doi :10.3969/j.issn.1671-1122.2016.01.015

微博自动分类系统设计

张士豪，顾益军，张俊豪

（中国人民公安大学网络安全保卫学院，北京 102623）

摘　要：文章提出了一种热门微博分类的新思路，通过对热门微博的转发用户进行聚类分析，并根据不同的用户聚集状态来区分不同种类的热门微博。在用户聚类中采用了基于K-means聚类算法的改进算法X-means，并根据微博用户数据特点对X-means算法进行了进一步改进，将属性差异和用户节点差异考虑在聚类过程当中。其中，在对X-means算法改进过程中，对于用户属性的加权采用了基于对数函数的加权方式，确保聚类结果更加科学、准确；在对用户自身权重的加权中，通过建立重点人员信息库的方式，实现了对特殊用户节点的加权，并利用HITS算法对重点人员信息库实现动态更新。在完成用户聚类之后，将得到的重要用户的信息分领域录入重点人员信息库，实现聚类过程与信息库的反馈机制。另外，实验将相同数据分别代入改进前后的K-means算法与X-means算法中，并通过轮廓系数评价聚类结果，证明了改进后的X-means算法在微博用户聚类中更有优势。

关键词：微博分类；用户聚类；轮廓系数

中图分类号：TP309　　**文献标识码：**A　　**文章编号：**1671-1122（2016）01-0081-07

中文引用格式：张士豪，顾益军，张俊豪．微博自动分类系统设计[J]. 信息网络安全，2016(1)：81-87.

英文引用格式：ZHANG Shihao, GU Yijun, ZHANG Junhao, An Automatic Classification System for Microblogging [J]. Netinfo Security, 2016 (1) : 81-87.

An Automatic Classification System for Microblogging

ZHANG Shihao, GU Yijun, ZHANG Junhao

(*School of Cybersecurity,People's Public Security University of China, Beijing* 102623, *China*)

Abstract: This paper proposed a new idea for popular microblogging classification, by analyzing the users who forwarded the popular microblogging to obtain the clustering result, and distinguishing the different kinds of popular microblogging depending on the aggregation state of user. The user clustering algorithm is called X-means algorithm which improved on the basis of K-means clustering algorithm, and improved further according to the characteristics of the microblogging user. Taking into account the difference of the user themselves and their attributes, this paper used a weighted approach based on the logarithmic function in the process of improving X-means algorithm ,which can ensure that the clustering results more scientific and accurate. Simultaneously , this paper achieved a weighted approach for the special nodes by the way of establishing a Key-Personnel- Database, then this paper achieved the dynamic updates of the database with the HITS algorithm. After completing the user clustering, the experiment put the important user information into the Key-Personnel- Database in different fields, by which can achieve the feedback mechanism between the clustering processes and the database. In addition, clustered the microblogging user with the X-means algorithm and the k-means algorithm as well as their improved algorithm, and ultimately proved the improved X-means algorithm has more advantages in the microblogging user clustering.

Key words: microblogging classification; user clustering; outline coefficient

收稿日期：2015-11-16

基金项目：公安部重点研究计划[2011ZDYJGADX016]

作者简介：张士豪（1992—），男，山西，硕士研究生，主要研究方向为网络安全与数据挖掘；顾益军（1968—），男，江苏，副教授，博士，主要研究方向为网络安全与数据挖掘；张俊豪（1991—），男，河南，硕士研究生，主要研究方向为网络安全与数据挖掘。

通信作者：张士豪 296666546@163.com

0 引言

在新兴的互联网时代，微博作为一种短内容交互式的社交平台已经成为人们发表意见、共享信息的一种主要工具。而随着微博的普及，在微博中也出现了许多影响社会稳定、危害社会治安的舆情事件，其中包括诽谤、谣言、反动言论以及恐吓威胁等内容。公安网监部门必须仔细地对微博进行实时监控，才能及时发现违法行为，尽早处理。然而，新浪微博拥有庞大的用户群，单靠人工搜索或者简单的筛选无法处理这样的海量数据。因此，对获得的微博数据进行自动化处理分类十分必要，不仅可以节省警力，还可以提高网监部门的舆情监控效率[1-3]。

