



(12)发明专利申请

(10)申请公布号 CN 106295702 A

(43)申请公布日 2017.01.04

(21)申请号 201610668449.4

(22)申请日 2016.08.15

(71)申请人 西北工业大学

地址 710068 陕西省西安市友谊西路127号

(72)发明人 於志文 马超 王柱 郭斌

(74)专利代理机构 西安利泽明知识产权代理有限公司 61222

代理人 刘伟

(51)Int.Cl.

G06K 9/62(2006.01)

G06Q 50/00(2012.01)

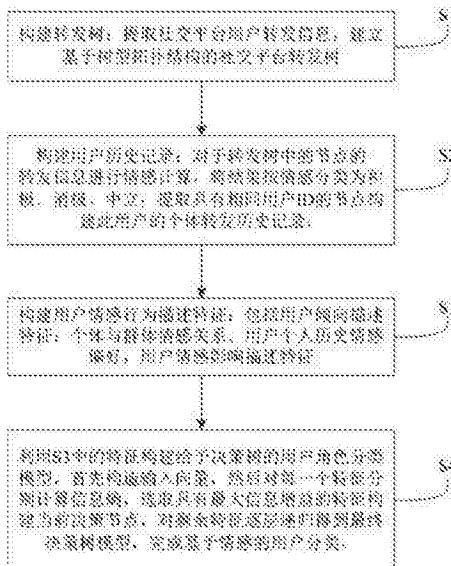
权利要求书2页 说明书6页 附图5页

(54)发明名称

一种基于个体情感行为分析的社交平台用户分类方法

(57)摘要

本发明公布了一种基于情感行为分析的社交平台用户分类方法,包括以下步骤:S1、构建转发树;S2、构建用户历史记录,提取具有相同用户ID的节点构建此用户的个体转发历史记录;S3、构建用户情感行为描述特征;S4、利用S3中的特征构建给予决策树的用户角色分类模型,完成基于情感行为分析的社交平台用户分类。本发明构建了较为全面的用户情感行为描述模型,可以更全面的考虑用户的个人历史信息;该方法充分利用了微博当中的用户个人信息,传播结构信息,情感信息以及动态时域信息。由于采用以上措施,本发明能够获得更好的分类准确率。



1. 一种基于个体情感行为分析的社交平台用户分类方法,包括以下步骤:

S1、构建转发树:提取社交平台用户转发信息,建立基于树型拓扑结构的社交平台转发树;

S2、构建用户历史记录:对于转发树中的节点的转发信息进行情感计算,将结果按情感分类为积极、消极、中立;提取具有相同用户ID的节点构建此用户的个体转发历史记录;

S3、构建用户情感行为描述特征:包括用户倾向描述特征:个体与群体情感关系 ER_u 、用户个人历史情感偏好 HP_u ;用户情感影响描述特征 EI_u ;

S4、利用S3中的特征构建给予决策树的用户角色分类模型,首先构造输入向量 $U_u = \langle ER_u, HP_u, EI_u \rangle$,然后对每一个特征分别计算信息熵 $Entropy(U) = \sum_{j=1}^3 -U^j \log_2 U^j$, U^j 为第j个特征,选取具有最大信息增益的特征构建当前决策节点,对剩余特征逐层递归得到最终决策树模型,进而完成基于情感的用户分类。

2. 根据权利要求1所述的一种基于个体情感行为分析的社交平台用户分类方法,其特征在于:所述的S1中的转发信息包括原始文本信息、转发文本信息、参与用户的个体信息。

3. 根据权利要求1所述的一种基于个体情感行为分析的社交平台用户分类方法,其特征在于:所述的S1按照层级由底向上进行文本情感解析,逐层添加转发节点,构建转发树。

4. 根据权利要求1所述的一种基于个体情感行为分析的社交平台用户分类方法,其特征在于:所述的S2中的情感计算采用多规则集模型,通过文本点互信息自底向上建立情感词典、语法规则,所述的自底向上是指按照从词语、短语、短句、整句的顺序依次分析。

5. 根据权利要求1所述的一种基于个体情感行为分析的社交平台用户分类方法,其特征在于: S3中所述的个体和群体情感关系是基于个体的情感选择与群体情感的分布,描述为个体与当前一条文本信息的情感关系因子 $ER_u(w)$,其取值范围为 $-1 \sim 1$,该值越大表示当前关系越趋近积极,该值越小表示当前关系越趋近消极,如下表示:

$$ER_u(w) = \begin{cases} 0.5 + \frac{N(w) - P(w) - |N(w) - O(w)|}{2S(w)}, & E_u(w) = P \\ \frac{P(w) - N(w)}{2S(w)}, & E_u(w) = O \\ -0.5 + \frac{N(w) - P(w) - |P(w) - O(w)|}{2S(w)}, & E_u(w) = N \end{cases}$$

其中, $N(w)$, $P(w)$, $O(w)$ 分别表示当前转发树内的消极情感分布,积极情感分布,中立情感分布, $S(w)$ 表示转发树规模。

6. 根据权利要求1所述的一种基于个体情感行为分析的社交平台用户分类方法,其特征在于:所述的S3中个体历史情感偏好 $HP_u(e)$ 是基于用户历史记录中的情感分布以及历史转发中的用户评论参与度 $C_u(w)$,用以下公式表示:

$$HP_u(e) = \sum_{\substack{E_u(w)=e, \\ w \in N_u}} \log(C_u(w) + 2) \exp\{-\theta_1(t_0 - t_w)\}$$

