doi:10.11959/j.issn.1000-436x.2016239

基于层次分析的微博短文本特征计算方法

邹学强[1,2,3]，包秀国[2]，黄晓军[4]，马宏远[2]，袁庆升[1,2,3]

（1. 中国科学院信息工程研究所，北京 100093；2. 国家计算机网络应急技术处理协调中心，北京 100029；
3. 中国科学院大学，北京 100049；4. 北京邮电大学信息与通信工程学院，北京 100876）

摘 要：为了建立用户精准兴趣模型以有效发现具有相似兴趣的用户群，提出了一种针对微博的短文本特征计算方法用于聚类算法，提升聚类效果以更好地挖掘微博用户的相似兴趣集合。该方法融合了微博转发数、评论数、点赞数等多个关键指标来度量微博短文本特征的重要性。同时，引入层次分析技术，改进了传统的 tf-idf 特征计算方法，并利用经典文本聚类算法进行实验。实验结果表明，改进后的短文本特征计算方法与传统的 tf-idf 特征计算方法相比，在类内集中度和类间分散度上取得了更好的效果。

关键词：层次分析；特征计算；文本聚类；短文本

中图分类号：TP391.1 **文献标识码**：A

Calculating the feature method of short text based on analytic hierarchy process

ZOU Xue-qiang[1,2,3], BAO Xiu-guo[2], HUANG Xiao-jun[4], MA Hong-yuan [2], YUAN Qing-sheng[1,2,3]

(1. Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China;
2. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China;
3. University of Chinese Academy of Sciences, Beijing 100049, China;
4. School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China）

Abstract: In order to model the accurate interest preference of microblog users and discover user groups with similar interest, a new method was proposed which considered the total amount of retweets, comments and attitudes of each microblog for text feature calculation with utilizing classic analytical hierarchy process method. The proposed method used three indicators to evaluate the importance of the text feature representation and made an improvement on traditional tf-idf feature calculation method to fit for short text. Furthermore, this method was also implemented in the traditional clustering algorithm. Experimental results show that, compared with the traditional tf-idf method, the improved approach has a better clustering effect on the average scattering for clusters and the total separation between clusters.

Key words: analytic hierarchy process, feature calculation, text clustering, short text

1 引言

近年来，快速发展的社交网络已成为人们交流信息的重要平台。其中，微博作为主流社交平台之一，吸引了越来越多的网民参与其中。随着微博用户规模的迅速膨胀，微博平台上产生和流动的大量数据(如朋友关系、用户发布的内容等)具有重要的研究意义和应用价值。其中，如何基于微博平台中的各种信息来辅助发现和定位广告目标用户群，进而提升微博营销服务和产品的有效性，已成为近年来的研究热点之一。为实现精准的广告投放，必须科学分析用户的兴趣并创建合理有效的推荐模型。因此，用户兴趣建模技术是实现精准广告投放和产品推荐必不可少的前提条件和核心技术之一。

收稿日期：2016-05-05；**修回日期**：2016-11-24

基金项目：国家高技术研究发展计划（“863”计划）基金资助项目（No.SS2014AA012303）；国家自然科学基金资助项目（No.61300206, No.61402123）

Foundation Items: The National High Technology Research and Development Program (863 Program) (No.SS2014AA012303), The National Natural Science Foundation of China (No.61300206, No.61402123)

微博信息的内容分析是用户兴趣建模的关键环节，而有效的文本分析依赖于良好的文本表示和特征计算方法。向量空间模型中的 tf-idf 方法是目前广泛使用且效果较好的一种文本特征计算方法。由于微博长度一般在 100 个字左右，微博文本实际上是由海量短文本构成的集合。相比传统的文本内容，其具有数量大、长度短、特征词少及富含噪声等特点，使传统 tf-idf 方法在短文本特征计算上面临巨大困难和挑战。