目前有关微博自动分类的研究主要集中在基于微博文本内容以及微博标签的分类，通过这种方式可以区分不同话题的微博种类，但是这种分类方式对于舆情监控不能起到很好的效果[4,5]。因此，本文提出一种基于用户聚类的微博分类方法，更加贴合公安实际工作。聚类算法采用了X-means聚类算法，并且结合微博用户的实际情况，对X-means算法进行了适当的改进。同时对聚类结果进行分析比较。另外，对用户聚类结果中少量的重点用户信息进行数据库存储，并且将不同领域内的用户加上不同领域的标签，将数据库中的信息反馈到用户聚类算法中去，从而使得聚类结果更加科学。

1 相关研究成果

在微博分类方法选择上，许多研究都采用了从文本分类的角度对微博自动化分类进行研究。江斌通过对微博文本分词方法的不断完善，构建了不同种类微博的特征模式库，然后通过对微博分词结果与特征模式库进行比对来实现微博的自动分类，这种技术手段主要应用于用户喜好实现微博推送的情况[2]；周咏梅等人使用了句法分析以及条件随机场（Conditional Random Files,CRFs）方法抽取特定的评价词语对象，然后使用多策略的支持向量机（SVM）算法进行微博分类，通过对词语的筛选，提高了微博文本情感分类的准确性[4]；曹海涛利用基础情感词汇（Pleasure Arousal Dominance, PAD）对每一个微博词汇进行情感倾向性分析，计算出词汇PAD值，进而对微博进行文本情感分类[6]；谢丽星等同样利用了SVM的方法进行微博分类，但是她将层次结构的思想引入了微博情感分析和特征抽取的过程中，提高了分类准确性[7]；杜伟夫提出一种领域相关的词语情感倾向性的计算框架，把语义倾向性计算转化为词语优化问题，利用词语之间的相似度建立词语无向图，并对图进行切分求解[8]；高永兵等采取了基于K-means算法的改进算法，改进了短文本矩阵的向量稀疏性问题，将“微话题”内容进行比较分析并改进节点相似度计算方法，一定程度上解决了K-means聚类算法手动选取K值和随机选取聚类中心的问题[9]。以上这些方法可以很有效地对不同话题的微博进行分类，但用在舆情监控方面，基于用户的微博分类比基于文本的分类方法显得更加有效。

在用户聚类算法选择上，张雪凤等人也对传统聚类算法K-means进行改进，改进主要围绕K-means聚类算法的聚类准则函数进行，以簇间加权标准差之和代替簇内误差平方和作为准则函数，权重则由簇内节点数目占总节点数目的百分比来决定，这种改进思路十分具有借鉴意义[10]；王荣等从用户关键字的角度对用户进行聚类分析，采用了基于Rock聚类算法的改进算法，并利用独创性的相似权重和平均邻居的概念，简化了算法的复杂度，同时还提高了聚类的准确性，但是这种方法并没有考虑到用户节点自身的有效性[11]；李磊等人采用了CLARA（Cluster Larger Application）聚类算法,这种算法基于K-mediods（K-中心点）算法，根据用户的属性值（粉丝数、关注数、微博数）将用户分为三类，在特定话题下对重点用户进行重点分析，这种思路非常符合舆情监控的需要，不过方法中没有将用户和微博联系起来进行分析，用户属性选择也比较少[12]；曹鹏等人基于X-means聚类算法提出一种自适应随机子空间组合分类算法，将整体的数据集利用X-means算法分为多个子空间，分别进行处理，其中构建分类器的思路可以用于用户聚类之后的节点更新问题[13]；赵峥则分别使用K-means算法和Two-Step算法对新浪微博中的用户信息进行分析研究，其中，在K-means算法的使用过程中还对算法进行了改进，将聚类结果与C-H指数结合分析，选择最优的聚类个数，并最终得出K-means聚类算法在微博用户分析中比Two-Step算法更加有效的结论[14]；何黎等人在个性化营销策略的研究中对于核心用户的挖掘思路对本文中用户权重分配有很大的借鉴意义[15]；杨凯等人分析了微博用户从注

册期到稳定期的数据，并对用户关系使用 K-means 算法进行挖掘，进而对用户进行分类[16]。

本文提出的基于 X-means 聚类算法的改进算法与上述的算法相比可以更好地处理微博数据，采用了基于用户属性的聚类方式，并引入属性权重以及用户自身权重对用户节点进行复合加权，可以得到更加科学的聚类结果。