其中, $\exp\{-\theta_1(t_0 - t_w)\}$ 为控制用户偏好的时间衰减, $\log(C_u(w) + 2)$ 为通过评论长度描述用户的参与程度。

7. 根据权利要求1所述的一种基于个体情感行为分析的社交平台用户分类方法,其特征在在于:S3中所述的情感影响 EI_u 是基于转发树的结构特点 $SF_u(w)$ 、转发树的时域影响 $TF_u(w)$ 、用户的情感变化 $EF_u(w)$,如下表示:

$$EI_u = \frac{\sum_{w \in W_u} \alpha SF_u(w) + \beta TF_u(w) + \gamma EF_u(w)}{HR_u} \left(1 + \frac{HR_u}{HR_u + NR_u} \right)$$

HR_u 表示用户转发作为内部节点的个数, NR_u 表示用户转发作为叶子节点的个数。

8. 根据权利要求7所述的一种基于个体情感行为分析的社交平台用户分类方法,其特征在在于:所述的转发树的结构特点 $SF_u(w)$ 基于转发树的绝对规模 $S(w)$ 、相对规模 $S_u(w)$ 以及子树深度 $DP_u(w)$,如下表示:

$$SF_u(w) = \alpha_1 \frac{S_u(w)}{S(w)} + (1 - \alpha_1) \exp \left\{ -\delta \frac{DP_u(w)}{\log S_u(w)} \right\}$$

9. 根据权利要求7所述的一种基于个体情感行为分析的社交平台用户分类方法,其特征在在于:所述转发树的时域影响 $TF_u(w)$ 为转发树在时间角度对信息传播的贡献,所述贡献体现在子树相对于整个转发树的存活时间、子树相对于原始文本的时间延迟两个方面;

$$TF_u(w) = \beta_1 \frac{LP_u(w)}{LP(w)} + (1 - \beta_1) \exp \{-\varepsilon(t_u - t_w)\}$$

其中 $LP_u(w)$ 为子树生命周期, $LP(w)$ 为转发树生命周期, $\frac{LP_u(w)}{LP(w)}$ 为子树相对于整个转发树的存活时间, $\exp\{-\varepsilon(t_u - t_w)\}$ 为子树出现的时域延迟。

10. 根据权利要求7所述的一种基于个体情感行为分析的社交平台用户分类方法,其特征在在于:所述的用户的情感变化 $EF_u(w)$ 以当前用户的转发行为作为时间分界点,通过计算用户转发前后的情感分布差异,并通过指数函数对参数进行标准化,用以下公式表示:

$$EF_u(w) = -\exp \left\{ \sum_{e \in E} |B_u(w, e) - A_u(w, e)| \right\} + 1$$

其中, $B_u(w, e)$, $A_u(w, e)$ 分别为用户转发前后的情感分布。

一种基于个体情感行为分析的社交平台用户分类方法

技术领域

[0001] 本发明属于社交网络技术领域,特别涉及一种基于个体情感行为分析的社交平台用户分类方法。

背景技术

[0002] 随着互联网技术的发展,以微博为代表的在线社交网络得到大规模的使用。用户可以在其上自行发布信息,也可以通过转发,评论,点赞等方式与其它信息进行交互,与真实社交网络相同,在线社交网络的用户行为传达出的不仅仅是字面信息,它同时包含着用户的情感态度,这种情感态度因用户个人背景与习惯的不同而不同,并贯穿于用户的所有交互行为当中,我们把用户所具有的这种情感特征称之为用户的情感角色。

[0003] 目前针对在线社交网络用户的研究主要包括以下几个方面,1、用户影响力的挖掘,此类研究着力于通过对用户个人属性以及信息传播特征的分析,建立描述用户社交影响力的模型或算法,实现用户影响力计算,发现社交领导者;2、用户在线行为的预测,此类研究通过对用户历史,上下文环境以及社交关系等因素的考虑对用户进行建模,实现对用户特定行为或偏好的预测,例如是否参与转发,是否感兴趣等。3、用户情感分析,此类研究以某一个时刻用户会有怎样的情感作为出发点,通过多种数据源(包括文本,图片,视频,音乐等),线上线下结合以及社交关系等因素实现用户情感的分析与预测。以上研究在一定程度上为我们揭示了用户的在线行为规律和社交网络的内在运作规律,但缺乏对用户情感的综合考虑。

发明内容

[0004] 针对以上问题,本发明通过从用户个人情感角度进行分析,提供一种基于个体情感行为分析的社交平台用户分类方法,具体技术方案为:

一种基于个体情感行为分析的社交平台用户分类方法,包括以下步骤:

S1、构建转发树:提取社交平台用户转发信息,建立基于树型拓扑结构的社交平台转发树;

S2、构建用户历史记录:对于转发树中的节点的转发信息进行情感计算,将结果按情感分类为积极、消极、中立;提取具有相同用户ID的节点构建此用户的个体转发历史记录;

S3、构建用户情感行为描述特征:包括用户倾向描述特征:个体与群体情感关系 ER_u 、用户个人历史情感偏好 HP_u ;用户情感影响描述特征 EI_u ;

S4、利用S3中的特征构建给予决策树的用户角色分类模型,首先构造输入向量 $U_u = \langle ER_u, HP_u, EI_u \rangle$,然后对每一个特征分别计算信息熵 $Entropy(U) = \sum_{j=1}^3 -U^j \log_2 U^j$, U^j 为第j个特征,选取具有最大信息增益的特征构建当前决策节点,对剩余特征逐层递归得到最终决策树模型,进而完成基于情感的用户分类。

