就意味着它们拥有更大的注意力。此外,到该层的链接的宽度表示该层上的模式的总频率。

[0107] 2. 在每个层上布局上下文关键词。如图8中的a所示,本发明将单词按出现顺序放置在中心关键词左右的羽毛上,并按时间顺序从层的顶部到底部排列模式。本发明将上下文关键词与中心词垂直对齐。与中心词在同一水平线上的关键词形成一个模式。层边的时间刻度表示同一行中图案的对应时间段。关键词的大小编码了包含该单词后的模式频率。因此,最后一个关键词频率表示模式频率。

[0108] 3. 合并选定的上下文关键词。在放置过程中,发现在同一列中有许多重复的关键词,使得不容易比较序列中不同的语义信息。例如,选中的上下文关键词如“flu”不会明显显示出来,因为“pandemic”也会重复出现,而且更接近中心(图9中的a)。因此,需要避免其他上下文关键词的影响,并强调所选的层的上下文关键词。如图9中的b所示,本发明将这些关键词合并到同一列中,保持了整体结构,避免了误解。合并后,上下文关键词的频率演化信息将丢失。因此,本发明还增加了一个迷你走势图,以可视化词频随时间的变化,以增强信息的展示。

[0109] 4. 添加词语之间的连接线。利用树形结构的思想,本发明通过添加线来连接相同模式的单词,以便更好地理解。如图9中的(a)表明,存在上下文关键词是模式的最终词的情况(图9中的a“flu”)。如果合并重复的单词,这个单词的位置可能会变成空白,就会看起来好像没有单词在正确的层级上,会造成误解。因此,本发明添加线条来连接上下文关键词和水平线上的下一个空白位置,以表示上下文关键词的存在,如图9所示。

[0110] 本发明提供的ContextWing系统包括图10中所示的主题视图、控制视图、模式视图和细节视图。

[0111] 1. 主题视图

[0112] 本发明提供一个主题视图来选择关键词作为模式视图的输入。如图11所示,顶部视图是一个直方图,显示两个数据流的推文的变化百分比。两个流的符号分别放置在视图的顶部和底部。主题用不同的颜色标示在左下方的按钮上。在气泡图中,关键词被聚合并划分为几个时段。由于每个关键词实际上都可以生成一个模式翼结构,本发明将关键词气泡设计为一个翼状雕文。大小和不透明度表示单词出现的频率,颜色表示单词所属的主题。气泡的垂直位置代表与两个关键角色的相关性。一些重要的指标,如频率和情绪分布,显示在工具提示。为了直观地观察主题的连贯性,本发明为在不同阶段频繁出现的关键词添加了连接线。用户可以将关键词悬停在屏幕上观察频率和相关度。直方图可以通过刷屏来选择时间段,所选时间段的数据将被重新聚合,并显示在多个池中。主题视图的设计也可以扩展到流设置。直方图、气泡池和主题按钮以预设的间隔同步更新(例如,1分钟)。根据建模结果,如果出现新的主题,旧的主题将被替换并用新的颜色突出显示。主题按钮的颜色和名称总是与更新气泡的类别相对应,可以帮助用户更直观地感知主题的动态变化。在动态变化的情况下,用户很难对之前的信息在脑海中保持一幅地图。因此,本发明将直方图和气泡图结合起来,可以帮助用户查看实时历史数据。用户也可以点击“暂停”按钮来暂停/继续更新。

[0113] 2. 控制视图

[0114] 本发明设置了数据集选项和分析模式,用户可以选择在静态和流分析模式之间切换。此外,从主题视图(图11所示)开始,有两种方法可以探索(每个上下文词还可以作为中心词有上下文关键词,所以在探索模式下可以继续往下点击每一个关键词)中心词及其上下文关键词。用户可以点击“更改模式”,打开迭代控制面板(图3所示):(1)探索模式:用户可以通过点击,不断向下钻取一个中心词的上下文关键词。(2)分析模式:用户可以点击一个关键词作为中心词,然后选择其上下文关键词。为了保持信息的一致性,控制视图中所选关键词的颜色仍然表示其主题。然后,点击“Go Pattern”可以在右边的模式视图中观察到派生的模式。例如,图1显示了分析模式下的选择操作,它支持对中心词的上下文关键词选择。在这个过程中,用户可以重新点击气泡来更新选择,点击“返回”和“重启”到之前或初始状态,进行迭代探索。

