(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 104899267 B
(45)授权公告日 2017.12.19

(21)申请号 201510268991.6

(22)申请日 2015.05.22

(65)同一申请的已公布的文献号
申请公布号 CN 104899267 A

(43)申请公布日 2015.09.09

(73)专利权人 中国电子科技集团公司第二十八研究所
地址 210007 江苏省南京市苜蓿园东街1号1406信箱07分箱

(72)发明人 徐琳 王犇 葛唯益 刘畅 徐欣

(74)专利代理机构 江苏圣典律师事务所 32237
代理人 胡建华

(51)Int.Cl.
G06F 17/30(2006.01)
G06Q 50/00(2012.01)

(56)对比文件
CN 101983383 A,2011.03.02,
CN 102200987 A,2011.09.28,
CN 104239338 A,2014.12.24,
US 2009198723 A1,2009.08.06,
US 2007038659 A1,2007.02.15,

审查员 林欣

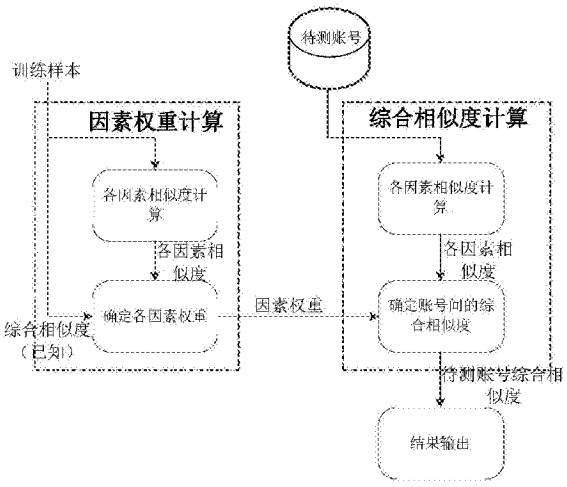
权利要求书3页 说明书12页 附图1页

(54)发明名称

一种社交网站账号相似度的综合数据挖掘方法

(57)摘要

本发明公开了一种社交网站账号相似度的综合数据挖掘方法，该方法能够用于网络舆情监控中，解决识别同一用户的多个社交网站账号的问题。本发明综合考虑了影响社交网站账号综合相似度的三大类因素：个人属性、交互行为和内容，并利用训练样本确定各因素相似度在综合相似度计算中所占的权重。与现有技术相比，本发明的技术优势在于：(1)为识别同一用户的多个社交网站账号提供量化、可靠、全面的参考，并且适用于大数据环境下的计算机自动处理；(2)采用训练样本确定各因素相似度在综合相似度计算中所占的权重，能够保持与人工处理结果的一致性。



CN 104899267 B

1.一种社交网站账号相似度的综合数据挖掘方法，其特征在于，包括社交网站账号综合相似度的计算的方法，步骤如下：

步骤1，计算两个不同社交网站账号的个人属性因素的相似度；

步骤2，计算两个不同社交网站账号交互行为的相似度；

步骤3，计算两个不同社交网站账号的内容的相似度；

步骤4，通过对步骤1～3中计算得到的相似度的加权平均，计算两个社交网站账号的综合相似度；

步骤1中，按照个人属性因素的数据类型，将个人属性因素分为字符串型、文本型、布尔型、枚举型、地址型、时间型和整型，对于各种个人属性因素，根据其类型采用相应的相似度计算方法；对于两个不同的用户社交网站账号a和b，$sim_{f_i}(a,b)$表示a和b在因素f_i上的相似度，相似度是一个大小在[0,1]间的数，其中i＝1,2,…,N,N为影响社交网站账号相似度的因素数量，$s_{f_i,a}$和$s_{f_i,b}$分别表示社交网站账号a和b在影响因素f_i上的取值，各种数据类型个人属性因素的相似度计算方法如下：

(a)字符串型：

当$s_{f_i,a}$和$s_{f_i,b}$为字符串时，相似度$sim_{f_i}(a,b)$计算方法为：

$$sim_{f_i}(a,b)=1-\frac{edis(s_{f_i,a},s_{f_i,b})}{\max(strlen(s_{f_i,a}),strlen(s_{f_i,b}))},$$

其中，函数edis(A,B)表示求字符串A和B的编辑距离，是指字符串A转换成字符串B所需的最少操作次数，函数strlen(A)表示求字符串A的长度；$\max(strlen(s_{f_i,a}),strlen(s_{f_i,b}))$表示求得$strlen(s_{f_i,a})$和$strlen(s_{f_i,b})$之间的最大值；

(b)文本型：

当$s_{f_i,a}$和$s_{f_i,b}$为文本时，相似度计算方法如下：

(b-1)提取文本$s_{f_i,a}$和$s_{f_i,b}$中出现的词语，构成一个词语集合；

(b-2)分别统计文本$s_{f_i,a}$和$s_{f_i,b}$中各个词语出现的词频，按顺序排列构成词频向量$\overrightarrow{L_a}$和$\overrightarrow{L_b}$；

(b-3)求向量$\overrightarrow{L_a}$和$\overrightarrow{L_b}$的余弦值，计算得到相似度：

$$sim_{f_i}(a,b)=\frac{\overrightarrow{L_a}\bullet\overrightarrow{L_b}}{\left|\overrightarrow{L_a}\right|\bullet\left|\overrightarrow{L_b}\right|}$$

其中，符号||为向量取模运算；

(c)布尔型、枚举型或者地址型：

当$s_{f_i,a}$和$s_{f_i,b}$为布尔型、枚举型或者地址型时，相似度计算方法为：

$$sim_{f_i}(a,b)=\begin{cases}1 & s_{f_i,a}=s_{f_i,b}\\ 0 & s_{f_i,a}\neq s_{f_i,b}\end{cases};$$

(d)时间型：

当$s_{f_i,a}$和$s_{f_i,b}$为时间型时，相似度计算方法为：

$$sim_{f_i}(a,b)=1-\frac{\min(thr,\left|s_{f_i,a}-s_{f_i,b}\right|)}{thr},$$

其中，thr是相似度门限，即当社交网站账号a和b因素f_i相差超过时间thr时，即认为社交网站账号的因素f_i没有关联，thr取值范围是0～+∞，$\min(thr,\left|s_{f_i,a}-s_{f_i,b}\right|)$表示求得thr和$\left|s_{f_i,a}-s_{f_i,b}\right|$之间的最小值；