为了解决信息稀疏问题，Amr[1]、David[2]和 Bollegal 等[3]分别从背景知识及语义解析等方面对短文本特征进行扩展，Sun[4]和 Ramge[5]研究了微博的去噪问题。在前人研究的基础上，本文通过分析微博短文本结构和数据的特点，针对微博特有的转发数、评论数、点赞数等属性特征，提出了一种基于层次分析的短文本特征计算方法。实验结果表明，本文改进后的特征计算方法的聚类效果明显优于传统的 tf-idf 特征计算方法，可以提高针对微博短文本的用户兴趣建模的准确性，为实现个性化推荐提供技术基础。

2　相关工作

由于微博中用户特性复杂、行为表现存在差异性，使微博数据中夹杂了大量的噪声数据，如何针对每个用户建立精准的个性化模型是一个研究热点，也是一个难点。目前，已有的研究主要是通过分析用户的行为和属性信息，获取描述用户兴趣的特征或关键词，从而对用户兴趣进行建模。相关工作大致分为 3 类：基于主题发现的兴趣模型、基于用户标签的兴趣模型和基于用户关联关系的兴趣模型。

Weng 等[6]把每个用户发布的所有 Twitter（加上 Twitter 网址的 citation）看作一个大集合，通过在 PageRank 算法中引入基于用户兴趣的用户相似度，利用主题模型生成用户兴趣，从而找出 Twitter 的某个话题下具有影响力的用户。Abel 等[7]提取微博中的散列标签等与传统的新闻媒体进行关联，丰富了微博的语义，进而判定出用户的实际兴趣。Welch 等[8]利用用户间的关注和转发 2 种关系构建得到兴趣关系图，发现用户转发的微博可以有效反映用户对话题的兴趣程度。Liu 等[9]利用基于机器翻译和词频统计相结合的方法，通过从微博文本中抽取关键词来挖掘用户的兴趣。

邱云飞等[10]提出微博短文本重构的概念，根据微博文本包含的特殊符号对文本内容进行扩展，以抽象的文本向量为基础进行聚类，缓解短文本带来的信息稀疏性问题，提高了文本聚类效果，从而改善了用户兴趣集合的划分效果。宋巍等[11]以支持向量机(SVM)作为分类模型，采取词语层次特征与主题层次特征相结合的策略，构建训练分类器特征，实现基于微博分类的用户兴趣识别方法。方维[12]通过统计相关数据以及问卷调查的方式对微博用户的行为进行分析，采用文本分类与主题词匹配相结合的方法, 有效检测识别了用户兴趣。张俊林等[13]利用微博用户间的关联关系，以用户为节点建立图模型，引入标签传播算法，根据邻居标签修正自身标签，通过多次迭代最终为每个用户推荐合适的兴趣标签。

本文提出的基于层次分析的短文本特征计算方法可以更加准确地从用户的微博中提取用户标签来描述用户的兴趣。此外，由于改进后的短文本特征计算方法可以提高用户微博短文本聚类效果，因此，可以更好地挖掘微博用户的相似兴趣集合。

3　短文本特征计算方法

3.1　文本特征表示

文本聚类的目标是使同类文档的相似度尽量大，而不同类文档的相似度尽量小。其首要问题是如何将文本内容进行形式化表示，以转换成计算机可以理解的形式，从而进行文档的相似度度量。在众多的文本表示方法中，由 Salton 等[14]于 20 世纪 70 年代提出向量空间模型(VSM, vector space model)，因其较强的可计算性和可操作性而受到广泛应用。在该模型中，文档的内容被映射为多维空间中的一个点，通过向量的形式给出。其核心思想是将文档分解为由词条特征构成的向量，具体做法是将文档进行分词，然后计算文档中每个词条的权值，即用特征词条及其权值表示文档信息。具体表示形式如下

$$V(d)=(t_1,w_1(d),\cdots,t_i,w_i(d),\cdots,t_n,w_n(d)) \quad (1)$$

其中，d 表示一个文档，t_i 表示文档集合中的某个词条，$w_i(d)$ 表示词条 t_i 在 d 中的权值。

这种表示形式简单直接，但是由于微博文本的特点使文档特征向量的维数可以达到数万甚至数十万，从而导致信息稀疏问题，且如此高维向量空间使聚类算法的处理时间大大增加，并对算