2 算法描述

2.1 用户聚类算法

X-means 算法是基于 K-means 算法的一种改进算法[17]，而 K-means 算法也叫做 K 均值算法，是一种具有代表性的硬聚类算法，算法实现的过程为：首先，由用户根据需求指定最后要得到的聚类个数 K，并提供初始的数据集；其次，算法在数据集中随机选择 K 个聚类中心，然后计算每一个数据集中的节点到这些聚类中心的距离，将每一个节点分配给离它最近的聚类中心，形成 K 个簇，然后重新计算每个簇的聚类中心，并重新计算节点到中心的距离；最后，如果每一个簇的聚类中心不再变化，也没有任何一个节点发生重新分配的情况，那么就可以认为得到了最终的聚类结果。

K-means 聚类算法的距离函数采用了欧氏距离函数，而聚类准则函数采用了误差平方和函数（SSE），其计算过程如公式（1）所示。

$$SSE=\sum_{j=1}^{k}\sum_{x\in C_j}dist(x,m_j)^2 \quad \cdots\cdots(1)$$

其中，k 表示聚类个数，C_j 表示标号为 j 的聚类，$dist(x,m_j)$ 表示第 j 个簇中的节点与聚类中心 m_j 的欧式距离。

当聚类达到最优解的时候，也就是 SSE 的值收敛到一个极小值的时候。误差平方和函数的意义在于，当 SSE 值收敛为一个极小值的时候，此时聚类得到最优的聚类结果。

K-means 聚类算法虽然简单易行，但是在实际应用过程中，用户为了得到最佳聚类结果，往往需要进行多次试验，尝试不同的 K 值，再对结果进行分析比较，得到最终的最优解；另外，它的聚类准则函数太过简单，需要引入更加科学的聚类结果评价机制，因此，基于这两点，X-means 算法做出了相应的改进。

X-means 算法在保留 K-means 算法原理的同时，引入了 BIC 准则作为聚类结果的评价标准，算法实现的过程为：首先，由用户指定最终的聚类个数的取值范围 $[K_1,K_2]$，并提供初始数据集；其次，算法以 K_1 作为聚类个数运行一次 K-means 算法，得到 K_1 个聚类，然后对每一个聚类都使用 K-means 算法聚为两个子类，然后比较所有 K_1 组父类和孩子类的 BIC 得分（贝叶斯系数），找出分差最大的一组父类和子类，将子类保留，剩余组保留原来的父类，从而得到 K_1+1 个聚类，不断重复上述过程；最后，当最终的聚类个数达到用户定义的聚类个数最大值 K_2 时返回最佳的聚类结果。

BIC 得分的运算方法如公式（2）所示：

$$BIC(M_j)=L_j(D)-p(j)/2*\log R \quad \cdots\cdots(2)$$

其中，D 表示已知数据集，$L_j(D)$ 表示算法在进行第 j 次迭代过程之后得到的聚类结果 M_j 的对数似然函数所能取到的极大值，$p(j)$ 是 M_j 中对应的参数数量，R 为已知数据集 D 中的节点个数。BIC 得分标准也被称为 Schwarz 标准，与 BIC 得分标准类似的还有 AIC 标准或者是 MDL 算法等都可以对算法进行优化，只不过针对本文实验中所涉及的数据集来说，最适合的还是 BIC 得分标准。

2.2 属性加权算法

在对用户数据进行聚类之前，需要对得到的用户数据进行预处理。首先，由于用户的各个属性在用户类别判定中的重要程度不同，因此需要对用户的不同属性值进行加权处理。属性权重的值可以利用层次分析法来确定，通过量化每两个属性之间的重要程度的对比情况，得出层次分析法的初始矩阵，然后计算出用户属性权重的值 R。

在计算出用户属性权重之后，不能简单地利用属性权重与用户属性值相乘得到新属性值的方法进行属性加权，因为在之后的数据预处理环节中我们将要对属性值进行归一化处理，因此不论属性值放大几倍，对最终代入聚类算法的属性值都没有任何影响。本文实验所采取的属性加权方法利用了对数函数自变量越趋近于 1 变化率越大的特点，计算公式如公式（3）所示。

$$t_{ij}=X_{ij}\ln[(1-R_i)X_{ij}+1] \quad \cdots\cdots(3)$$

公式（3）中，X_{ij} 表示未处理前的属性值，t_{ij} 表示经过属性加权之后的属性值，R_i 表示第 i 维属性的权值。由于层次分析法计算出的 R 值是永远小于 1 的，而上述公式得出的属性值变化幅度与对数函数中 X_{ij} 的系数是成反比的，

