基于奇异值分解的新闻标题聚类研究

文晓艺,郝程程
(上海对外经贸大学 统计与信息学院,上海 201600)

摘 要:汉语分词技术和文本聚类是自然语言处理的重要环节,在文本信息的组织、摘要和导航中应用广泛。文本聚类作为一种无监督学习算法,其依据是聚类假设:同类的文档相似程度大,不同类的文档相似程度小。文中主要研究汉语文本聚类算法在新闻标题类文本中的应用。首先对采集到的若干条新闻标题进行分词和特征提取,将分词后的文本转化为词条矩阵;然后使用 TF-IDF 技术处理词条矩阵,得到基于分词权重的新的词条矩阵,对新的词条矩阵进行奇异值分解,得到主成分得分矩阵,提取主成分分析文本特征并根据主成分得分矩阵进行 K-均值和分层聚类分析;最后将聚类结果用词云图的形式展示出来并评价聚类效果的好坏。实证显示,对词条矩阵的奇异值分解能降低向量空间的维数,提高聚类的精度和运算速度。

关键词:汉语分词;词云图;奇异值分解;潜在语义分析;K-means 聚类
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2020)02-0042-05
doi:10.3969/j.issn.1673-629X.2020.02.009

Study on News Header Clustering Based on Singular Value Decomposition

WEN Xiao-yi, HAO Cheng-cheng
(School of Statistics and Information, Shanghai University of International Business and Economics, Shanghai 201600, China)

Abstract: Chinese word segmentation and text clustering are important in natural language processing, which are widely used in text information organization, summarization and navigation. As an unsupervised learning algorithm, text clustering is based on the clustering hypothesis: documents of same category are more similar, while documents of different categories are less similar. We mainly study the application of Chinese text clustering algorithms in news headers. First of all, we divide the collected news headlines into word segmentation and feature extraction, and convert the text after word segmentation into term line matrix. Then the term line matrix is processed by TF-IDF technology and a new lexical matrix based on word segmentation weight is obtained. The new lexical matrix is decomposed by singular value and the principal component scoring matrix is obtained. The text features of principal component analysis are extracted and K-means and hierarchical cluster analysis are performed according to the scoring matrix of principal component analysis. Finally, the clustering results are displayed in the form of a word cloud map and the quality of the clustering effect is evaluated. The experiment shows that the singular value decomposition of the lexical matrix can effectively reduce the dimension of the vector space, thus improving the accuracy and speed of the clustering.

Key words: Chinese word segmentation; word cloud diagram; singular value decomposition; latent semantic analysis; K-means clustering

0 引 言

文本聚类技术作为文本挖掘技术的重要分支之一,有着非常广泛的应用。文本聚类的一个难点在于文本特征词提取,对此 Dumais(1998)提出了隐含语义索引 LSI 来构造向量空间模型,通过对原文本词条矩阵进行奇异值分解来进行降维,提高聚类的效率[1]。在国内的文本挖掘相关领域中,姜宁和史忠植(2002)在对聚类分析模型进行比较的基础上,提出了贝叶斯后验模型选择方法,给出了一个用于文本聚类分析的概率模型[2];徐建锁等(2004)应用动态自组织映射神经网络来实现文本聚类,这种方法不必预先给定聚类个数,使得聚类过程更加灵活[3-4];姚清耘等(2008)探讨了基于向量空间模型的文本聚类方法,并提出了一种文本聚类的改进方法——LP 算法[5];宋涛等(2010)根据潜在语义分析模型,提出了截断奇异值分解中 K 值的选取方法来降低文本空间的维度[6]。

收稿日期:2019-03-04 **修回日期**:2019-07-08 **网络出版时间**:2019-11-07
基金项目:上海市大学生创新训练项目(201810273116)
作者简介:文晓艺(1997-),女,通讯作者,研究方向为统计学、生存分析、高维数据分析;郝程程,博士,副教授,研究方向为多元纵向数据。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20191107.0908.020.html

文中主要研究对新闻标题这种汉语短文本的聚类问题。短文本与长文本不同,因为每条文本包含的信息较少,基于现有的以词条矩阵为基础的文本聚类,点与点之间的距离更近,研究表明现有聚类技术的效果并不明显。针对汉语简短文本,介绍了目前主要的自动分词、特征词提取以及 k-均值聚类方法,并通过利用一组新闻标题数据,对主流汉语简短文本的聚类方法进行了评估。

1　汉语短文本的自动分词与特征选择

与一般文本相比,文中研究的新闻标题文本有两个突出特点:一是文本长度短,主要指新闻标题的一般不超过 30 个汉字;二是未登录词多,由于新闻类文本的即时性与专业性,其普通新词、专有名词与专业术语的出现频率远高于语料库一般文本。因此,文中在进行自动分词与特征提取时,重点关注这两个问题。