法的准确性产生不利影响。因此，对特征空间进行降维处理显得非常必要和关键。最有效的特征降维方法就是通过特征选取，去掉某些表征文档能力差的词。具体到文档相似度计算，就是减少词语的数量。无监督的特征选择方法，如主成分分析法(PCA)、隐语义索引(LSI)、奇异值分解(SVD)等，并不能选择出具有表征力的特征词，并且无法得到每一维特征的实际含义。有监督的方法通常需要类信息，常用的有监督特征选择方法的评估函数有文档频率、信息增益、期望交叉熵、互信息、卡方统计法等。本文将采用词频和文档频率作为特征选取的评估函数，认为非常稀缺的词或经常在多个文档中出现的词对聚类产生的影响较小。

3.2　文本特征计算

在构建文本向量空间过程中，需要为划分好的词条赋予适当的权值，权值代表该词条对表征文本内容的能力，权值越大，说明该特征项对文本的区分度或分类越好。因此，为了提高文本聚类结果，在构建文本向量时应该尽可能保留原有的文档信息。常见的特征计算方法主要有布尔值、tf（term frequency）和 tf-idf（term frequency-inverse document frequency）等。其中，tf-idf 方法得到了广泛的应用，词频 *tf* 反映特征项在同一文本内部的分布情况，逆文档频率 *idf* 反映同一特征项在不同文本上的分布情况。传统的 tf-idf 特征计算公式如下

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2}$$

$$idf_i=\ln\left(\frac{N}{N_i}+0.01\right) \tag{3}$$

$$w_{ij}=tf_{i,j}\cdot idf_i \tag{4}$$

其中，$n_{i,j}$ 表示词条 i 在文本 j 中出现的频次，N_i 表示词条 i 在整个文本空间中出现的频次，N 表示整个文本空间中文本的数目。

3.3　基于层次分析的短文本特征计算方法

传统的 tf-idf 特征计算方法比较适用于长文本，无法兼顾微博文本作为短文本本身所包含有用信息少、可供抽取信息贫乏的特点。因此，需要借助微博平台提供的其他信息来改进短文本特征的计算方法。

通过观察分析如图 1 所示的微博示例发现，一条有价值的微博很可能被其他用户转发、评论和点赞，而这 3 种用户互动行为能够有效地体现该微博在整个微博文档集合中的重要性。因此，本文尝试利用微博的转发数、评论数和点赞数作为特征参数对传统的 tf-idf 计算方法进行改进，使改进的 tf-idf 文本特征计算方法更加符合短文本的特点。

图 1　微博示例

每条微博的转发数、评论数、点赞数可以看作是衡量微博重要性的 3 个指标，但这 3 个指标在表示能力上有所差别。为了反映这 3 个指标在表征微博重要性上的差异，需要定量度量指标的权值。因此，本文引入层次分析法[15]（AHP, analytic hierarchy process）以量化这 3 个指标在反映微博重要性时的权值，这样进行综合计算时得到的结果可以更好地反映实际情况。层次分析法是一种应用广泛且效果较好的权值确定方法，它把复杂问题中的各因素划分成相关联的有序层次，形成条理化的多目标、多准则的决策方法，是一种将定量分析与定性分析相结合的有效方法。基于层次分析的短文本特征计算方法具体步骤如下。

1) 构造判断矩阵。判断矩阵是层次分析法的基本信息，是进行权值计算的重要依据。本文使用 Saaty 提出的传统 1~9 标度法对指标进行两两比较得到量化的判断矩阵。在矩阵中，第 i 行、第 j 列所表达的含义如表 1 所示。在判断矩阵中，$a_{i,j}=\dfrac{1}{a_{j,i}}$ 。

表 1　指标两两比较时权值等级及其赋值

标度 $a_{i,j}$	含义
1	i 因素与 j 因素同等重要
3	i 因素比 j 因素略重要
5	i 因素比 j 因素较重要
7	i 因素比 j 因素非常重要
9	i 因素比 j 因素绝对重要
2,4,6,8	介于以上判断之间的状态的标度