[0005] 进一步地,一种基于个体情感行为分析的社交平台用户分类方法S1中的转发信息包括原始文本信息、转发文本信息、参与用户的个体信息。

[0006] 进一步地,一种基于个体情感行为分析的社交平台用户分类方法S1按照层级由底向上进行文本情感解析,逐层添加转发节点,构建转发树。

[0007] 进一步地,一种基于个体情感行为分析的社交平台用户分类方法S2中的情感计算采用多规则集模型,通过文本点互信息自底向上建立情感词典、语法规则,所述的自底向上是指按照从词语、短语、短句、整句的顺序依次分析。

[0008] 进一步地,一种基于个体情感行为分析的社交平台用户分类方法S3中所述的个体和群体情感关系是基于个体的情感选择与群体情感的分布,描述为个体与当前一条文本信息的情感关系因子 $ER_u(w)$,其取值范围为 $-1\sim 1$,该值越大表示当前关系越趋近积极,该值越小表示当前关系越趋近消极,如下表示:

$$ER_u(w) = \begin{cases} 0.5 + \frac{N(w) - P(w) - |N(w) - O(w)|}{2S(w)}, & E_u(w) = P \\ \frac{P(w) - N(w)}{2S(w)}, & E_u(w) = O \\ -0.5 + \frac{N(w) - P(w) - |P(w) - O(w)|}{2S(w)}, & E_u(w) = N \end{cases}$$

其中, $N(w)$, $P(w)$, $O(w)$ 分别表示当前转发树内的消极情感分布,积极情感分布,中立情感分布, $S(w)$ 表示转发树规模。

[0009] 进一步地,一种基于个体情感行为分析的社交平台用户分类方法S3中个体历史情感偏好 $HP_u(e)$ 是基于用户历史记录中的情感分布以及历史转发中的用户评论参与度 $C_u(w)$,用以下公式表示:

$$HP_u(e) = \sum_{\substack{E_u(w)=e, \\ w \in W_u}} \log(C_u(w) + 2) \exp\{-\theta_1(t_0 - t_w)\}$$

其中, $\exp\{-\theta_1(t_0 - t_w)\}$ 为控制用户偏好的时间衰减, $\log(C_u(w) + 2)$ 为通过评论长度描述用户的参与程度。

[0010] 进一步地,一种基于个体情感行为分析的社交平台用户分类方法S3中所述的情感影响 EI_u 是基于转发树的结构特点 $SF_u(w)$ 、转发树的时域影响 $TF_u(w)$ 、用户的情感变化 $EI_u(w)$,如下表示:

$$EI_u = \frac{\sum_{w \in W_u} \alpha SF_u(w) + \beta TF_u(w) + \gamma EF_u(w)}{HR_u} \left(1 + \frac{HR_u}{HR_u + NR_u}\right)$$

HR_u 表示用户转发作为内部节点的个数, NR_u 表示用户转发作为叶子节点的个数。

[0011] 进一步地,一种基于个体情感行为分析的社交平台用户分类方法中转发树的结构特点 $SF_u(w)$ 基于转发树的绝对规模 $S(w)$ 、相对规模 $S_u(w)$ 以及子树深度 $DP_u(w)$,如下表示:

$$SF_u(w) = \alpha_1 \frac{S_u(w)}{S(w)} + (1 - \alpha_1) \exp\left\{-\delta \frac{DP_u(w)}{\log S_u(w)}\right\}。$$

[0012] 进一步地,一种基于个体情感行为分析的社交平台用户分类方法中转发树的时域影响 $TF_u(w)$ 为转发树在时间角度对信息传播的贡献,所述贡献体现在子树相对于整个转发树的存活时间、子树相对于原始文本的时间延迟两个方面;

$$TF_u(w) = \beta_1 \frac{LP_u(w)}{LP(w)} + (1 - \beta_1) \exp\{-\varepsilon(t_u - t_w)\}$$

其中 $LP_u(w)$ 为子树生命周期, $LP(w)$ 为转发树生命周期, $\frac{LP_u(w)}{LP(w)}$ 为子树相对于整个转发树的存活时间, $\exp\{-\varepsilon(t_u - t_w)\}$ 为子树出现的时域延迟;

进一步地,一种基于个体情感行为分析的社交平台用户分类方法中用户的情感变化 $EF_u(w)$ 以当前用户的转发行为作为时间分界点,通过计算用户转发前后的情感分布差异,并通过指数函数对参数进行标准化,用以下公式表示:

$$EF_u(w) = -\exp\left\{\sum_{e \in M} |B_u(w, e) - A_u(w, e)|\right\} + 1$$

其中, $B_u(w, e)$, $A_u(w, e)$ 分别为用户转发前后的情感分布。

[0013] 本发明具有以下有益效果:

为了能够系统的描述用户在线情感行为,本发明定义了六类微博用户情感角色,分别是积极领导者,积极追随者,消极领导者,消极追随者,中立领导者,中立追随者,并提出一种基于个体情感行为分析的社交平台用户分类方法,该方法从两个维度(情感倾向与情感影响)建立用户情感行为描述模型。

[0014] 由于采用了技术方案中的用户情感倾向特征和用户影响特征,构建了较为全面的用户情感行为描述模型,可以更全面的考虑用户的个人历史信息;该方法充分利用了微博当中的用户个人信息,传播结构信息,情感信息以及动态时域信息。由于采用以上措施,本发明能够获得更好的分类准确率。