因此通过 $1-R_i$ 改变了函数针对变量 R_i 的单调性，通过上式可以实现微博用户的属性加权。

在对属性值进行加权之后，由于各个属性之间在数值上往往相差很大，所以为了避免不同属性值之间的巨大差异影响到最终的聚类结果，要对用户的各个属性值进行归一化处理，将用户属性值中最大值与最小值统计出来，并将每一个属性值都代入公式（4），得出新的属性值。

$$x_{ij}=\frac{t_{ij}-\mathrm{Min}_i}{\mathrm{Max}_i-\mathrm{Min}_i} \quad \cdots\cdots (4)$$

其中，x_{ij} 表示第 j 个节点的第 i 维属性的新属性值，t_{ij} 表示之前的属性值，Min_i 表示第 i 维属性的最小属性值，Max_i 表示第 i 维属性的最大属性值。将所有属性值都归为 [0,1] 范围内的值，通过这种方式就避免了属性值之间的量级差对结果的影响。

2.3 用户加权算法

将通过聚类得到的少数有影响力的用户信息录入数据库，并将重点用户分领域、分类别地存储在重点人员信息库中，同时在重点人员信息库的不断完善过程中，利用重点人员之间的相互关系，得出信息库中每一名用户对于该领域的影响因子，由影响因子的大小对重点人员的权重进行动态浮动调整。

完整的微博分类系统设计思路如图 1 所示。

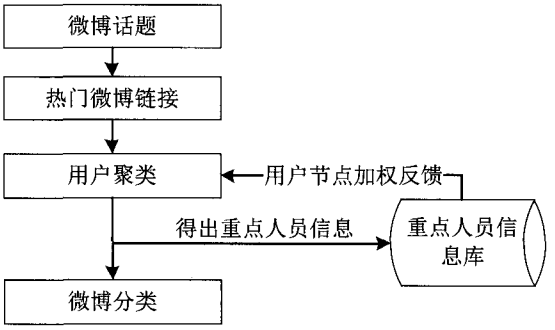

图1 系统设计

其中对用户节点的分析是整个系统的核心内容，分析过程如图 2 所示。

其中 HITS（Hypertext – Induced Topic Search）算法是一种网页评级算法，一般应用于用户查询当中，依照搜索引擎为用户返回的网页列表进行逐个评价，在数据库用户加权过程中，也可以使用 HITS 算法进行节点权重确定：将每一名用户都看做一个单独的网页，然后将用户之间的相互关注关系作为网页的链入链接和链出链接，根据用户