[0115] 3. 模式视图

[0116] 一旦构建出机翼结构,用户可以从不同方面进行比较。本发明提供了以下四种交互方式来进行详细的比较。首先,为了支持相同上下文关键词的模式的时间比较,用户可以悬停任何关键词,相应的模式将被突出显示,而其他模式将被隐藏。因此,用户可以更好地观察单层上的时间标记的单个模式。第二,从所选的上下文关键词的角度来比较模式。用户可以点击任意时间刻度,突出显示该时间段内不同层次的图案。此外,当用户将鼠标悬停在层侧面时,会显示出一个迷你走势图(用没有轴的线表示的小数据图),表示所选上下文关键词在整个周期内的演化频率。最后,本发明还提供了一个工具提示,让用户可以点击任何关键词来查看每个图案的频率和情感分布。模式视图还支持实时更新,以与主题视图中相同的时间间隔显示模式,垂直排列以对应当前时刻的几个时刻。

[0117] 4. 细节视图

[0118] 为了帮助用户理解模式,我们提供了一个细节视图(图7),可以显示原始推文的时间、情绪评分等信息。在模式视图中,用户可以选择一个模式,原始推文将显示在细节视图中。此外,用户可以选择一个时间段和键入他们感兴趣的单词。

[0119] 需要声明的是,上述具体实施方式仅仅为本发明的较佳实施例及所运用技术原理。本领域技术人员应该明白,还可以对本发明做各种修改、等同替换、变化等等。但是,这些变换只要未背离本发明的精神,都应在本发明的保护范围之内。另外,本申请说明书和权利要求书所使用的一些术语并不是限制,仅仅是为了便于描述。