(e)整型：

当$s_{f_i,a}$和$s_{f_i,b}$为整型时，相似度计算方法为：

$sim_{f_i}(a,b)=1-\frac{\left|s_{f_i,a}-s_{f_i,b}\right|}{\max(s_{f_i,a},s_{f_i,b},1)}$，$\max(s_{f_i,a},s_{f_i,b},1)$表示求得$s_{f_i,a}$、$s_{f_i,b}$和1中的最大值；

步骤2中，将社交网站账号之间的交互行为都视为一条有向边，每种交互行为在社交网站账号之间构成一张有向图，每种交互行为有两类影响社交网站账号相似度的方式：正向认同和反向认同、连通性和距离，每种交互行为对社交网站账号综合相似度的每类影响方式，均作为影响社交网站账号综合相似度的因素，两个不同社交网站账号交互行为的相似度的计算方法为：

(1)正向认同和反向认同：

正向认同和反向认同的相似度计算方法为：

$$sim_{f_i}(a,b)=\frac{num\left(F(a)\cap F(b)\right)}{num\left(F(a)\cup F(b)\right)},$$

其中，对于正向认同关系，F(a)表示从社交网站账号a出发的所有有向边指向的社交网站账号的集合，F(b)表示从社交网站账号b出发的所有有向边指向的社交网站账号的集合；对于反向认同关系，F(a)表示到达社交网站账号a的所有有向边另一端的社交网站账号集合，F(b)表示到达社交网站账号b的所有有向边另一端的社交网站账号集合，num()表示统计括号内集合的元素的数量；

(2)连通性和距离：

连通性和距离的相似度计算方法为：

$$sim_{f_i}(a,b)=\begin{cases}\frac{1}{d} & a和b相互可达\\ 0 & a和b相互不可达\end{cases},$$

其中，a和b相互可达是指：若将一种账号之间发生的交互行为作为一条有向边，则账号a能够通过一条以上有向边到达账号b；账号b也能够通过一条以上有向边到达账号a，可达账号之间的距离d是指账号a和b之间间隔的最小有向边数量；

步骤3中采用如下方法计算两个社交网站账号间的内容相似度：

对于社交网站账号a内容的集合Φ_a和社交网站账号b内容的集合Φ_b，

$\Phi_a=\left\{s_{a,1},s_{a,2},s_{a,3},\cdots,s_{a,M_a}\right\}$，

其中，$s_{a,j}$是文本型数据，表示文本的内容，$1\leq j\leq M_a$，M_a是社交网站账号a的内容数量，

$\Phi_b=\left\{s_{b,1},s_{b,2},s_{b,3},\cdots,s_{b,M_b}\right\}$，

其中，$s_{b,k}$是文本型数据，表示文本的内容，$1 \leqslant k \leqslant M_b$，$M_b$是社交网站账号b的内容数量，

两个不同社交网站账号的内容相似度的计算步骤为：

(3-1)采用计算文本型个人属性因素相似度的计算方法，两两计算集合Φ_a中每个元素$s_{a,j}$与集合Φ_b中每个元素$s_{b,k}$的相似度，构成一个集合记为$\{s_{j,k}\}$，

(3-2)令j＝1；m＝0，j为计数器，m是a和b两个账号相同的内容的数量，初始值为0；

(3-3)若$\max(s_{j,k} \mid 1 \leqslant k \leqslant M_b) \geqslant tr$，则将m更新为m+1，其中，tr为用户配置门限，取值在(0,1)间，即两个文本型的内容相似度超过tr时，则判定文本是相同的；

(3-4)将j更新为j+1；

(3-5)若$j \leqslant M_a$，返回(3-3)，否则进入(3-6)；

(3-6)计算社交网站账号a和b内容相似度，计算表达式为：

$sim_{f_i}(a,b)=\dfrac{m}{\max(M_a,M_b)}$，$M_a$和$M_b$分别是账号a和b拥有的内容数量，是定值，$\max(M_a,M_b)$表示求得$M_a$和$M_b$之间的最大值；

步骤4中采用如下方法计算两个社交网站账号间的综合相似度：

$$sim(a,b)=\sum_{i=1}^{N} w_i sim_{f_i}(a,b),$$

其中，w_i是各个影响因素的权重，必须满足$\sum_{i=1}^{N} w_i = 1$；

所述各个影响因素的权重w_i的计算的方法如下：

输入P个训练样本，第1个训练样本综合相似度记为Y_1，因素f_i的相似度记为$X_{i,1}$，其中1＝1,2,…,P，i＝1,2…,N，根据训练样本计算因素权重的步骤如下：

步骤4-1：输入P个训练样本；

步骤4-2：利用P个训练样本的综合相似度Y_1构造矩阵$\overline{Y}$，$\overline{Y}=(1,Y_1,Y_2\cdots,Y_P)$；

步骤4-3：计算所有P个训练样本各个因素的相似度$X_{i,1}$；

步骤4-4：利用$X_{i,1}$构造因素相似度矩阵$\overline{X}$：

$$\overline{X}=\begin{pmatrix} 1 & X_{11} & X_{1,2} & \cdots X_{1,P} \\ 1 & X_{21} & X_{2,2} & \cdots X_{2,P} \\ 1 & \vdots & & \ddots \ \vdots \\ 1 & X_{N,1} & X_{N,2} & \cdots X_{N,P} \end{pmatrix};$$

步骤4-5：利用线性回归公式$\overline{W}=\left(X^T X\right)^{-1} X^T Y$，得到权重矩阵$\overline{W}$，$\overline{W}=(w_1,w_2,\cdots w_N)$，矩阵$\overline{W}$中的各个元素即对应各个影响因素的权重$w_i$的值。

一种社交网站账号相似度的综合数据挖掘方法

技术领域

[0001] 本发明属于计算机互联网数据挖掘技术，用于计算机互联网数据传播控制，特别是一种社交网站账号相似度的综合数据挖掘方法。

背景技术

[0002] 以微博为代表的社交网站的兴起，极大增加了互联网信息传播的速度和广度。社交网站用户之间通过“互粉”、转发、评论、“@”等操作使得信息能够在极短的时间内在社交网上大规模扩散。这种短时间、大范围的信息传播既给用户获取信息带来了极大的便利，但是也带来了网络谣言泛滥的严重问题。