这里使用下标 i=1,2,3 对应微博转发数、评论数

和点赞数。通过对微博用户使用特点分析可以发现，用户的点赞行为往往表示对微博内容的赞赏，转发行为往往表示用户希望将所浏览的信息传递给其他用户，评论行为往往表示用户对微博信息的看法。因此，本文认为用户点赞行为的重要性最高，转发行为的重要性高于评论行为。综合分析结果，本文假定了转发数、评论数和点赞数的标度，根据经验设置判断矩阵如下所示。

$$\boldsymbol{A}=\begin{bmatrix} 1 & 3 & \frac{1}{3} \\ \frac{1}{3} & 1 & \frac{1}{4} \\ 3 & 4 & 1 \end{bmatrix}$$

2) 计算重要性排序。根据判断矩阵，使用方根法计算判断矩阵的特征向量。

首先，利用式(5)计算判断矩阵每行所有元素的几何平均值，得到 $\overline{\boldsymbol{w}}=(\overline{w_1},\overline{w_2},\overline{w_3})^{\mathrm{T}}$，然后，利用式(6)对 $\overline{w_i}$ 进行归一化。

$$\overline{w_i}=\sqrt[n]{\prod_{j=1}^{n}a_{i,j}} \tag{5}$$

$$w_i=\frac{\overline{w_i}}{\sum_{i=1}^{n}w_i} \tag{6}$$

通过以上求解，可以得到矩阵 $\boldsymbol{A}$ 的权值向量。

$$\boldsymbol{w}=(w_1,w_2,w_3)^{\mathrm{T}}=(0.272\,1,0.119\,9,0.608\,0)^{\mathrm{T}}$$

3) 一致性检验。对判断矩阵进行一致性检验，以检验权值的分配是否合理。一致性检验主要使用判断矩阵的一般一致性（*CI*）和随机一致性比率（*CR*）2 个指标。

$$CI=\frac{\lambda_{\max}-n}{n-1} \tag{7}$$

$$CR=\frac{CI}{RI} \tag{8}$$

其中，$\lambda_{\max}$ 是指判断矩阵的最大特征根，*RI* 表示判断矩阵的随机一致性均值，与矩阵的阶数有关。1~9 阶判断矩阵的 *RI* 值如表 2 所示。

当 $CR\leqslant 0.1$ 或 $\lambda_{\max}=n$，$CI=0$ 时，判断矩阵具有满意的一致性，认为判断矩阵一致性可以接受；否则，必须对判断矩阵进行修改调整，直至 $CR\leqslant 0.1$ 使其具有良好的一致性。

表 2　随机一致性指标 *RI*

n	*RI*
1	0
2	0
3	0.58
4	0.90
5	1.12
6	1.24
7	1.32
8	1.41
9	1.45

按照上述方法对矩阵 $\boldsymbol{A}$ 进行一致性检验，计算得到其 *CR* 值为 0.071 3，满足判断矩阵的一致性条件。

利用求解出的短文本特征向量 $\boldsymbol{w}$，对微博短文本的特征计算方法进行改进。由于单条微博短文本特征稀疏，可以分别将同一个用户某个时间周期内发布的微博文本归纳为一个文档来进行短文本特征计算。设 x_{jk1}、x_{jk2}、x_{jk3} 分别表示第 j 篇微博文档中第 k 条微博归一化后的转发数、评论数、点赞数指标，对 $n_{i,j}$ 做如下改进。

$$n_{i,j}=\sum_{k=1}^{m}[1+\lambda(w_1x_{jk1}+w_2x_{jk2}+w_3x_{jk3})]I_{i,k} \tag{9}$$