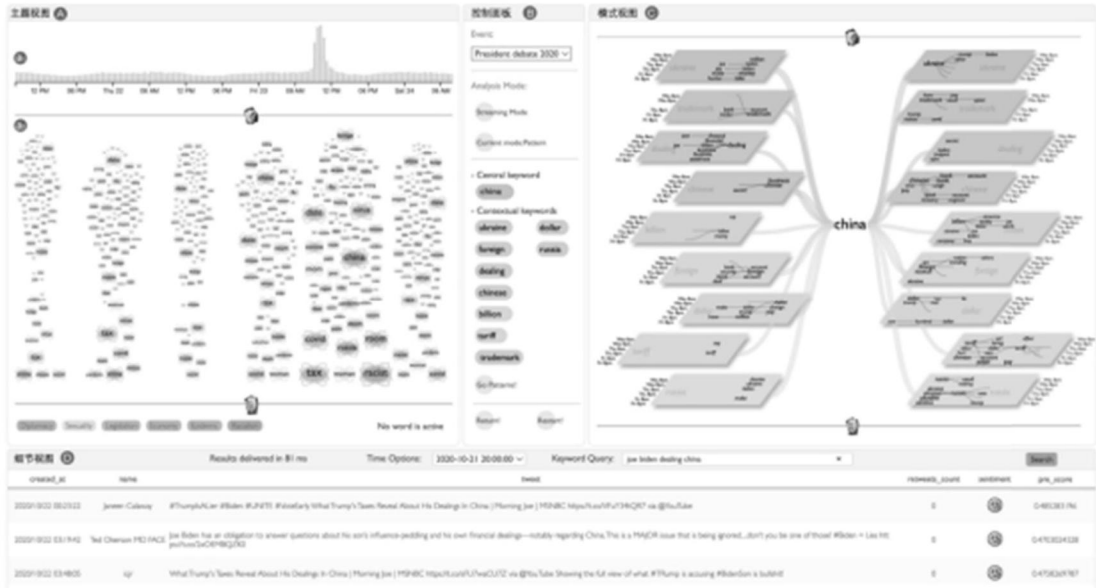


图1

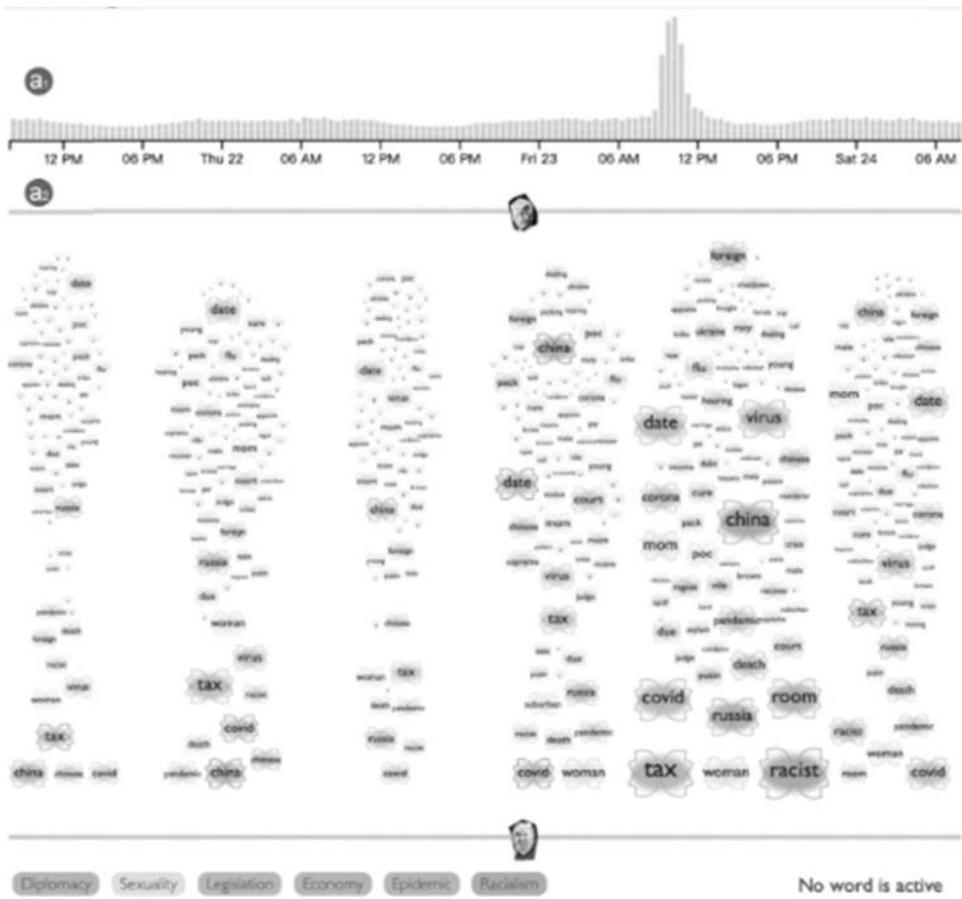


图2



图3

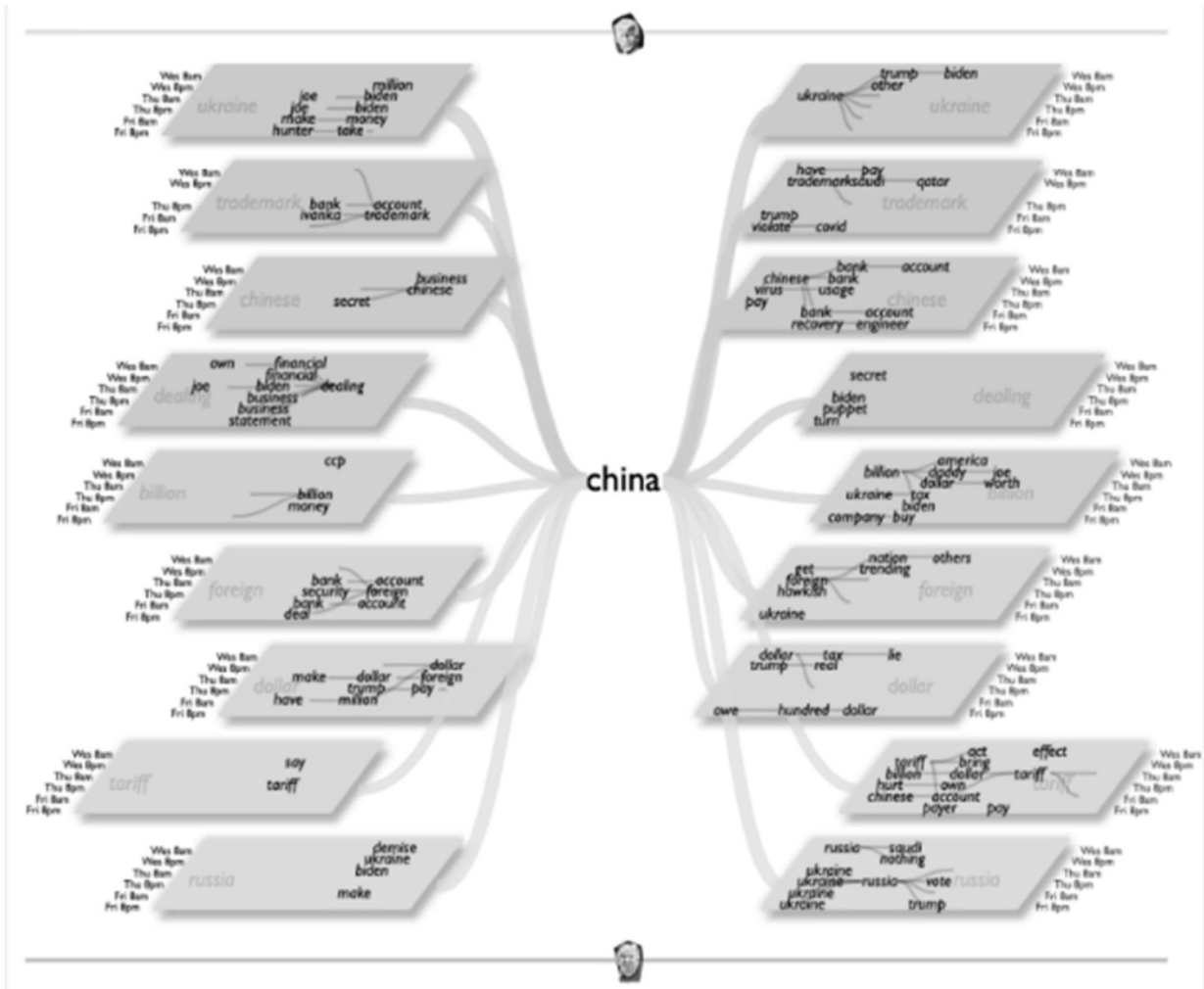


图4

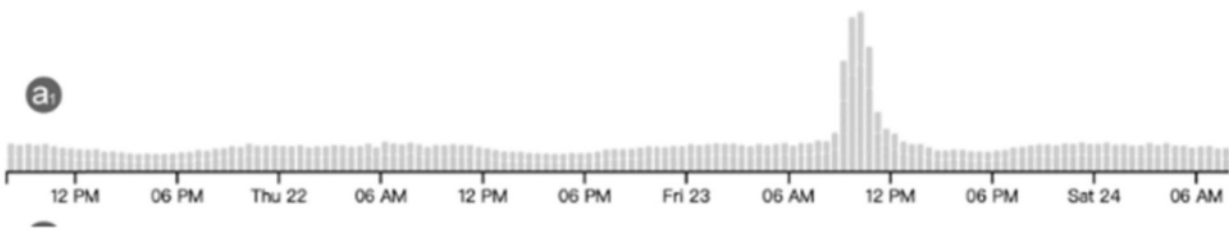


图5

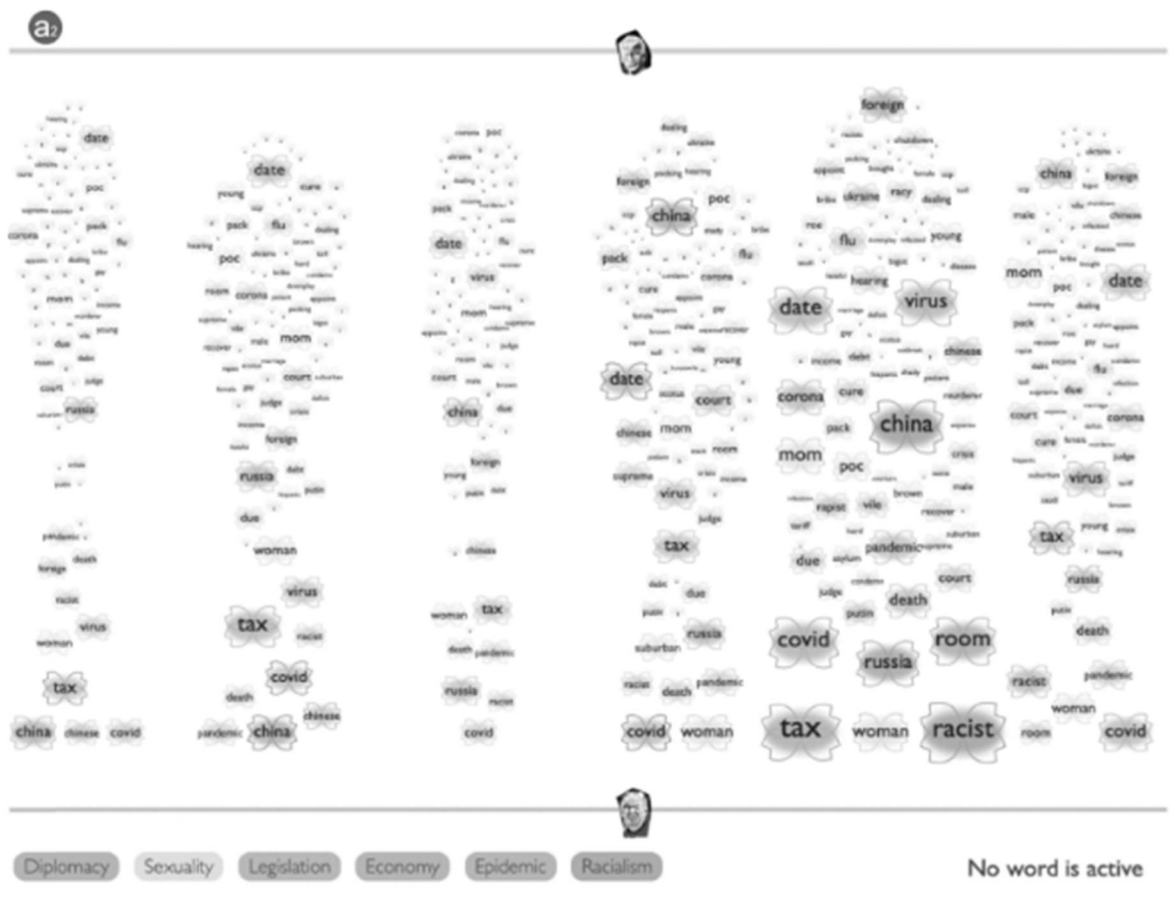


图6

created_at	name	tweet	retweets_count	sentiment	pre_score
2020-10-21 08:53:10	Jackie Duval	If Hunter Biden's name was Donald Trump & the emails and texts and corrupt business dealings would be ALL over the news! Bio bias bias! #Biden #Democrats #Hilary #Benghazi #HillaryEmails #SCOTUShearings #OscarsBiden #HunterLapTop #HunterBidenCoverUp #JoeBiden #Trump2020	0	👎	0.492721429
2020-10-21 09:37:53	Dr. Michael K. Crane	Joe Biden has an obligation to answer questions about his sons' influence peddling and his own financial dealings—notably regarding China. https://t.co/7uA916rY @NYC #Biden #Factor2020 #Corruption #HunterBiden #China #Russia #FBI #GOP #Trump	0	👎	0.652288634
2020-10-21 09:37:28	StopFakeClaims	Everyday despises #Trump links to a new law Today he asked the DOJ to investigate Joe Biden, which is proof beyond any reasonable doubt of #Trump's authoritarian tendency! And we learned about his secret bank account in China showing his shady business dealings! #BidenInDut	1	👎	0.444198726
2020-10-21 09:41:04	Thomas Sabin	@JoeBiden, in his public financial disclosures, along with the income tax returns he voluntarily released, show no income or business dealings of his own in #China. #Trumpas #BidenWasTrue #BidenIsNot2020 #JoeBidenIsIn	0	👎	0.444198726