[0003] 为了应对网络谣言泛滥的问题，舆情监控是社交网站管理不可缺少的环节。舆情监控包括两方面内容：一是内容的识别，二是传播的控制。为了实现传播控制，管理者可以通过采取禁言、封停社交网站账号等方式来避免谣言的扩散，但是谣言传播者也可以通过注册多个社交网站账号来逃避管理者的监管。同时，谣言传播者也可以通过注册多个社交网站账号的方式进一步加快谣言传播的速度，扩大谣言传播的范围。因此，识别同一用户注册的多个社交网站账号是社交网站舆情监控中必须解决的关键技术问题。

[0004] 识别同一用户注册的多个社交网站账号实际上是社交网站账号间的相似度分析。现有的社交网站数据挖掘方法无法直接运用于社交网站账号相似度分析，主要有两方面原因：1)由于现有数据挖掘方法主要用于用户关系分析，社交关系的紧密程度并不等同于用户社交网站账号间的相似程度；2)社交网站账号间相似度是一个受多种因素影响的综合性指标，包括：个人属性、发贴内容、转发模式等，目前的数据挖掘方法缺乏对影响关联性多种因素的综合考虑，因此不适用于社交网站账号间关联性分析。

发明内容

[0005] 发明目的：本发明所要解决的技术问题是针对现有技术的不足，提供一种社交网站账号相似度的综合数据挖掘方法，包括不同社交网站账号综合相似度计算方法。

[0006] 不同社交网站账号综合相似度计算方法实施步骤如下：

[0007] 步骤1：计算两个不同社交网站账号的个人属性因素的相似度。个人属性因素包括：用户名、性别、地区、最后发表时间、粉丝数、关注数、文本数、简介、联系方式等。按照个人属性因素的数据类型，将个人属性因素分为字符串型、文本型、布尔型、枚举型、地址型、时间型和整型。对于不同的个人属性因素，需要根据其类型采用相应的相似度计算方法。

[0008] 步骤2：计算两个不同社交网站账号交互行为的相似度。社交网站账号之间的交互行为包括：关注、转发、评论、“@”等。将社交网站账号之间的交互行为都视为一条有向边，则每种交互行为在社交网站账号之间构成一张有向图。每种交互行为(关注、转发、评论、“@”等)有2种影响社交网站账号相似度的方式：正向认同和反向认同、连通性和距离，需要分别计算每种交互行为的上述两项因素的相似度。

[0009] 步骤3：计算两个不同社交网站账号的内容的相似度。若两个社交网站账号经常发

出内容相同的文本、博客等，则两个社交网站账号的相似度就越高。

[0010] 步骤4：通过对各因素相似度的加权平均，计算两个社交网站账号的综合相似度。

[0011] 进一步地，步骤1中，按照个人属性因素的数据类型，将个人属性因素分为字符串型、文本型、布尔型、枚举型、地址型、时间型和整型，对于各种个人属性因素，根据其类型采用相应的相似度计算方法；对于两个不同的用户社交网站账号a和b，$sim_{f_i}(a,b)$表示a和b在因素f_i上的相似度，相似度是一个大小在[0,1]间的数，其中i＝1,2,…,N,N为影响社交网站账号相似度的因素数量，$s_{f_i,a}$和$s_{f_i,b}$分别表示社交网站账号a和b在影响因素f_i上的取值，各种数据类型个人属性因素的相似度计算方法如下：

[0012] (a)字符串型：

[0013] 当$s_{f_i,a}$和$s_{f_i,b}$为字符串时，相似度$sim_{f_i}(a,b)$计算方法为：

[0014] $$sim_{f_i}(a,b)=1-\frac{edis(s_{f_i,a},s_{f_i,b})}{\max(strlen(s_{f_i,a}),strlen(s_{f_i,b}))},$$

[0015] 其中，函数edis(A,B)表示求字符串A和B的编辑距离，是指字符串A转换成字符串B所需的最少操作次数，函数strlen(A)表示求字符串A的长度；$\max(strlen(s_{f_i,a}),strlen(s_{f_i,b}))$表示求得$strlen(s_{f_i,a})$和$strlen(s_{f_i,b})$之间的最大值；

[0016] (b)文本型：

[0017] 当$s_{f_i,a}$和$s_{f_i,b}$为文本时，相似度计算方法如下：

[0018] (b-1)提取文本$s_{f_i,a}$和$s_{f_i,b}$中出现的词语，构成一个词语集合；

[0019] (b-2)分别统计文本$s_{f_i,a}$和$s_{f_i,b}$中各个词语出现的词频，按顺序排列构成词频向量$\overrightarrow{L_a}$和$\overrightarrow{L_b}$；

[0020] (b-3)求向量$\overrightarrow{L_a}$和$\overrightarrow{L_b}$的余弦值，计算得到相似度：

[0021] $$sim_{f_i}(a,b)=\frac{\overrightarrow{L_a}\bullet\overrightarrow{L_b}}{\left|\overrightarrow{L_a}\right|\bullet\left|\overrightarrow{L_b}\right|}$$

[0022] 其中，符号| |为向量取模运算；

[0023] (c)布尔型、枚举型或者地址型：

[0024] 当$s_{f_i,a}$和$s_{f_i,b}$为布尔型、枚举型或者地址型时，相似度计算方法为：

[0025] $$sim_{f_i}(a,b)=\begin{cases}1 & s_{f_i,a}=s_{f_i,b}\\ 0 & s_{f_i,a}\neq s_{f_i,b}\end{cases};$$

[0026] (d)时间型：

[0027] 当$s_{f_i,a}$和$s_{f_i,b}$为时间型时，相似度计算方法为：

[0028] $$sim_{f_i}(a,b)=1-\frac{\min(thr,\left|s_{f_i,a}-s_{f_i,b}\right|)}{thr}$$

[0029] 其中，thr是相似度门限，即当社交网站账号a和b因素f_i相差超过时间thr时，即认

为社交网站账号的因素f_i没有关联，thr取值范围是0～+∞，，例如：取值为24小时，即时间差超过24小时则认为没有相似性；$\min(thr,|s_{f_i,a}-s_{f_i,b}|)$表示求得thr和$|s_{f_i,a}-s_{f_i,b}|$之间的最小值；

[0030] (e)整型：

[0031] 当$s_{f_i,a}$和$s_{f_i,b}$为整型时，相似度计算方法为：

[0032] $sim_{f_i}(a,b)=1-\frac{|s_{f_i,a}-s_{f_i,b}|}{\max(s_{f_i,a},s_{f_i,b},1)}$，$\max(s_{f_i,a},s_{f_i,b},1)$表示求得$s_{f_i,a}$、$s_{f_i,b}$和1中的最大值。