附图说明

[0015] 图1本发明一种基于个体情感行为分析的社交平台用户分类方法流程图;

图2本发明一种基于个体情感行为分析的社交平台用户分类方法用户历史记录实例;

图3本发明一种基于个体情感行为分析的社交平台用户分类方法结构特性分布;

图4本发明一种基于个体情感行为分析的社交平台用户分类方法时域特性分布;

图5本发明一种基于个体情感行为分析的社交平台用户分类方法参数学习结果;

图6本发明一种基于个体情感行为分析的社交平台用户分类方法情感变化特性分布;

图7本发明一种基于个体情感行为分析的社交平台用户分类方法个人与宏观情感关系分布;

图8本发明一种基于个体情感行为分析的社交平台用户分类方法历史情感偏好结果分布;

图9本发明一种基于个体情感行为分析的社交平台用户分类方法情感影响结果。

具体实施方式

[0016] 为了使本发明的目的及优点更加清楚明白,以下结合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。

实施例

[0017] S1、构建转发树：提取社交平台用户转发信息，建立基于树型拓扑结构的社交平台转发树。

以微博为例，抓取微博上的转发数据，保留数据当中的用户信息，转发信息以及原始微博信息，根据微博转发的标识符“//@”以及上级用户昵称，按照层级由底向上进行文本解析，逐层添加转发节点，构建微博转发树。总共收集到19389名用户信息，构建转发树7096颗。

[0018] S2、构建用户历史记录：对于转发树中的节点的转发信息进行情感计算，将结果按情感分类为积极、消极、中立；提取具有相同用户ID的节点构建此用户的个体转发历史记。

[0019] 利用多规则集模型，对转发树中每一个节点所包含的文本信息进行情感计算，得到三种结果，分别是积极，消极和中立。之后，利用每一个微博转发节点所包含的用户信息，将具有相同用户ID的节点提取出来构建用户的个人历史转发记录并以XML文件形式进行存储。图2为一个用户的历史记录示例，<uid_1796678344>代表一个用户，<retweet>为当前用户的一条转发，<org_id>、<org_text>、<org_time>、<org_emotion>、<p_name>、<p_id>、<w_id>、<w_text>、<w_time>、<w_emotion>表示对应转发的相关属性。

[0020] S3、构建用户情感行为描述特征：包括用户倾向描述特征：个体与群体情感关系 ER_u 、用户个人历史情感偏好 HP_u ；用户情感影响描述特征 EL_u 。

[0021] 从个人与宏观情感关系以及用户个人历史情感偏好两个角度构建用户情感倾向，对于前者，以 $ER_u(w)$ 表示用户与当前一条微博的情感关系因子取值范围在-1~1之间，该值越大表示当前关系越趋近积极，反之趋近消极，为使中立情感位于0附近，设定积极与消极的原点分别是0.5和-0.5，

$$ER_u(w) = \begin{cases} 0.5 + \frac{N(w) - P(w) - |N(w) - O(w)|}{2S(w)}, & E_u(w) = P \\ \frac{P(w) - N(w)}{2S(w)}, & E_u(w) = O \\ -0.5 + \frac{N(w) - P(w) - |P(w) - O(w)|}{2S(w)}, & E_u(w) = N \end{cases}$$

$N(w)$ 、 $P(w)$ 、 $O(w)$ 分别表示当前转发树内的三类情感分布(消极，积极，中立)， $S(w)$ 表示转发树规模。

[0022] 用户个人历史情感偏好 $HP_u(e)$ 基于用户历史记录中的情感分布以及历史转发中的用户评论参与度 $C_u(w)$ ，指数部分用于控制用户偏好的时间衰减，以最近的微博发布时间 t_0 作为参考点，对数部分通过评论长度描述用户的参与程度：

$$HP_u(e) = \sum_{\substack{E_u(w)=e, \\ w \in W_u}} \log(C_u(w) + 2) \exp\{-\theta_1(t_0 - t_w)\}$$

[0023] 从转发的结构特性，时域特性以及情感变化角度描述用户情感影响，微博转发的结构特点 $SF_u(w)$ 权衡转发树的绝对规模 $S(w)$ 、相对规模 $S_u(w)$ 以及子树深度 $DP_u(w)$ ：

$$SF_u(w) = \alpha_1 \frac{S_u(w)}{S(w)} + (1 - \alpha_1) \exp\left\{-\delta \frac{DP_u(w)}{\log S_u(w)}\right\}$$

[0024] 图3描述了 $SF_u(w)$ 的计算结果分布,我们认为,在具有相同转发规模的情况下,子树越深意味着子树越稀疏,反之则越茂密,而更加茂密的子树往往具有更大范围的影响作用。

[0025] 与结构特性不同,时域影响 $TF_u(w)$ 用来描述转发树在时间角度对信息传播的贡献,这种贡献集中体现在两个方面,第一,子树相对于整个转发的存活时间;第二,子树相对于原始微博的时间延迟。 $TF_u(w)$ 综合考虑子树生命周期 $LP_u(w)$ 、转发树生命周期 $LP(w)$ 以及子树出现的时域延迟 $\exp\{-\varepsilon(t_u-t_w)\}$ 。 ε 图用于控制衰减速度:

$$TF_u(w) = \beta_1 \frac{LP_u(w)}{LP(w)} + (1 - \beta_1) \exp\{-\varepsilon(t_u - t_w)\}.$$