1.1　汉语新闻标题自动分词算法

自动分词问题是聚类孤立语与黏着语(如汉语、日语等)文本时的首要基础性工作。文本分词,即将整段文本切割为词,是获取特征词的频率及其文档频率,从而将文本数字化不可缺少的环节。然而,与西方屈折语文本不同的是,孤立语与黏着语缺乏显示标志指示词与词的分隔,因此需要计算机系统进行自动分词。汉语文本分词的主要难点是分词歧义[7]。但是,对于新闻标题类文本,未登录词对于分词精度的影响远高于歧义切分。

国内外众多学者在孤立语与黏着语自动分词领域已进行了大量研究[8]。诸多国内学者(如文献[9])将现有研究分为两大类:基于词表的分词方法和基于统计模型的分词方法。前者包括正向最大匹配法、双向扫描法、组词遍历法等,后者可分为基于词的生成式模型(word-based generative model)与基于字的区分式模型(character-based discriminative model)两大类。文中选用了三种基于统计模型的分词方法。

1.1.1　基于词的二元语法模型

基于词的 n 元文法模型属于生成式分词方法,是目前主流的统计分词方法之一。基于词的二元语法模型的最大概率法表述如下:

假设 S 为一个汉字序列,$W=w_1\,w_2\cdots w_N$ 是 S 可能切分出的词序列,N 为词序列长度。最大概率分词过程实际上即求解使概率 $P(W\mid S)$ 最大的切分词序列 W^*。根据贝叶斯公式,即:

$$W^*=\arg\max_W P(W\mid S)=\arg\max_W P(W)P(S\mid W) \tag{1}$$

其中,$P(S\mid W)$ 为生成模型,$P(W)$ 为语言模型,若 $P(W)$ 采用二元语法,可以表示为:

$$P(W)=P(w_1)\prod_{i=1}^{N}P(w_i\mid w_{i-1})$$

1.1.2　基于字的一阶隐马尔可夫模型

另一种统计分词方法是基于字构词的区分式分词方法。基于字构词的一阶隐马尔可夫模型法表述如下:

对于 $W=w_1\,w_2\cdots w_N$,采用文献[7]的做法,把字序列 $S=s_1\,s_2\cdots s_n$ 转换成可能的词位序列 $T=t_1\,t_2\cdots t_n$,其中规定每个字只有 4 个词位:词首(B)、词中(M)、词尾(E)和单独成词(S)。则最大概率分词过程即求解:

$$T^*=\arg\max_T P(T\mid S)=\arg\max_T P(T)P(S\mid T) \tag{2}$$

若采用二元语法,且假设生成模型满足一阶隐马尔可夫模型,则

$$P(T)=P(t_1)\prod_{i=1}^{N}P(t_i\mid t_{i-1})$$

$$P(S\mid T)=\prod_{i=1}^{N}P(s_i\mid t_i)$$

其中,式(2)中的未知参数利用训练语料库进行最大似然估计,词位状态 T^* 利用 Viterbi 动态规划算法求解。

1.1.3　混合模型

一般而言,模型(1)对于词典词的处理可以获得较好的表现,而对于未登陆词的分词效果欠佳。对于新闻类文本该缺点尤其重要。模型(2)恰好相反,对于词典词不能很好地识别。为了加强对未登录词的识别,在最终分词时,利用了结合最大概率法和隐马尔可夫模型的混合分词方法。

1.2　汉语新闻标题特征选择方法

文本特征选择是文本聚类的重要准备,不仅可以降低计算维度,提高计算效率,而且可能由于去除了数据噪声而提高分类的准确率。考察了基于 TF-IDF(term frequency-inverse document frequency)的特征选择方法,以及基于 SVD(singular value decomposition)近似的特征选择方法。

1.2.1　基于 TF-IDF 的特征选择方法

TF-IDF 是一种用于信息搜索和信息挖掘的常用加权技术,由 Salton(1988)提出,在搜索、文献分类和其他相关领域都有广泛应用。其中词频(term frequency,TF)指的是某一个给定的词语在该文件中出现的次数,计算公式为[10]:

$$\mathrm{tf}_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{3}$$

其中,$n_{i,j}$ 是该词在文件中的出现次数,而 $\sum_k n_{k,j}$ 则是在文件中所有字词的出现次数之和。

逆向文件频率(inverse document frequency,IDF)是一个词语普遍重要性的度量。某一特定词语的 IDF 计算公式如下:

$$\mathrm{idf}_i = \log \frac{|D|}{|\{j:t_i \in d_j\}|} \tag{4}$$

其中,$|D|$ 表示语料库中的文件总数,$|\{j:t_i \in d_j\}|$ 表示包含词语的文件数目,t_i 表示第 i 个特征词,d_j 表示第 j 条文本。如果该词语不在语料库中,就会导致式(4)分母为零。因此一般情况下,使用 $1+|\{j:t_i \in d_j\}|$ 作为分母。最终,TF-IDF 权重的计算公式为:

$$w_{i,j} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i \tag{5}$$

其中,$w_{i,j}$ 表示第 i 个特征词在第 j 条文本中的 TF-IDF 权重。

TF-IDF 计算简单、易行,然而很多情况下,TF-IDF 矩阵大部分元素都为 0,过于稀疏。为了保证聚类效果,对 TF-IDF 矩阵进行奇异值分解,对求得的主成分得分矩阵进行聚类分析。