[0033] 步骤2中，将社交网站账号之间的交互行为都视为一条有向边，每种交互行为在社交网站账号之间构成一张有向图，每种交互行为有两类影响社交网站账号相似度的方式：正向认同和反向认同、连通性和距离，每种交互行为对社交网站账号综合相似度的每类影响方式，均作为影响社交网站账号综合相似度的因素，两个不同社交网站账号交互行为的相似度的计算方法为：

[0034] (1)正向认同和反向认同：

[0035] 正向认同和反向认同的相似度计算方法为：

[0036] $$sim_{f_i}(a,b)=\frac{num\left(F(a)\cap F(b)\right)}{num\left(F(a)\cup F(b)\right)},$$

[0037] 其中，对于正向认同关系，F(a)表示从社交网站账号a出发的所有有向边指向的社交网站账号的集合，F(b)表示从社交网站账号b出发的所有有向边指向的社交网站账号的集合；对于反向认同关系，F(a)表示到达社交网站账号a的所有有向边另一端的社交网站账号集合，F(b)表示到达社交网站账号b的所有有向边另一端的社交网站账号集合，num( )表示统计括号内集合的元素的数量；

[0038] (2)连通性和距离：

[0039] 连通性和距离的相似度计算方法为：

[0040] $$sim_{f_i}(a,b)=\begin{cases}\frac{1}{d} & a和b相互可达\\ 0 & a和b相互不可达\end{cases},$$

[0041] 其中，a和b相互可达是指：若将一种账号之间发生的交互行为作为一条有向边，则账号a能够通过一条以上有向边到达账号b；账号b也能够通过一条以上有向边到达账号a，可达账号之间的距离d是指账号a和b之间间隔的最小有向边数量。

[0042] 步骤3中采用如下方法计算两个社交网站账号间的内容相似度：

[0043] 对于社交网站账号a内容的集合Φ_a和社交网站账号b内容的集合Φ_b，

[0044] $$\Phi_a=\left\{s_{a,1},s_{a,2},s_{a,3},\cdots,s_{a,M_a}\right\},$$

[0045] 其中，$s_{a,j}$是文本型数据，表示文本的内容，$1\leq j\leq M_a$，M_a是社交网站账号a的内容数量，

[0046] $\Phi_b=\{s_{b,1},s_{b,2},s_{b,3},\cdots,s_{b,M_b}\}$，

[0047] 其中，$s_{b,k}$是文本型数据，表示文本的内容，$1\leq k\leq M_b$，M_b是社交网站账号b的内容数量，

[0048] 两个不同社交网站账号的内容相似度的计算步骤为：

[0049] (3-1)采用计算文本型个人属性因素相似度的计算方法，两两计算集合Φ_a中每个元素$s_{a,j}$与集合Φ_b中每个元素$s_{b,k}$的相似度，构成一个集合记为$\{s_{j,k}\}$，

[0050] (3-2)令j＝1；m＝0，j为计数器，m是a和b两个账号相同的内容的数量，初始值为0；

[0051] (3-3)若$\max(s_{j,k}|1\leq k\leq M_b)\geq tr$，则将m更新为m+1，其中，tr为用户配置门限，取值在(0,1)间，即两个文本型的内容相似度超过tr时，则判定文本是相同的；

[0052] (3-4)将j更新为j+1；

[0053] (3-5)若$j\leq M_a$，返回(3-3)，否则进入(3-6)；

[0054] (3-6)计算社交网站账号a和b内容相似度，计算表达式为：

[0055] $sim_{f_i}(a,b)=\dfrac{m}{\max(M_a,M_b)}$，$M_a$和$M_b$分别是账号a和b拥有的内容数量，是定值，$\max(M_a,M_b)$表示求得$M_a$和$M_b$之间的最大值。

[0056] 步骤4中采用如下方法计算两个社交网站账号间的综合相似度：

[0057] $sim(a,b)=\sum_{i=1}^{N}w_i sim_{f_i}(a,b)$，

[0058] 其中，w_i是各个影响因素的权重，必须满足$\sum_{i=1}^{N}w_i=1$。

[0059] 本发明提供的一种社交网站账号相似度的综合数据挖掘方法还包括各个影响因素的权重w_i的计算方法：

[0060] 一个训练样本是已知综合相似度的两个社交网站账号。通过输入的训练样本来“训练”系统，得出影响综合相似度各因素的权重，进而实现待测社交网站账号的综合相似度的自动计算。假设共输入P个训练样本，第1个训练样本综合相似度记为Y_1，因素f_i的相似度记为$X_{i,1}$，其中1＝1,2,…,P，i＝1,2…,N。

[0061] 令$\overline{X}=\begin{pmatrix}1 & X_{11} & X_{1,2} & \cdots X_{1,P}\\ 1 & X_{21} & X_{2,2} & \cdots X_{2,P}\\ 1 & \vdots & & \ddots\ \vdots\\ 1 & X_{N,1} & X_{N,2} & \cdots X_{N,P}\end{pmatrix}$，$\overline{Y}=(1,Y_1,Y_2\cdots,Y_P)$，$\overline{W}=(w_1,w_2,\cdots w_N)$

[0062] 根据训练样本计算因素权重的步骤如下：

[0063] 步骤4-1：输入P个训练样本；

[0064] 步骤4-2：利用P个训练样本的综合相似度Y_1构造矩阵$\overline{Y}$；

[0065] 步骤4-3：计算所有P个训练样本各个因素的相似度$X_{i,1}$；

[0066] 步骤4-4：利用$X_{i,1}$构造因素相似度矩阵$\overline{X}$；

[0067] 步骤4-5：利用线性回归公式$\overline{W}=(X^TX)^{-1}X^TY$，得到权重矩阵$\overline{W}$，矩阵$\overline{W}$中的元

素即权重w_i的值。

[0068] 该方法能够用于网络舆情监控中，解决识别同一用户的多个社交网站账号的问题。本发明综合考虑了影响社交网站账号综合相似度的三大类因素：个人属性、交互行为和内容，并利用训练样本确定各因素相似度在综合相似度计算中所占的权重。

[0069] 有益效果：与现有技术相比，本发明的技术优势在于：为识别同一用户的多个社交网站账号提供量化、可靠、全面的参考，并且适用于大数据环境下的计算机自动处理。