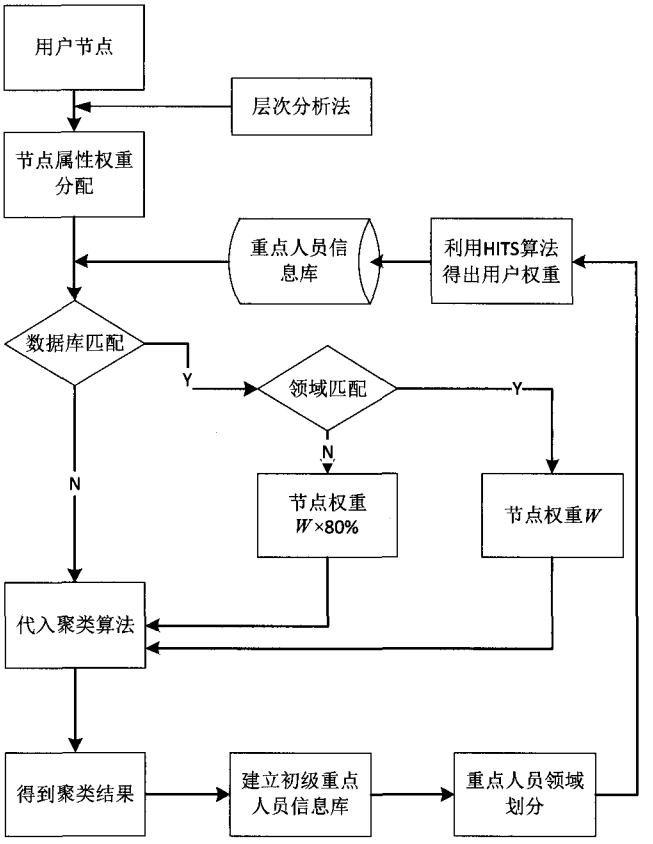

图2 用户节点分析流程

关系得出单一领域内的节点有向图，并转换成矩阵模型代入 HITS 算法，最终得出用户权重。

HITS 算法主要围绕两个关键参数，分别是权威等级和中心等级。一个节点的权威等级越高，说明有越多的用户关注了该节点；而一个节点的中心等级越高，就表示此节点关注的用户也越多。权威性和中心性是一种相互依存的关系，一个好的权威节点必定被很多好的中心节点关注，而一个好的中心节点也必然会关注许多好的权威节点。

用 $\boldsymbol{S}$ 集表示重点人员信息库中的所有节点，用 $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$ 表示 $\boldsymbol{S}$ 集所形成的有向图（如图 3 所示），$\boldsymbol{V}$ 表示用户节点，$\boldsymbol{E}$ 表示有向边的集合（有向边表示用户之间的单向关注关系）。

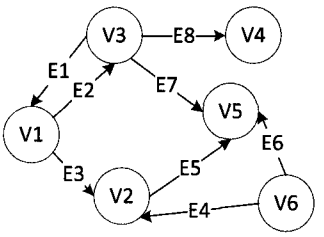

图3 有向图S示意图

用 $\boldsymbol{L}$ 作为 G 的邻接矩阵，有如公式（5）所示的关系：

$$L_{ij}=\begin{cases}1 & 当(i,j)\in E\\ 0 & 其他\end{cases} \quad (5)$$

在有向图中，每一名用户的权威值表示为 $a(i)$，中心值表示为 $h(i)$，这两个参数的值是相互影响的，它们之间的关系如公式（6）和公式（7）所示：

$$a(i)=\sum_{(j,i)\in E}h(j) \quad (6)$$

$$h(i)=\sum_{(j,i)\in E}a(j) \quad (7)$$

将所有节点的权威值和中心值的写成矩阵的形式，有权威值列向量 $\boldsymbol{a}=(a(1),a(2),\cdots,a(n))^{T}$ 和中心值列向量 $\boldsymbol{h}=(h(1),h(2),\cdots,h(n))^{T}$，由上面的权威值与中心值的关系可以推导出公式（8）和公式（9）所示结论：

$$\boldsymbol{a}=\boldsymbol{L}^{T}\boldsymbol{h} \quad (8)$$

$$\boldsymbol{h}=\boldsymbol{L}\boldsymbol{a} \quad (9)$$

然后采用幂迭代的方法计算权威值和中心值，用 $\boldsymbol{a}_k$ 和 $\boldsymbol{h}_k$ 作为第 k 次迭代的权威值和中心值，有公式（10）、公式（11）所示的迭代关系：

$$a_k=\boldsymbol{L}^{T}\boldsymbol{L}a_{k-1} \quad (10)$$

$$h_k=\boldsymbol{L}\boldsymbol{L}^{T}h_{k-1} \quad (11)$$

定义 $\boldsymbol{a}_0=\boldsymbol{h}_0=(1,1,\cdots,1)$ 作为迭代的初始条件，经过若干次迭代，当 a_k 与 a_{k-1} 的差值趋近于某些极小向量 ε 时，迭代终止，算法返回每一个用户节点的权威值和中心值，并将用户权威值与中心值的总和作为用户节点的重点人员权重 W。

通过不定期地利用 HITS 算法进行数据库权重更新，可以完善数据库的建设和维护，热门微博转发用户在被聚类分析时需要与已知的数据库进行比对，如果被分析节点存在于数据库中，那么需要将其属性值进行加权，反之则将 W 值置为 1，改动之后两点的距离定义如公式（12）所示：

$$d=\mathrm{sqrt}\left(\sum_{i=1}^{n}(W_1x_{i1}-W_2x_{i2})^2\right) \quad (12)$$

另外，在建立重点人员信息库之后，将会在用户入库时对节点进行领域划分。在分析用户节点时，如果被分析用户与数据库中节点比对成功，那么就要进行领域比对，如果被分析节点领域与目标微博讨论领域相同，则在用户聚类中赋予其全值权重，反之，则赋予该节点 80% 的重点人员权重，这里需要注意的是削弱后的权重不能小于 1，也就是说如果发生了 $0.8\times W$ 的值小于 1 的情况，就将其看做是普通节点，权重赋为 1。