图7

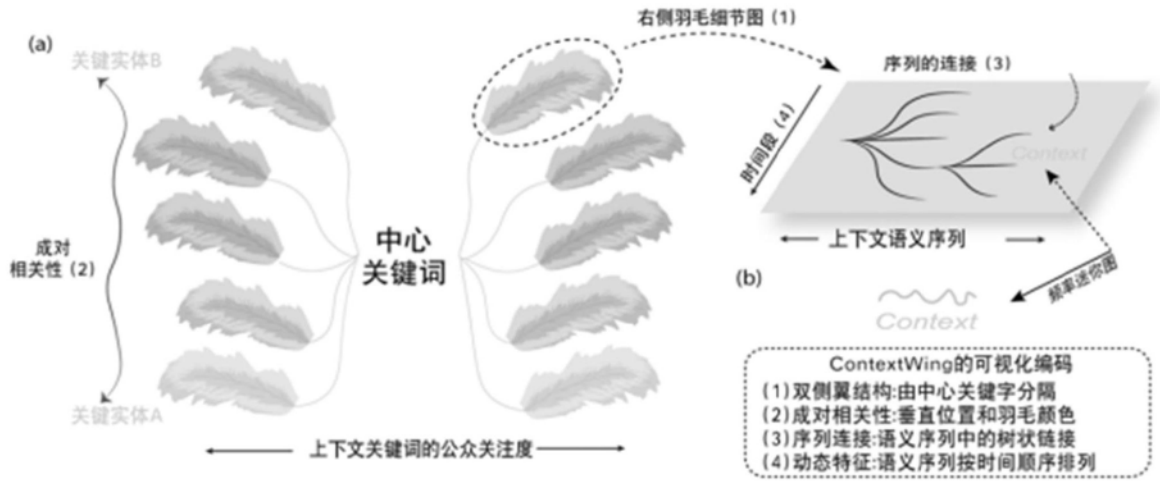


图8

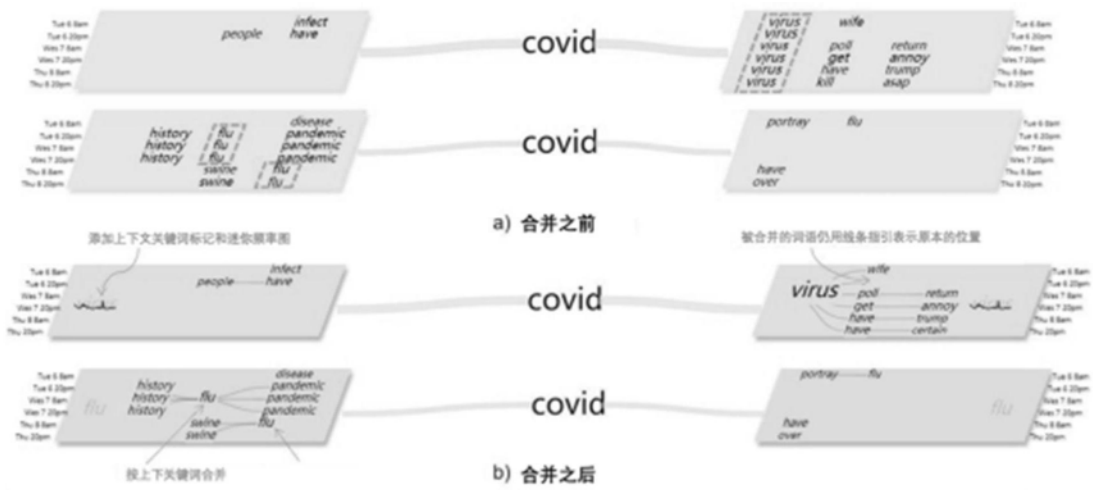