附图说明

[0070] 下面结合附图和具体实施方式对本发明做更进一步的具体说明，本发明的上述和/或其他方面的优点将会变得更加清楚。

[0071] 图1为综合相似度计算过程。

[0072] 图2与微博账号“南京正在发生”综合相似度最高10个账号。

具体实施方式

[0073] 社交网站账号相似度受多种因素影响，在计算综合相似度时，必须综合考虑多种影响因素，并确定每种因素的权重。结合图1，本发明首先根据输入的训练样本确定影响综合相似度的每种因素的权重；然后利用得到的因素权重自动完成待测账号的综合相似度计算。本发明可以分为两部分，第一部分是不同社交网站账号综合相似度计算方法，第二部分基于训练样本的因素权重的计算方法。

[0074] 不同社交网站账号综合相似度计算的实施步骤如下：

[0075] 假设a和b分别为两个不同的用户社交网站账号，$sim_{f_i}(a,b)$表示a和b在因素f_i上的相似度，是一个大小在[0,1]间的数。其中i＝1,2,…,N,N为影响社交网站账号相似度的因素数量。不同社交网站账号综合相似度计算方法计算步骤为：

[0076] 步骤1：计算两个不同社交网站账号的个人属性因素的相似度。

[0077] 个人属性因素包括：用户名、性别、地区、最后发表时间、粉丝数、关注数、文本数、简介、联系方式等。其中，联系方式可以包括多个，如QQ、MSN、邮箱、手机号等。按照个人属性因素的数据类型，将个人属性因素分为无空格字符串型、文本型、布尔型、枚举型、地址型、时间型和整型。对于不同的个人属性因素，根据其类型采用相应的相似度计算方法。假设$s_{f_i,a}$和$s_{f_i,b}$分别表示社交网站账号a和b在影响因素f_i上的取值。

[0078] (1)字符串型

[0079] 不同社交网站账号的字符串型个人属性因素之间的编辑距离越小，则该个人属性因素相似度越大。数据类型为字符串型的个人属性因素如：用户名、昵称等。当$s_{f_i,a}$和$s_{f_i,b}$为字符串时，相似度计算方法可以表示为：

[0080] $$sim_{f_i}(a,b)=1-\frac{edis(s_{f_i,a},s_{f_i,b})}{\max(strlen(s_{f_i,a}),strlen(s_{f_i,b}))}$$

[0081] 其中，$s_{f_i,a}$和$s_{f_i,b}$为字符串，函数edis(A,B)表示求字符串A和B的编辑距离，是指字符串A转换成字符串B所需的最少操作次数；函数strlen(A)表示求字符串A的长度。该式子反映编辑距离越大，影响因素的相似度越高。

[0082] (2) 文本型

[0083] 不同社交网站账号的有文本型个人属性因素之间的词向量余弦越小，则该个人属性因素相似度越大。数据类型为文本型的个人属性因素如：个人简介、个性签名等。当$s_{f_i,a}$和$s_{f_i,b}$为文本时，相似度计算方法如下：

[0084] 1) 提取文本$s_{f_i,a}$和$s_{f_i,b}$中出现的词语，构成一个词语集合；

[0085] 2) 分别统计文本$s_{f_i,a}$和$s_{f_i,b}$中各个词语出现的词频，按顺序排列构成词向量$\overrightarrow{L_a}$和$\overrightarrow{L_b}$；

[0086] 3) 求向量$\overrightarrow{L_a}$和$\overrightarrow{L_b}$的余弦值，得到相似度，即：

[0087] $$sim_{f_i}(a,b)=\frac{\overrightarrow{L_a}\bullet\overrightarrow{L_b}}{\left|\overrightarrow{L_a}\right|\bullet\left|\overrightarrow{L_b}\right|}$$

[0088] 其中，符号| |为向量取模运算。

[0089] (3) 布尔型、枚举型、地址型

[0090] 不同社交网站账号的布尔型、枚举型和地址型的个人属性因素只有在完全相同时才能认为其具有关联。布尔型的个人属性因素如：性别；枚举型的个人属性因素如：国家、城市等；地址型的个人属性因素如：QQ、MSN、邮箱、手机号邮编、地址等联系方式。当$s_{f_i,a}$和$s_{f_i,b}$为布尔型、枚举型和地址型时，相似度计算方法可以表示为：

[0091] $$sim_{f_i}(a,b)=\begin{cases}1 & s_{f_i,a}=s_{f_i,b}\\ 0 & s_{f_i,a}\neq s_{f_i,b}\end{cases}$$

[0092] (4) 时间型

[0093] 不同社交网站账号的时间型个人因素属性之间的差值越小，则相似度越高。数据类型为时间型的个人属性因素如：最后发帖时间。当$s_{f_i,a}$和$s_{f_i,b}$为时间型时，相似度计算方法可以表示为：

[0094] $$sim_{f_i}(a,b)=1-\frac{\min(thr,\left|s_{f_i,a}-s_{f_i,b}\right|)}{thr}$$

[0095] 其中，thr是相似度门限，即当社交网站账号a和b因素f_i相差超过时间thr时，即认为社交网站账号的因素f_i没有关联。thr是可配置参数，取值范围是0～+∞，一般可以取24小时。

[0096] (5) 整型

[0097] 不同社交网站账号的整型个人因素属性之间的差值越小，则相似度越高。数据类型为整型的个人属性因素如：粉丝数、关注数、发帖数等。当$s_{f_i,a}$和$s_{f_i,b}$为整型时，相似度计算方法可以表示为：

[0098] $$sim_{f_i}(a,b)=1-\frac{\left|s_{f_i,a}-s_{f_i,b}\right|}{\max(s_{f_i,a},s_{f_i,b},1)}$$

[0099] 步骤2：计算两个不同社交网站账号交互行为的相似度。

[0100] 社交网站账号之间的交互行为包括：关注、转发、评论、“@”等。将每一次交互行为都视为一条有向边，则每种交互行为在社交网站账号之间构成一张有向图。例如：社交网站账号a关注了社交网站账号b，则社交网站账号a到社交网站账号b之间有一条指向社交网站账号b的“有向边”。

[0101] 每种交互行为(关注、转发、评论、“@”等)有2种影响社交网站账号相似度的方式：正向认同和反向认同、连通性和距离，下面分别阐述：

[0102] (1)正向认同和反向认同

[0103] 一个社交网站账号的有向边指向另一个社交网站账号，则这两个社交网站账号之间是正向认同关系。若两个社交网站账号与同一个社交网站账号发生正向认同关系，则这两个社交网站账号具有一定的相似性。例如：社交网站账号a和社交网站账号b同时关注了社交网站账号c，则社交网站账号a和b之间具有一定的相似性。