本方法中通过试验准确度,将其设为0.2,图4描述了 $TF_u(w)$ 的计算结果分布。 α_1 与 β_1 为学习参数,通过对特征采取单独分类验证,以0.1为步长,选择准确性最高的值作为参数实际数值,此理中采用决策树的分类方法测试结果如图5所示,因此参数值分别设为0.6和0.7。

[0026] 情感变化 $EF_u(w)$ 以当前用户的转发行为作为时间分界点,用户转发前后的情感分布分别以 $B_u(w, e)$, $A_u(w, e)$ 表示,通过 $|B_u(w, e) - A_u(w, e)|$ 计算情感分布差异,并通过指数函数对参数进行标准化:

$$EF_u(w) = -\exp\left\{\sum_{e \in M} |B_u(w, e) - A_u(w, e)|\right\} + 1$$

图6描述了 $EF_u(w)$ 的计算结果分布。

[0027] S4、利用S3中的特征构建给予决策树的用户角色分类模型,首先构造输入向量 $U_u = \langle ER_u, HP_u, EI_u \rangle$,然后对每一个特征分别计算信息熵 $Entropy(U) = \sum_{j=1}^3 -U^j \log_2 U^j$, U^j 为第j个特征,选取具有最大信息增益的特征构建当前决策节点,对剩余特征逐层递归得到最终决策树模型,进而完成基于情感的用户分类。

[0028] 根据S3得到的结果进行特征融合,得到综合描述用户情感倾向 ER_u 、 HP_u 与情感影响 EI_u 的特征作为模型输入:

$$ER_u = \frac{\sum_{w \in W_u} ER_u(w)}{HR_u + NR_u}$$

$$HP_u = \frac{HP_u(P) - HP_u(N)}{\sum_{e \in M} HP_u(e)}$$

$$EI_u = \frac{\sum_{w \in W_u} \alpha SF_u(w) + \beta TF_u(w) + \gamma EF_u(w)}{HR_u} \left(1 + \frac{HR_u}{HR_u + NR_u}\right)$$

其中 EI_u 对三类影响特征进行融合,并考虑叶子节点并未产生任何影响这一情况,引入 $\left(1 + \frac{HR_u}{HR_u + NR_u}\right)$ 作为去噪因子 HR_u 表示用户转发作为内部节点的个数, NR_u 表示用户转发作为叶子节点的个数,图7展示了当前数据集 ER_u 的计算结果分布,图8展示了 HP_u 的计算结果分布,图9展示了 EI_u 的计算结果分布。最终通过基于决策树的分类方法,得到6种情感角色分类,分类结果如表1所示。

[0029] 表1实施例情感角色分类结果

情感角色	准确度
积极领导者(PL)	0.87
积极追随者(PF)	0.90
中立领导者(OL)	0.83
中立追随者(OF)	0.86
消极领导者(NL)	0.91
消极追随者(NF)	0.92