其中，λ 是权值影响因子，m 表示第 j 篇微博文档包含的微博条数，$I_{i,k}$ 表示词项 i 在第 k 条微博中出现的频次。

4　实验分析

4.1　实验数据

微博用户中存在大 V 用户、普通用户、营销账号、官方用户和僵尸账号等不同类型的用户，这些用户的质量参差不齐，如果不加区分地对用户进行兴趣建模和信息推荐，必然会影响最终的推荐效果，因此，需要对有效的普通活跃（非大 V）用户进行识别。本文使用 LIBSVM 软件[16]训练 SVM 分类器，利用训练好的分类器对未分类用户进行识别分类，针对识别出的普通微博用户进行兴趣建模。

本文利用新浪微博提供的 API 接口，爬取并识别获得 1 879 名普通活跃微博用户及其发布的 2 113 703 条微博数据。微博数据需要进行预处理和数据清洗，包括去除噪声信息（微博中的链接、转发标志、@用户名和表情符号等）、分词、停用词。由于名

词更能反映用户的兴趣偏好，这里仅针对名词进行特征选择，最终获得 18 686 个特征词。

4.2 评价指标

集中度和分散度是评价聚类方法常用的 2 个指标。其中，集中度是指簇中的成员必须尽可能靠近，分散度是指簇之间的距离，要尽可能大。文献[17]提出的 *SD* 有效性指标可以作为对比改进前后聚类效果的指标。*SD* 有效性指标定义如下

$$SD(C)=\alpha\cdot Scat(C)+Dis(C) \tag{10}$$

其中，α 为加权因子，用于权衡簇平均分散度 ($Scat(C)$) 和簇间总体离散度 ($Dis(C)$) 之间的相对重要性。$Scat(C)$ 和 $Dis(C)$ 的定义分别如下。

$$Scat(C)=\frac{\frac{1}{C}\sum_{i=1}^{C}\left\|\boldsymbol{\sigma}(\boldsymbol{v}_i)\right\|}{\left\|\boldsymbol{\sigma}(\boldsymbol{X})\right\|} \tag{11}$$

$$Dis(C)=\frac{D_{\max}}{D_{\min}}\sum_{k=1}^{C}\left(\sum_{z=1}^{C}\left\|\boldsymbol{v}_k-\boldsymbol{v}_z\right\|\right)^{-1} \tag{12}$$

其中，C 是聚类的簇数，$\left\|\boldsymbol{\sigma}(X)\right\|$ 是整个数据集的方差向量的模，$\left\|\boldsymbol{\sigma}(v_i)\right\|$ 是第 i 个簇的方差向量的模，$D_{\max}$ 是各个簇中心之间的最大距离，$D_{\min}$ 是各个簇中心之间的最近距离，$\left\|\boldsymbol{v}_k-\boldsymbol{v}_z\right\|$ 表示第 k 个和第 z 个簇中心之间的距离。

由式(10)~式(12)可得，对于相同的数据集，如果利用不同的聚类算法得到的聚类结果的平均分散度和簇间总体离散度越小，则其聚类性能会越好，当 $SD(C)$ 使值最小时，此时的 C 为最优的聚类个数。

4.3 实验结果

针对微博文本数据集，分别利用传统的 tf-idf 方法和改进的 tf-idf 方法进行文档表示，利用 *K*-means 算法对微博文档集进行聚类，对二者的聚类效果进行比较，如图 2 所示。实验参数选择为 $\alpha=0.5$，$\lambda=3$。

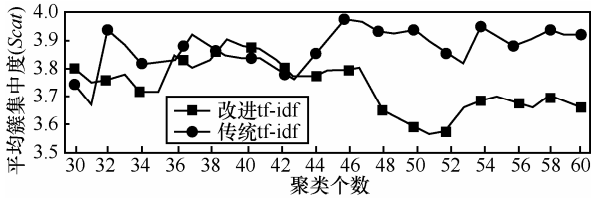


图 2 传统 tf-idf 特征计算方法和改进 tf-idf 特征计算方法聚类平均分散度对比

Scat 指标用来表征聚类结果的平均簇集中度，其对应的值越小，说明平均每个类的类内距离越小。由图 2 可知，同传统的 tf-idf 特征计算方法相比，改进后方法的 *Scat* 指标整体上更小，当聚类个数超过 42 个以后，这种差距则体现得更为明显。这表明改进后的 tf-idf 特征计算方法提高了类内的相似性，减小了类内文档间的距离。