[0104] 一个社交网站账号被另一个社交网站账号的有向边所指，则这两个社交网站账号之间是反向认同关系。若两个社交网站账号与同一个社交网站账号发生反向认同关系，则这两个社交网站账号具有一定的相似性。例如：社交网站账号c同时关注了社交网站账号a和社交网站账号b，则社交网站账号a和b之间具有一定的相似性。

[0105] 正向认同和反向认同的相似度可以表示为：

[0106] $$sim_{f_1}(a,b)=\frac{num\left(F(a)\cap F(b)\right)}{num\left(F(a)\cup F(b)\right)}$$

[0107] 其中，对于正向认同关系，F(a)表示从社交网站账号a出发的所有有向边指向的社交网站账号的集合，F(b)表示从社交网站账号b出发的所有有向边指向的社交网站账号的集合；对于反向认同关系，F(a)表示到达社交网站账号a的所有有向边另一端的社交网站账号集合，F(b)表示到达社交网站账号b的所有有向边另一端的社交网站账号集合。num( )表示统计括号内集合的元素的数量。

[0108] (2)连通性和距离

[0109] 每种交互行为在社交网站账号之间构成一张有向图，若两个社交网站账号在有向图中通过若干条有向边是互相可达的，则认为两个社交网站账号是强相关的。例如：社交网站账号a关注了社交网站账号b，社交网站账号b关注了社交网站账号c，社交网站账号c关注了社交网站账号a，则存在两条路径a→b→c和c→a使得社交网站账号a和社交网站账号c是相互可达的。

[0110] 相互可达的社交网站账号之间的相似度受社交网站账号之间的距离的影响。社交网站账号之间的距离是两个社交网站账号之间的最短路径上的有向边数量，距离越长，相似度越低。

[0111] 连通性和距离的相似度可以表示为：

[0112] $$sim_{f_1}(a,b)=\begin{cases}\frac{1}{d} & a\text{和}b\text{相互可达}\\ 0 & a\text{和}b\text{相互不可达}\end{cases},$$

[0113] 其中，d是相互可达社交网站账号之间的距离，指账号a和b之间间隔的最小有向边数量。

[0114] 步骤3：计算两个不同社交网站账号的内容的相似度。

[0115] 若两个社交网站账号经常发出内容相同的文本、博客等，则两个社交网站账号的相似度就越高。

[0116] 假设Φ_a和Φ_b分别是社交网站账号a和社交网站账号b内容的集合

[0117] $$\Phi_a=\left\{s_{a,1},s_{a,2},s_{a,3},\cdots,s_{a,M_a}\right\},$$

[0118] 其中，$s_{a,j}$是文本型数据，表示文本的内容，如：社交网站账号a的1篇博客或1篇微博等；M_a是社交网站账号a的内容数量，$1\leqslant j\leqslant M_a$。

[0119] 类似的，$$\Phi_b=\left\{s_{b,1},s_{b,2},s_{b,3},\cdots,s_{b,M_b}\right\},$$

[0120] 其中，$s_{b,k}$是文本型数据，表示文本的内容，如：社交网站账号b的1篇博客或1篇微博等；M_b是社交网站账号b的内容数量，$1\leqslant k\leqslant M_b$。

[0121] 两个不同社交网站账号的内容相似度的计算步骤为：

[0122] 步骤3-1，两两计算Φ_a中每个元素$s_{a,j}$与Φ_b中每个元素$s_{b,k}$的相似度，构成一个集合记为$\{s_{j,k}\}$，计算方法与步骤1中文本型个人属性因素相似度的计算方法相同；

[0123] 步骤3-2，令$j=1$；$m=0$；

[0124] 步骤3-3，若$\max(s_{j,k}\mid 1\leqslant k\leqslant M_b)\geqslant tr$，则$m=m+1$；其中，tr为用户配置门限，取值在(0,1)间，即两个文本型的内容相似度超过tr时，则认为文本是相同的；

[0125] 步骤3-4，将j更新为j+1；

[0126] 步骤3-5，若$j\leqslant M_a$，返回步骤3-3，否则进入步骤3-6；

[0127] 步骤3-6，计算社交网站账号a和b内容相似度，计算表达式为：

[0128] $$sim_{f_i}(a,b)=\frac{m}{\max(M_a,M_b)},$$

[0129] 步骤4：计算两个社交网站账号的综合相似度：

[0130] $$sim(a,b)=\sum_{i=1}^{N}w_i sim_{f_i}(a,b),$$

[0131] 其中，w_i是各个影响因素的权重，必须满足$\sum_{i=1}^{N}w_i=1$。w_i的取值通过基于训练样本的因素权重的计算方法确定。

[0132] 基于训练样本的因素权重的计算方法的实施步骤如下：

[0133] 一个训练样本是已知综合相似度的两个社交网站账号。通过输入的训练样本来“训练”系统，得出影响综合相似度各因素的权重，进而实现待测社交网站账号的综合相似度的自动计算。假设共输入P个训练样本，第l个训练样本综合相似度记为Y_l，因素f_i的相似度记为$X_{i,l}$，其中$l=1,2,\cdots,P$，$i=1,2\cdots,N$。

[0134] 令$$\overline{X}=\begin{pmatrix}1 & X_{11} & X_{1,2} & \cdots X_{1,P}\\ 1 & X_{21} & X_{2,2} & \cdots X_{2,P}\\ 1 & \vdots & & \ddots \vdots\\ 1 & X_{N,1} & X_{N,2} & \cdots X_{N,P}\end{pmatrix},\ \overline{Y}=\left(1,Y_1,Y_2\cdots,Y_P\right),\ \overline{W}=\left(w_1,w_2,\cdots w_N\right),$$

[0135] 根据训练样本计算因素权重的步骤如下：

[0136] 步骤4-1：输入P个训练样本；

[0137] 步骤4-2：利用P个训练样本的综合相似度Y_1构造矩阵$\overline{Y}$；

[0138] 步骤4-3：计算所有P个训练样本各个因素的相似度$X_{i,1}$；

[0139] 步骤4-4：利用$X_{i,1}$构造因素相似度矩阵$\overline{X}$；

[0140] 步骤4-5：利用线性回归公式$\overline{W}=\left(X^TX\right)^{-1}X^TY$，得到权重矩阵$\overline{W}$，矩阵$\overline{W}$中的元素即权重$w_i$的值。