3 实验

3.1数据获取阶段

本实验选取了新浪微博作为数据源，为了保证实验的准确性和实时性，在第三方数据与微博数据实时抓取这两种数据获取方式中选择了后者。脚本程序编写采用了目前比较流行的 python 语言。

将程序得到的 JSON 格式的用户信息存储为 CSV 格式的数据库文件，然后利用层次分析法确定用户属性权重，并通过属性权重对用户属性值进行属性加权，再对处理过的用户属性值进行归一化处理，完成数据预处理工作。

3.2 微博用户聚类

将经过预处理的数据代入怀卡托智能分析环境（Waikato Environment for Knowledge Analysis，WEKA）中，实现 X-means 聚类算法。

在聚类过程中要注意，实验需要实现对 X-means 聚类算法除属性加权之外的另一项改进，在标准的欧氏距离函数定义中加入一个匹配条件，如果能够与已知的重点人员信息库比对成功，那么就将距离函数中相关节点的属性值乘以其自身的用户权重再进行计算，否则就将其用户权重置为 1，然后将用户节点与微博进行领域匹配，如果匹配成功，原用户权重不变，如果未匹配成功，将其用户权重乘以 0.8 之后再代入距离函数定义中。

将实验数据获取阶段所得到的 CSV 格式数据在 WEKA 环境中打开，使用 X-means 算法聚类之后，得到结果如图 4 所示。

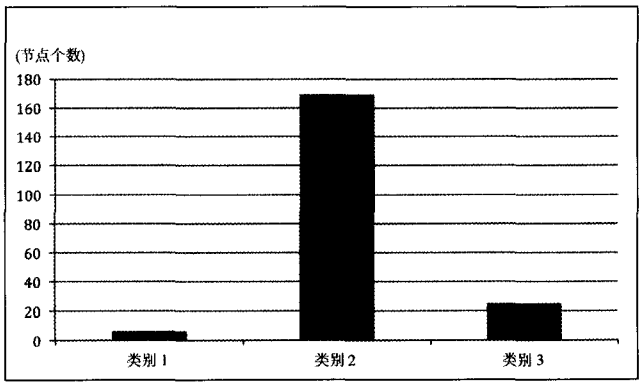

图4 X-means算法聚类结果

另外，为了更加明显地看出各个类簇之间的区别，我们将各个类的每一项属性平均值列于表 1 中。

表1 各个类簇的平均属性值

	类别1	类别2	类别3
微博数	5416.17	6753.39	15848.56
粉丝数	2626.17	2248.39	3401.08
关注数	389	259.06	1405.44
互粉数	79.6	106.64	537.68
转发数	3.5	0.12	0.16
评论数	0.83	0.15	0.08

为了将类簇之间的差异进行量化研究，将每一个聚类中心的各个属性值进行归一化处理，并进行属性加权，最后对所有加权属性值求和，这样就得到了每一个聚类的平均影响力系数 R_i（i 表示聚类类簇编号），如图5所示。

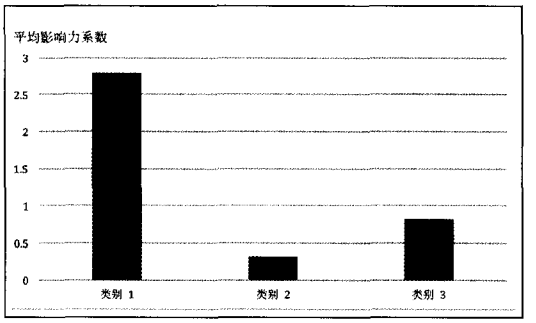

图5 聚类影响力系数对比图

由以上实验结果可以看出，类别1中的用户主要特点是除转发数和评论数高于其他两类很多之外，其他属性值都比较平均，而最终得到的类簇平均影响力系数是远大于其他两类的，这就说明这类用户在微博传播过程中起到的影响力更大；而第二类用户数据集中用户节点的数量最多，用户的平均影响力系数也最小，这类用户的特点是各个属性值都不会特别突出，此类用户就是微博上绝大多数的普通用户，他们对于目标微博的关注程度不是很高；第三类用户的微博数、互粉数和粉丝数都比较多，这一类用户是微博上比较活跃的用户，相对于一般用户的转发而言，这类用户的转发对微博传播的影响力更大，所以最终得出的平均影响力系数介于第一类与第二类之间。