Dis 指标表示簇间整体的分离程度，其对应的值越小，说明所有类整体的分散度越大。由图 3 可知，同传统的 tf-idf 特征计算方法相比，改进后方法的 *Dis* 指标整体上更小。这表明改进后的 tf-idf 特征计算方法提高了不同簇之间的差异性，增大了簇间距离。

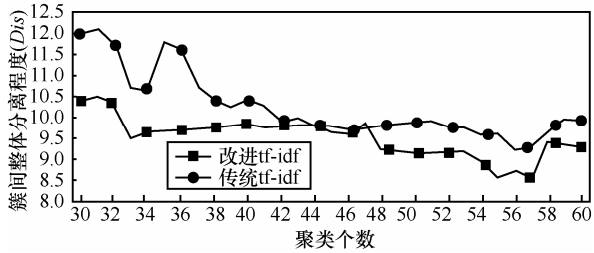


图 3 传统 tf-idf 特征计算方法和改进 tf-idf 特征计算方法聚类簇间总体离散度对比

由图 4 可知，当聚类个数在 56 左右时，2 类计算方法聚类效果达到最优，并且在最优点处，改进后方法的 *SD* 有效性指标要明显小于传统的 tf-idf 特征计算方法，这表明改进后的 tf-idf 特征计算方法具有更优的聚类效果。

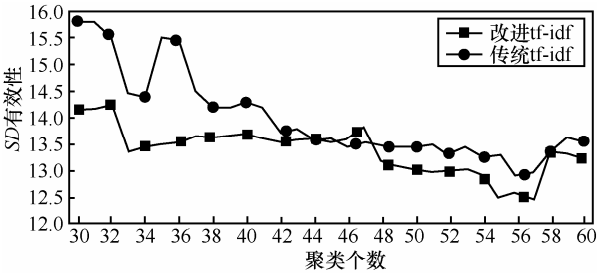


图 4 传统 tf-idf 特征计算方法和改进 tf-idf 特征计算方法聚类综合指标对比

综上所述，同传统的 tf-idf 特征计算方法相比，改进后的 tf-idf 特征计算方法在平均分散度、总体离散度和 *SD* 有效性 3 个指标上均有明显提升，聚类效果更为理想。

5 结束语

随着微博平台用户量和数据量的日益增大，微博内容分析和挖掘具有重要的研究价值和应用前景。由于微博短文本具有长度短、富含冗余信息等

特点，已有的文本表示和特征计算方法难以适用。因此，本文针对微博提出一种基于层次分析的短文本特征计算方法，借助微博短文本以外的其他数据和特征，对不同短文本特征的重要性进行度量。实验表明，该方法能够很好地对微博短文本进行表示，在一定程度上提升了微博文档的聚类效果，尤其在聚类簇数更高的设置上，本文提出的改进后的 tf-idf 特征计算方法表现出更好的性能。

本文的创新点在于采用层次分析法分析转发数、评论数和点赞数，并根据实际经验设置判断矩阵。最后通过实验证明了融入转发数、评论数、点赞数等特征和判断矩阵的有效性，在微博文档聚类中获得了积极效果。后续工作将研究如何设置最优判断矩阵等问题。

参考文献：

[1] AMR A, LIANG J H, ALEXANDER J S. Hierarchical geographical modeling of user locations from social media posts[C]//The 22nd International Conference on World Wide Web, 2013: 25-36.

[2] DAVID J. That's what friends are for: inferring location in online social media platforms based on social relationships[C]//The 7th International Conference on Weblogs and Social Media. 2013: 273-282.

[3] BOLLEGALA D, MATSUO Y, ISHIZUKA M. Measuring the similarity between implicit semantic relation using web search engines[C]//The 2nd ACM International Conference on Web Search and Data Mining WSDM'09, 2009: 104-113.

[4] SUN A. Short text classification using very few words[C]//The 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. NewYork, USA, 2012: 1145-1146.