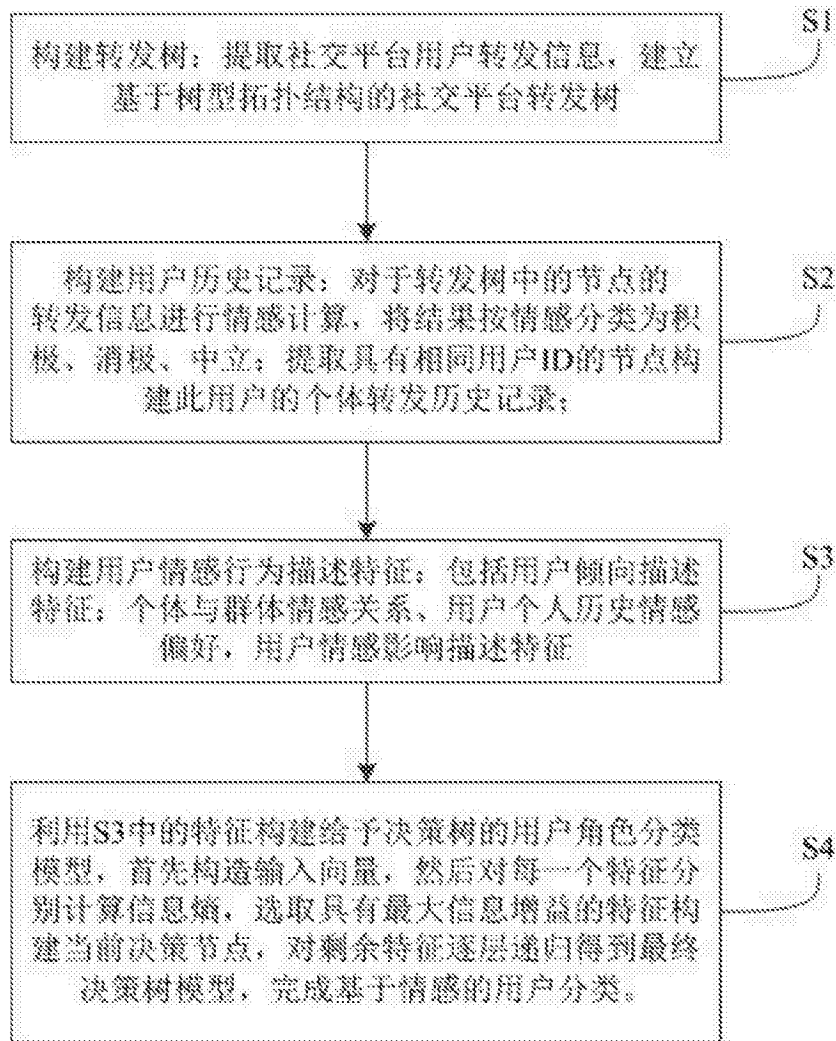


图1

```

<uid_1796678344>
<retweet>
<orig_id>3516700514423273</orig_id>
<orig_text>hanging out with my best friends!LOL</orig_text>
<orig_time>Mon Jan 26 19:43:47 +0800 2014</orig_time>
<orig_emotion>P</orig_emotion>
<orig_name>Nik</orig_name>
<orig_id>22132431243</orig_id>
<orig_text>looks nice!hope you enjoy your trip</orig_text>
<orig_time>Thu Jan 29 02:52:14 +0800 2014</orig_time>
<orig_emotion>P</orig_emotion>
</retweet>
<retweet>...</retweet>
<retweet>...</retweet>
</uid_1796678344>
  