通过这样的用户聚类情况，可以判定这条微博是一条普通的舆情类微博，并不存在网络水军以及网络推手聚集的情况，也不存在网络大V用户对其进行转发评论，但是微博普通用户参与较多，需要对此类舆情类微博进行适当关注和引导，有关微博分类的具体标准，第4部分中将会详细介绍。

3.3 数据库创建

在用户聚类过程中，把微博粉丝过万的用户录入重点人员信息库，并手动指定用户领域，最终得到初步的重点人员信息库，之后利用新浪微博的开放API对每个用户的粉丝以及关注用户信息进行爬取，得到各自领域用户节点的有向图，然后利用HITS算法对每一个用户节点计算权重，将权重列入用户属性中与用户节点一一对应，之后每运行一次用户聚类算法，需要加入新的节点到信息库时，就需要将新加入节点的关注信息和粉丝信息与现有节点进行比对，然后更新每个领域下的用户有向图，再计算出新加入用户的节点权重。

需要注意的一点是，随着重点人员信息库的不断完善，后期加入的节点对于数据库中已经存在的节点来说数量比较少，对数据库中原有节点的权重影响也比较小，因此只需对新加入的节点进行权重计算，而原有节点的权重保持不变。另外，我们可以设置一个新增节点阈值，每当新加入的节点达到现有节点数量的30%，就对整体的数据库进行一次权重更新，运行一次HITS算法，这样既保证数据库的时效性，又避免了不必要的计算过程，提高了数据库的运行效率和使用效率。

3.4 聚类结果分析

将处理前后的数据集先后代入X-means与K-means聚类算法中，并指定聚类个数为3个，得到对应的聚类结果，然后通过计算两种方法聚类结果的轮廓系数（Silhouette Coefficient）来证明改进之后的X-means对此类数据集来说是最优的聚类方法[18]。

轮廓系数是一种评价聚类有效性的方法，一个聚类结果的轮廓系数越高，说明该结果的数据拟合性更好，计算过程为：首先计算聚类结果中每一个节点的凝聚度和分离度来判断节点是否应该被放置于所在的簇类当中，然后得出在某一聚类结果下节点的平均轮廓系数作为该数据集的轮廓系数，上述4种聚类方法所得到的聚类结果的轮廓系数对比情况如图6所示。

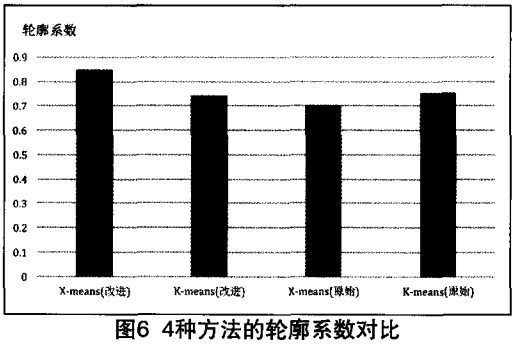

图6 4种方法的轮廓系数对比

由图6可以清楚地看出改进之后的X-means聚类结果具有更高的轮廓系数，由此可以证明本文中对微博用户数据聚类方法的改进是行之有效的。

4 热门微博分类标准

在进行微博用户聚类的实验中，已经对实验中涉及的数据集进行了详细的分析，得到该微博下的用户聚集情况，并以此为根据推测了微博的传播热度以及发展走势。下面通过一种量化办法来确定微博的具体分类标准。

如果我们用 P 来表示一个微博的重要程度，N_i 表示某微博下的转发用户聚类的第 i 个类簇的用户节点个数，N 表示该微博下所有用户个数，设微博用户共聚为 n 类，P 值计算方法如公式（13）所示。

$$P=\sum_{i=1}^{n}\frac{N_i}{N}R_i \quad \cdots\cdots(13)$$

得到最终的 P 值即为我们所分析的微博重要性指数，P 值越高，表示参与到该微博讨论中的有效用户更多，并且微博更容易传播发展。根据不同的 P 值，我们粗略设置了三类微博供公安部门进行微博鉴别，如表2所示。