[5] RAMGE D, DUMAIS S, LIEBLINGI D. Characterizing microblogs with topic models[C]//ICWSM, 2010: 130-137.

[6] WENG J S, LIM E P, JIANG J, et al. TwitterRank: finding topic-sensitive influential Twitterers[C]//The 3th ACM International Conference on Web Search and Data Mining. New York City, 2010: 261-270.

[7] ABEL F, GAO Q, HOU B G J, et al. Semantic enrichment of twitter posts for user profile construction on the social Web [C]//The 8th Extended Semantic Web Conference on the Semanic Web: Research and Pages (ESWC'11). 2011: 375-389.

[8] WELCH M J, SCHONFELD U, HE D, et al. Topical semantics of Twitter links [C]//The 4th ACM International Conference on Web Search and Data Mining (WSDM'11). 2011: 327-336.

[9] LIU Z, CHEN X, SUN M. Mining the interests of Chinese microbloggers via keyword extraction [J]. Frontiers of Computer Science in China, 2012, 6(1): 76-87.

[10] 邱云飞, 王琳颖, 邵良杉, 等. 基于微博短文本的用户兴趣建模方法[J]. 计算机工程, 2014, 40(2): 275-279.
QIU Y F, WANG L Y, SHAO L S, et al. User interest modeling approach based on short text of micro-blog[J]. Computer Engineering, 2014, 40(2): 275-279.

[11] 宋巍, 张宇, 谢毓彬, 等. 基于微博分类的用户兴趣识别[J]. 智能计算机与应用, 2013, 3(4): 80-83.
SONG W, ZHANG Y, XIE Y B, et al. Identifying user interests based on microblog classification[J]. Intelligent Computer and Applications, 2013, 3(4): 80-83.

[12] 方维. 微博兴趣识别与推送系统的研究与实现[D]. 华中科技大学, 2012.
FANG W. Research and implement of micro-blog interest found and pushing system[D]. Huazhong University of Science and Technology, 2012.

[13] 张俊林. 标签传播算法在微博用户兴趣图谱的应用[J]. 程序员, 2012, 1(7): 50-53.
ZHANG J L, Application of label propagation algorithm in user profiles of micro-blog[J]. Programmer, 2012, 1 (7): 50-53.

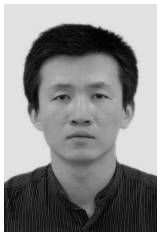
[14] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing [J]. Communications of the ACM CACM Homepage, 1975, 18(11): 613-620.

[15] 常建娥, 蒋太立. 层次分析法确定权重的研究[J]. 武汉理工大学学报（信息与管理工程版）, 2007, 29(1): 153-156.
CHANG J E, JIANG T L, Research on determining weights by analytic hierarchy process[J]. Journal of Wuhan University of Technology (Information & Management Engineering), 2007, 29(1): 153-156.

[16] CHANG C C, LIN C J. LIBSVM: a library for support vector machines [J]. ACM Transactions on Intelligent Systems & Technology, 2011, 2(3): 389-396.

[17] HALKIDI M, VAZIRGIANNIS M, BATISTAKIS Y. Quality scheme assessment in the clustering process[J]. Lecture Notes in Computer Science, 2000, 1910(1): 265-276.

作者简介：



邹学强（1978-），男，福建莆田人，中国科学院信息工程研究所博士生，主要研究方向为信息处理、信息安全、网络流量分析等。

包秀国（1962-），男，江苏如皋人，博士，中国科学院信息工程研究所教授、博士生导师，主要研究方向为信息网络安全、音视频处理、网络空间测绘等。

黄晓军（1990-），男，江西九江人，北京邮电大学硕士生，主要研究方向为数据挖掘、信息安全。

马宏远（1981-），男，辽宁朝阳人，博士，国家计算机网络应急技术处理协调中心高级工程师，主要研究方向为智能信息处理。

袁庆升（1980-），男，山东济南人，中国科学院信息工程研究所博士生，主要研究方向为多媒体大数据处理、网络与信息安全。