[0141] 实施例一：

[0142] 根据本发明提供的方法构建了新浪微博账号相似度计算系统，系统选取新浪微博账号的27个影响因素，其中个人属性因素14个，交互行为因素12个，内容因素1个，通过对上述影响因素相似度的计算，确定微博账号的综合相似度。采用上述系统对随机选取的超过40万个新浪微博账号进行了综合相似度自动检测。

[0143] 首先，向系统输入500个新浪微博账号训练样本，每个样本包含两个账号的全部信息及该样本两个账号的综合相似度Y_1，其中1＝1，2，…，500。采用如下方法确定影响综合相似度各因素的权重值：

[0144] 步骤1：输入500个训练样本；

[0145] 步骤2：利用500个训练样本的综合相似度Y_1构造矩阵$\overline{Y}$，其中；

[0146] 步骤3：计算所有500个训练样本27个影响因素的相似度$X_{i,1}$，其中i＝1，2，…，27；

[0147] 步骤4：利用$X_{i,1}$构造因素相似度矩阵$\overline{X}$；

[0148] 步骤5：利用线性回归公式$\overline{W}=\left(X^TX\right)^{-1}X^TY$，得到权重矩阵$\overline{W}$，矩阵$\overline{W}$中的元素即权重$w_i$的值。

[0149] 经过计算得到的权重w_1到w_{27}的值为：

[0150] w_1＝0.0197；w_2＝0.0160；w_3＝0.0041；w_4＝0.0400；

[0151] w_5＝0.0079；w_6＝0.0101；w_7＝0.0136；w_8＝0.0118；

[0152] w_9＝0.0140；w_{10}＝0.0259；w_{11}＝0.0181；w_{12}＝0.0119；

[0153] w_{13}＝0.0197；w_{14}＝0.0200；w_{15}＝0.0427；w_{16}＝0.0270；

[0154] w_{17}＝0.0470；w_{18}＝0.0514；w_{19}＝0.0516；w_{20}＝0.0818；

[0155] w_{21}＝0.0609；w_{22}＝0.0479；w_{23}＝0.0666；w_{24}＝0.0614；

[0156] w_{25}＝0.0542；w_{26}＝0.0838；w_{27}＝0.0909；

[0157] 然后，在完成对系统的训练后，系统对待测的约40万个新浪微博账号两两检测综合相似度，按照综合相似度从高到低，列出与每个新浪微博账号最相似的10个账号。例如：图2中显示了系统计算得出的与微博账号“南京正在发生”综合相似度最高10个账号。检测两个账号综合相似度的步骤为：

[0158] 步骤1：两两计算不同新浪微博账号的个人属性因素的相似度。

[0159] 纳入统计的新浪微博账号个人属性因素及其类型如表1，分别计算这些个人属性因素的相似度。

[0160] 表1

[0161]

序号*i*	个人属性因素	类型	说明
1	登录名（邮箱）	字符串型	按照字符串型因素计算方法计算相似度，相似度取值[0,1]
2	昵称	字符串型	按照字符串型因素计算方法计算相似度，相似度取值[0,1]
3	手机号	地址型	按照地址型因素计算方法计算相似度，相似度取值 0 或 1
4	性别	布尔型	按照布尔型因素计算方法计算相似度，相似度取值 0 或 1
5	所在地	地址型	按照地址型因素计算方法计算相似度，相似度取值 0 或 1
6	生日	地址型	按照地址型因素计算方法计算相似度，相似度取值 0 或 1
7	血型	枚举型	按照枚举型因素计算方法计算相似度，相似度取值 0 或 1
8	简介	文本型	按照文本型因素计算方法计算相似度，相似度取值[0,1]
9	QQ	地址型	按照地址型因素计算方法计算相似度，相似度取值 0 或 1
10	MSN	地址型	按照地址型因素计算方法计算相似度，相似度取值 0 或 1
11	粉丝数	整型	按照整型因素计算方法计算相似度，相似度取值[0,1]
12	关注数	整型	按照整型因素计算方法计算相似度，相似度取值[0,1]

[0162]

13	发帖数	整型	按照整型因素计算方法计算相似度，相似度取值[0,1]
14	最后发帖时间	时间型	按照时间型因素计算方法计算相似度，相似度取值[0,1]，相似度门限 *thr=24 小时*

[0163] 步骤2：两两计算不同微博账号交互行为的相似度。

[0164] (1)正向认同和反向认同

[0165] 考虑新浪关注、转发、评论、“@”四种交互行为所构成的正向认同和反向认同关系，分别计算其相似度。计算相似度时，集合F(a)和F(b)的定义如表2：

[0166] 表2：

[0167]

序号 i	交互行为	认同关系	说明
15	关注	正向认同	$F(a)$ 表示微博账号 a 关注的其他微博账号的集合 $F(b)$ 表示微博账号 b 关注的其他微博账号的集合
16	关注	反向认同	$F(a)$ 表示关注微博账号 a 的其他微博账号的集合 $F(b)$ 表示关注微博账号 b 的其他微博账号的集合
17	转发	正向认同	$F(a)$ 表示微博账号 a 转发的其他微博账号的集合 $F(b)$ 表示微博账号 b 转发的其他微博账号的集合
18	转发	反向认同	$F(a)$ 表示转发微博账号 a 的其他微博账号的集合 $F(b)$ 表示转发微博账号 b 的其他微博账号的集合
19	评论	正向认同	$F(a)$ 表示微博账号 a 评论的其他微博账号的集合 $F(b)$ 表示微博账号 b 评论的其他微博账号的集合
20	评论	反向认同	$F(a)$ 表示评论微博账号 a 的其他微博账号的集合 $F(b)$ 表示评论微博账号 b 的其他微博账号的集合
21	@	正向认同	$F(a)$ 表示微博账号 a “@”的其他微博账号的集合 $F(b)$ 表示微博账号 b “@”的其他微博账号的集合
22	@	反向认同	$F(a)$ 表示“@”微博账号 a 的其他微博账号的集合 $F(b)$ 表示“@”微博账号 b 的其他微博账号的集合

[0168] (2)连通性和距离

[0169] 如表3所示，考虑新浪微博关注、转发、评论、“@”四种交互行为构成的有向图的连通性和距离，即关注相互可达、转发相互可达、评论相互可达和“@”相互可达，分别计算其相似度。