表2 微博分类标准

P 值	微博种类	特性
[0,0.1]	网络水军类	此类微博转发用户中存在大量僵尸用户，围绕少量拥有众多僵尸粉丝的用户对原微博进行转发或者二次转发
[0.1,1]	普通舆情类	此类微博的转发用户大多是微博普通用户，也存在部分微博活跃用户，没有相关领域内的重点人员对其进行转发
[1,+∞]	社会热点类	此类微博的转发用户有大部分微博普通用户以及部分微博活跃用户，最重要的是此类微博通常会被微博话题相关领域的重点人员或者是网络上的一些大V用户转发，因此具有更广泛的社会影响力

5 结束语

本文提出了一种贴近公安实战的微博分类系统设计思路，首先对热门微博下的转发用户信息进行挖掘。其次根据其属性进行用户聚类，得到相应微博的用户聚集状态，聚类方法采用了现如今比较流行的X-means聚类算法，并对X-means算法进行了两方面改进，一方面对微博用户的属性值通过层次分析法进行加权；另一方面为重点人员建立信息库，并设定人员领域，利用HITS算法以及同领域内重点用户之间的相互关注关系得出重点人员权重，并将该权重返回至用户聚类过程中，使得用户聚类方法更加科学精确。最后，针对聚类得到的微博用户聚集状态计算微博的重要性系数 P，公安机关可以利用该系数值来判断目标微博是否值得进行舆情监控，从而针对不同种类的微博制定不同的应对策略。（责编 吴晶）

参考文献：

[1] 王明元，贾焰，周斌，等．一种基于主题相关性分类的微博话题立场研判方法 [J]. 信息网络安全，2014（9）：17-21.

[2] 江斌．微博自动分类方法研究及应用 [D]. 哈尔滨：哈尔滨工业大学，2012.

[3] 严岭，李逸群．网络舆情事件中的微博炒作账号发现方法研究 [J]. 信息网络安全，2014（9）：26-29.

[4] 周咏梅，杨佳能．面向文本情感分析的中文情感词典构建方法 [J]. 山东大学学报：工学版，2013（6）：27-33.

[5] 柳俊，周斌，黄九鸣．基于二部图投影的微博事件关联分析方法研究 [J]. 信息网络安全，2014（9）：44-49.

[6] 曹海涛．基于PAD模型的中文微博情感分析研究 [D]. 大连：大连理工大学，2013.

[7] 谢丽星，周明，孙茂松．基于层次结构的多策略中文微博情感分析和特征抽取 [J]. 中文信息学报，2012, 26(1):73-83.

[8] 杜伟夫．文本倾向性分析中的情感词典构建技术研究 [D]. 哈尔滨：哈尔滨工业大学，2010.

[9] 高永兵，郭文彦，周环宇，等．基于K-means的私人微博聚类算法改进 [J]. 微型机与应用，2014(14):78-81.

[10] 张雪凤，张桂珍，刘鹏．基于聚类准则函数的改进K-means算法 [J]. 计算机工程与应用，2011, 47(11):123-127.

[11] 王荣，李晋宏，宋威．基于关键字的用户聚类算法 [J]. 计算机工程与设计，2012, 33(9):3553-3557.

[12] 李磊，刘继．面向舆情主题的微博用户行为聚类实证分析 [J]. 情报杂志，2014(3):118-121.

[13] 曹鹏，李博，栗伟，等．结合X-means聚类的自适应随机子空间组合分类算法 [J]. 计算机应用，2013, 33（2）：550-553.

[14] 赵峥．基于两种改进的聚类算法对新浪微博用户信息的研究 [D]. 北京：首都经济贸易大学，2014.

[15] 何黎，何跃，霍叶青．微博用户特征分析和核心用户挖掘 [J]. 情报理论与实践，2011(11):121-125.

[16] 杨凯，张宁．微博用户关系网络的结构研究与聚类分析 [J]. 复杂系统与复杂性科学，2013, 10(2):37-43.

[17]PELEG D,MOORE A. X-means: Extending K-means with Efficient Estimation of the Number of Clusters [C]//Seventeenth International Conference on Machine Learning.San Francisco:Morgan Kaufmann Publishers, 2000: 89-97.

[18]DEY D, SOLORIO T, et al. Instance Selection in Text Classification Using the Silhouette Coefficient Measure.[J]. Lecture Notes in Computer Science, 2011(94):357-369