```

图2

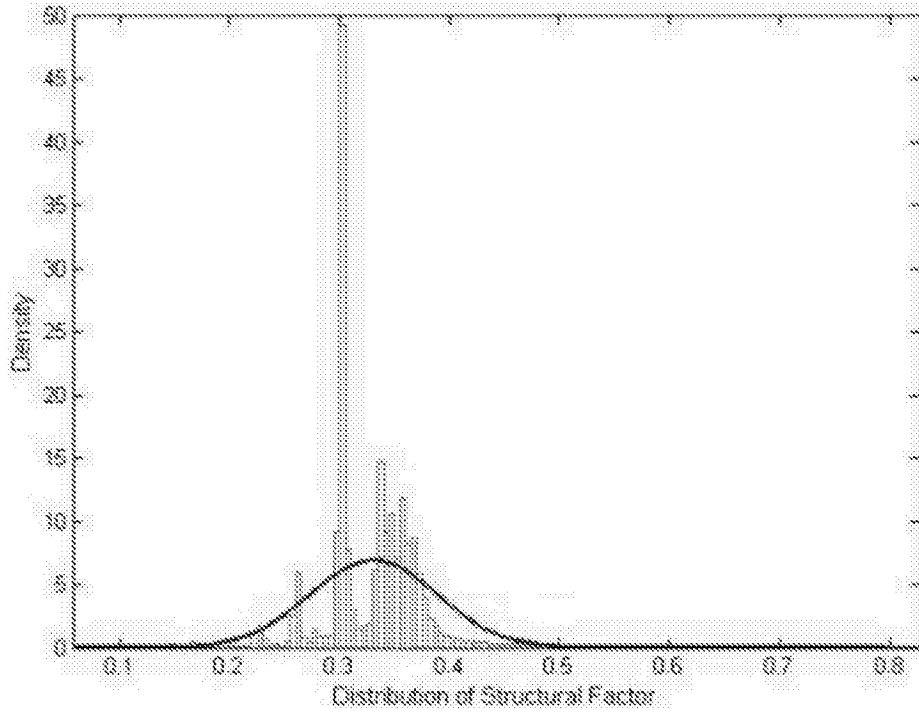


图3

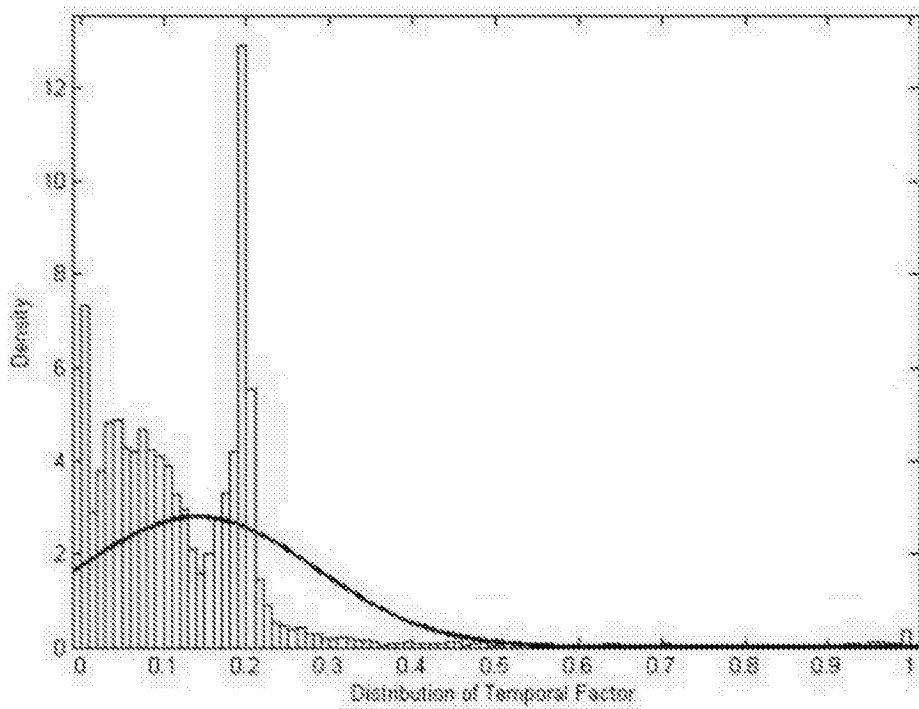


图4

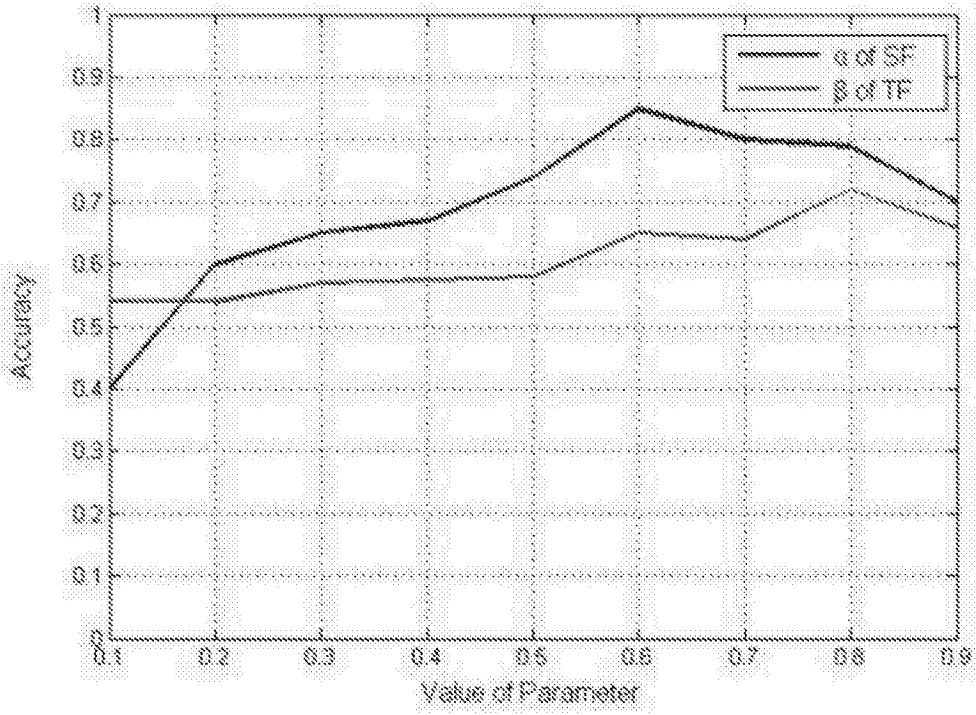


图5