[0170] 表3：

[0171]

序号i	交互行为	说明
23	关注	在关注行为构成的有向图中相互可达
24	转发	在转发行为构成的有向图中相互可达
25	评论	在评论行为构成的有向图中相互可达
26	@	在“@”行为构成的有向图中相互可达

[0172] 步骤3：两两计算不同新浪微博账号的内容的相似度

[0173] 每个新浪微博账号发出的每一条微博，作为该账号的一个内容，微博数量即该账号的内容数量。因此，在新浪微博的统计中，$s_{a,j}$表示账号a的第j篇微博，j满足$0 \leq j \leq M_a$，M_a为账号a的微博总数。

[0174]

序号i	说明
27	内容的相似度，在计算中tr取值为0.8，即内容80％相似时，两条微博判为相同。

[0175] 步骤4：两两计算不同新浪微博账号的综合相似度。

[0176] 计算时权重w_i即采用通过训练样本确定的权重值。

[0177] 本发明提供了一种社交网站账号相似度的综合数据挖掘方法，具体实现该技术方案的方法和途径很多，以上所述仅是本发明的优选实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本发明原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也应视为本发明的保护范围。本实施例中未明确的各组成部分均可用现有技术加以实现。

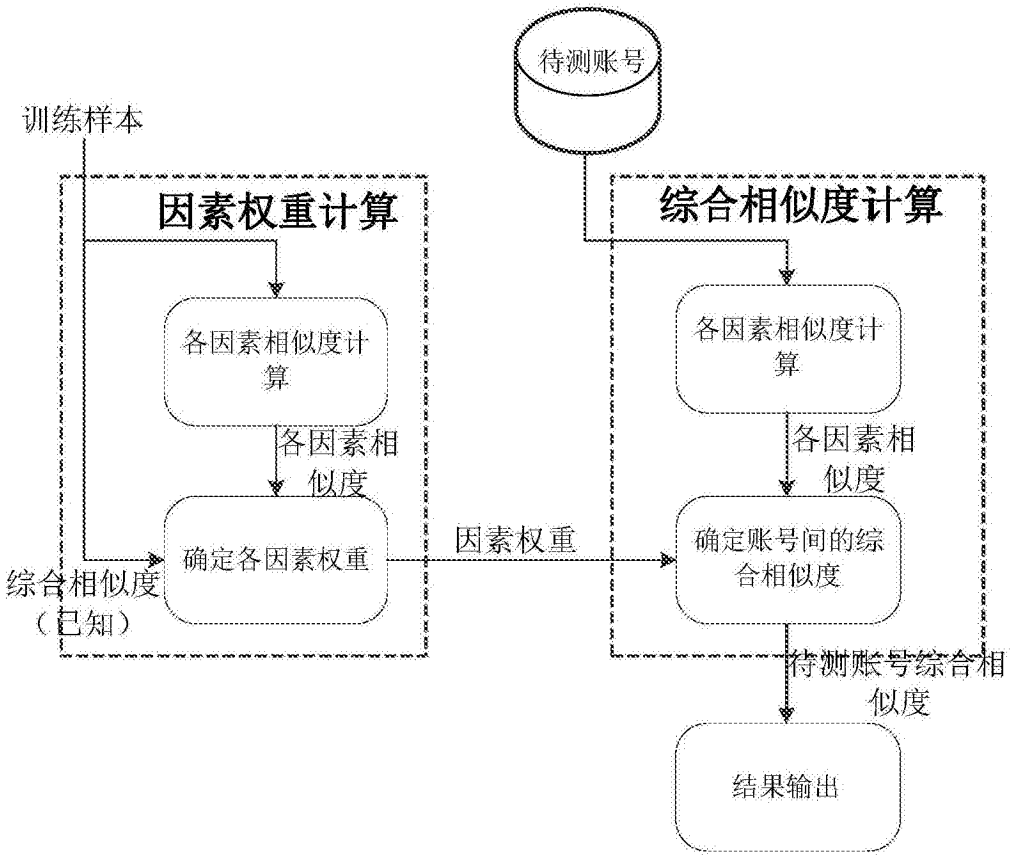

待测账号
训练样本
因素权重计算
综合相似度计算
各因素相似度计算
各因素相似度
确定各因素权重
因素权重
综合相似度（已知）
各因素相似度计算
各因素相似度
确定账号间的综合相似度
待测账号综合相似度
结果输出

图1

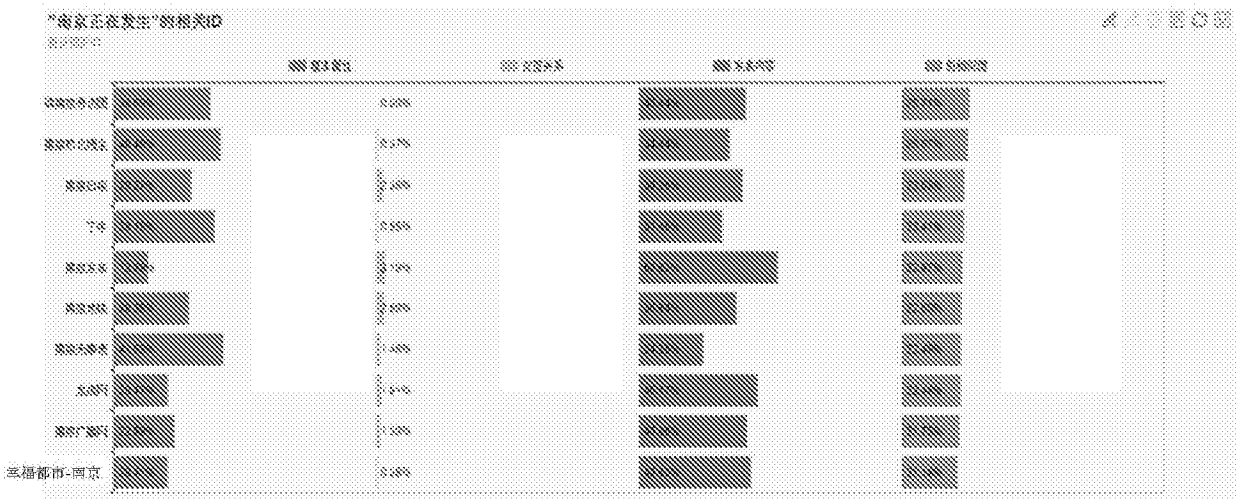


图2