图9

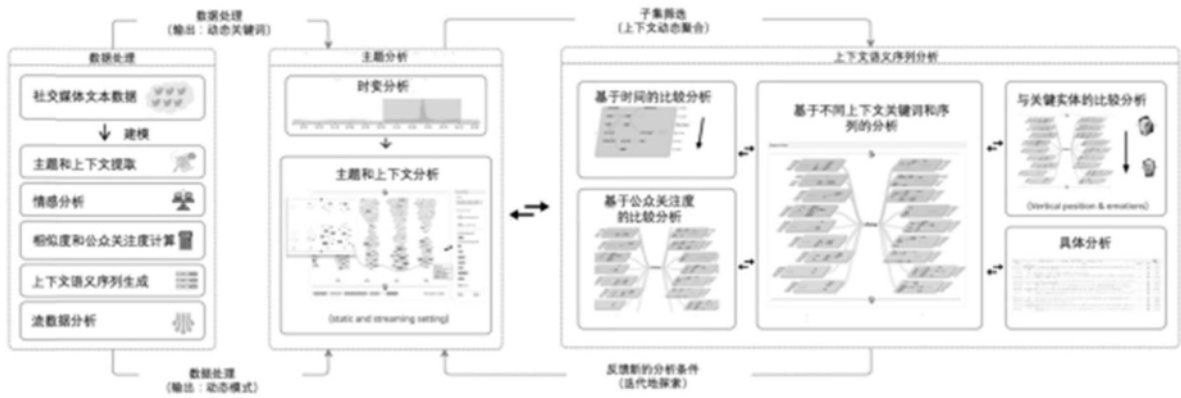


图10

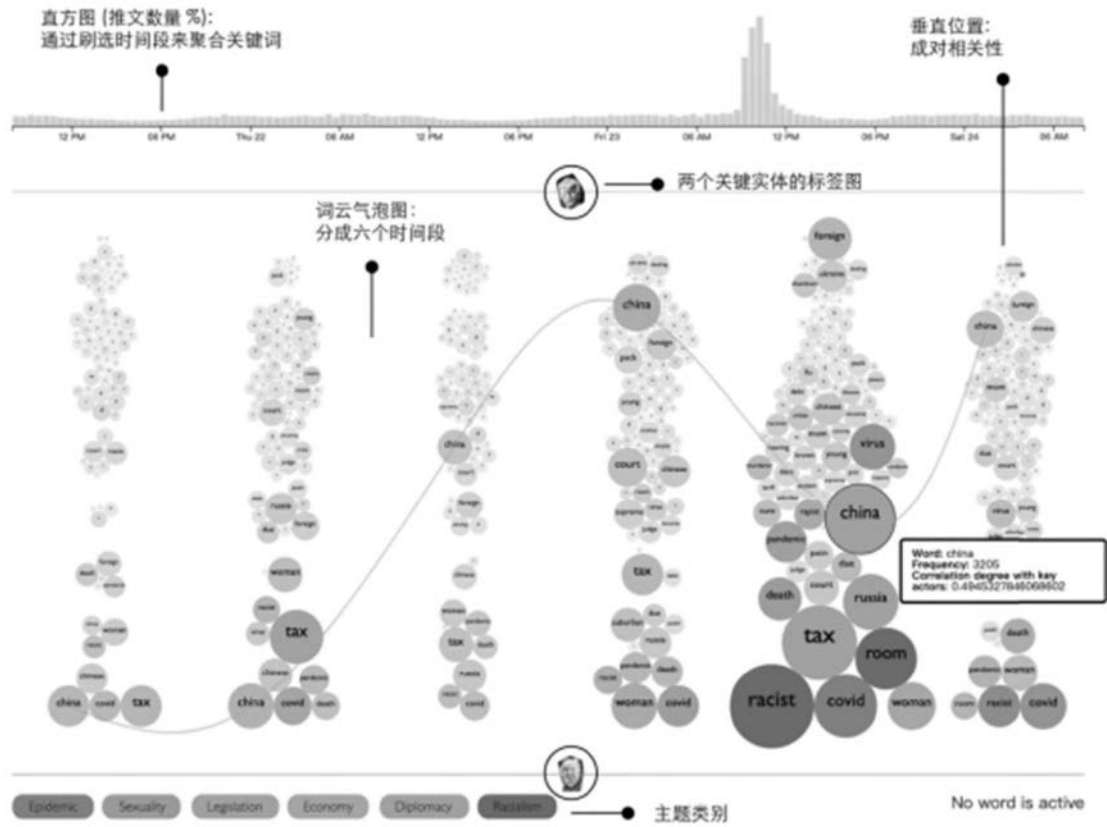


图11