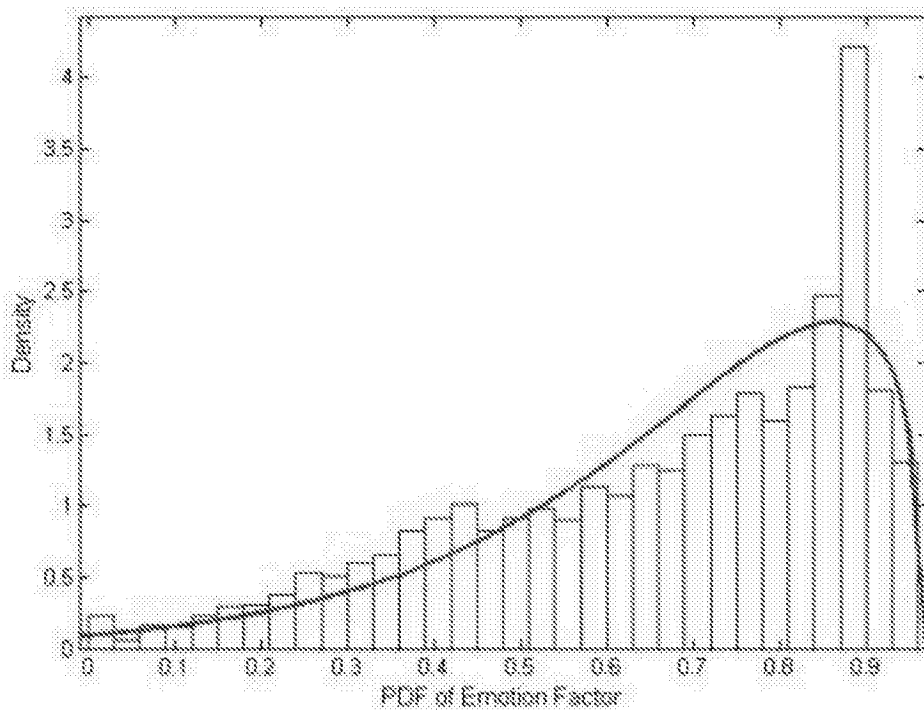


图6

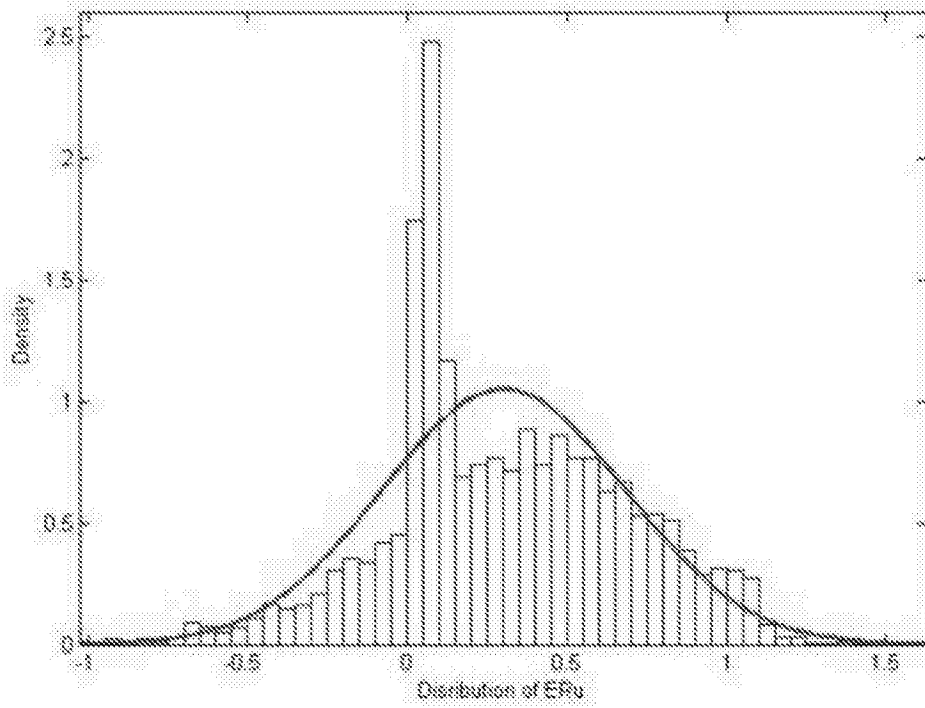


图7

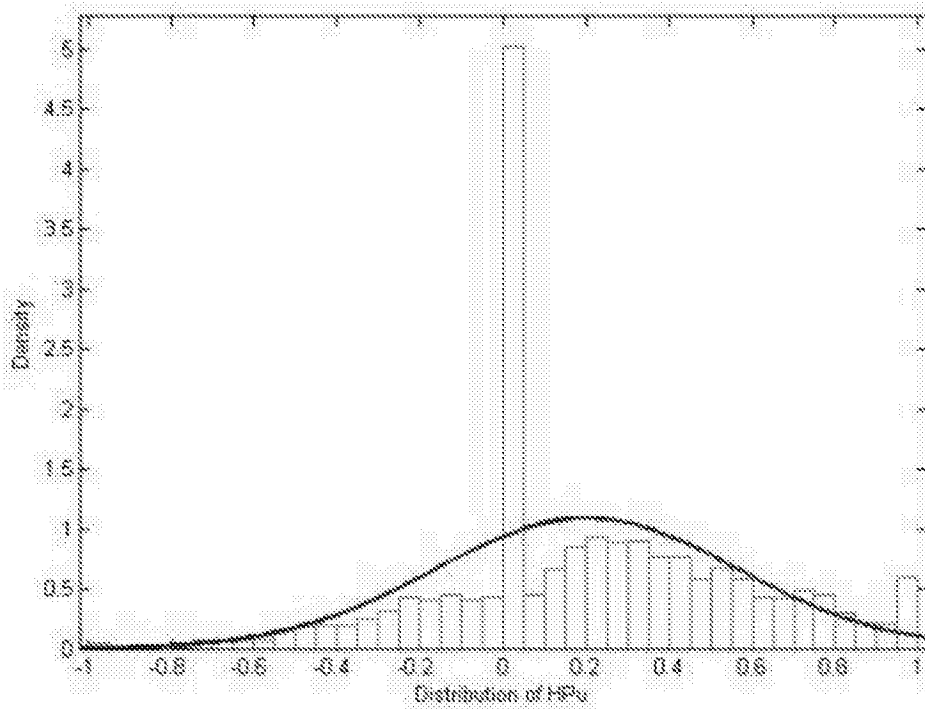


图8

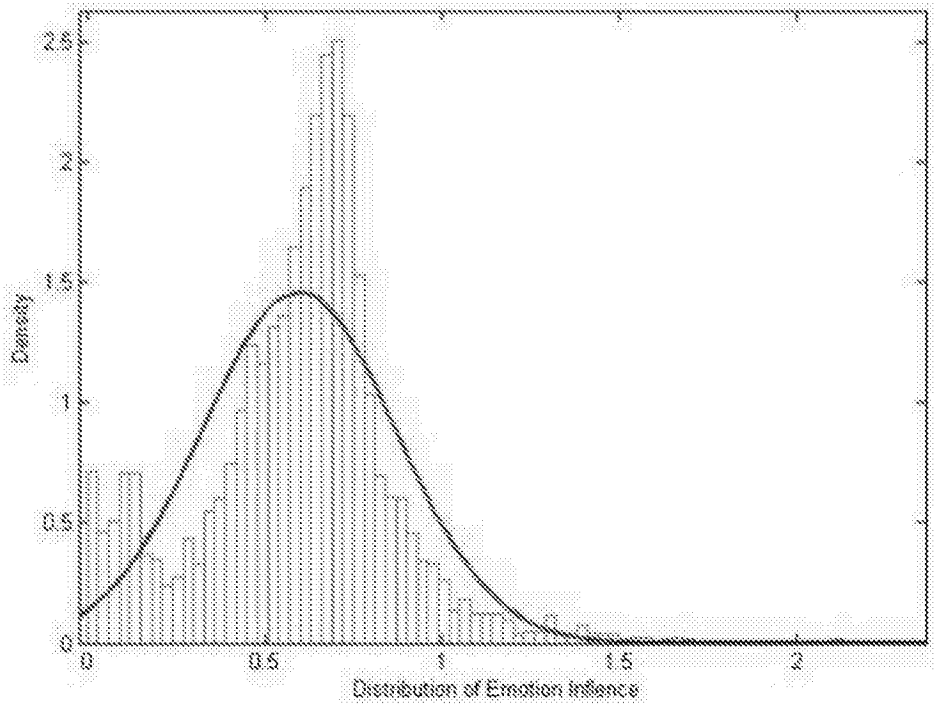


图9