1.2.2 基于 TF-IDF 的 SVD 改进

对特征词-文档矩阵进行矩阵降维是潜在语义分析(latent semantic analysis,LSA)的基本思想,即对于传统向量空间模型的特征词-文档矩阵,应将其从稀疏的高维特征空间映射到低维潜在语义的空间,从而可能去除原始向量空间中的噪音,部分提高文本聚类精确度[11]。文中采用 TF-IDF 矩阵的奇异值分解,对潜在语义空间进行求解。

假设有一个 $M \times N$ 的特征词-文档矩阵 $\boldsymbol{A}$,其中 M 为特征词的数量,N 为文档数目,记矩阵 $\boldsymbol{A}$ 的 SVD 分解为[12]:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} \tag{6}$$

其中,$\boldsymbol{U}$:$M \times R$ 和 $\boldsymbol{V}$:$R \times N$ 是正交矩阵,$\boldsymbol{\Sigma}$:$R \times R$ 是奇异值的对角阵,$R \leqslant \min\{M,N\}$ 是 $\boldsymbol{A}$ 的秩。对 $\boldsymbol{\Sigma}$ 的对角线上的值从大到小排列,取前 $K < M$ 个不变,其他设为 0,记该矩阵为 $\boldsymbol{\Sigma}_k$。将 $\boldsymbol{\Sigma}_k$ 替换式(6)中的 $\boldsymbol{\Sigma}$,得到矩阵 $\boldsymbol{A}$ 的低秩近似矩阵。

$$\boldsymbol{A}_k = \boldsymbol{U}_k\boldsymbol{\Sigma}_k\boldsymbol{V}_k^{\mathrm{T}} \tag{7}$$

其中矩阵 $\boldsymbol{A}_k$ 的秩为 K。

SVD 可以写为如下主成分分析形式:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}} = \sum_{i=1}^{p} \boldsymbol{\sigma}_i\boldsymbol{u}_i\boldsymbol{v}_i^{\mathrm{T}} \tag{8}$$

其中,$\boldsymbol{v}_i$ 被称为 $\boldsymbol{A}$ 的成分载荷,$\boldsymbol{A}$ 在 $\boldsymbol{v}_i$ 方向上的单位线性组合 $Z_1 = \boldsymbol{A}\boldsymbol{v}_1 = \boldsymbol{\sigma}_1\boldsymbol{u}_1$ 在所有 $\boldsymbol{A}$ 的单位线性组合中样本方差最大,$z_i = \boldsymbol{\sigma}_i\boldsymbol{u}_i$ 被称为第 i 主成分得分[13]。所以,将 TF-IDF 矩阵进行奇异值分解后,再用 $\boldsymbol{u}$ 矩阵和 $\boldsymbol{d}$ 矩阵组成的对角矩阵相乘,就能得到主成分得分矩阵。在实际计算中,对式(8)截取前 k 个主成分进行分析,这与计算式(7)数学上等价。

2 K-均值聚类算法

文中使用 K-均值算法进行聚类分析。K-均值算法是硬聚类算法,是典型的基于原型的目标函数聚类方法的代表,它是数据点到原型的某种距离作为优化的目标函数,利用函数求极值的方法得到迭代运算的调整规则。K-均值算法以欧氏距离作为相似度测度,它是求对应某一初始聚类中心向量 $\boldsymbol{V}$ 的最优分类,使得评价指标 J 最小。算法采用误差平方和准则函数作为聚类准则函数,即各类的聚类平方和最小[14]:

$$J = \sum_{k=1}^{k} \sum_{i=1}^{n} \| x_i - u_k \|^2 \tag{9}$$

算法过程如下:

(1)选取数据空间中的 K 个对象作为初始中心,每个对象代表一个聚类中心,可以选择前 s 个主成分做聚类,假设它们在 s 维的坐标为 $\alpha_i = (\alpha_{i1},\alpha_{i2},\cdots,\alpha_{is})$。

(2)对于样本中的数据对象,根据它们与这些聚类中心的欧氏距离,按距离最近的准则将它们分到距离它们最近的聚类中心(最相似)所对应的类。若选取的数据点坐标为 $(x_1,x_2,\cdots,x_s)$,其与第 i 个初始中心点的欧氏距离的计算公式为:

$$d = \sqrt{\sum_{k=1}^{s} (x_k - \alpha_{ik})^2} \tag{10}$$

(3)更新聚类中心:将每个类别中所有对象对应的均值作为该类别的聚类中心,计算目标函数的值。

(4)迭代步骤 2 ~ 步骤 3,直至新的质心与原质心相等或小于指定阈值,算法结束[15]。

3 数据介绍

文中从新浪新闻网爬取了财经、法治、国际、军事和社会共五类 255 条新闻标题,并进行了人为分类标识。这五类新闻数量不等,每类新闻的条数和平均字数如图 1 和图 2 所示。

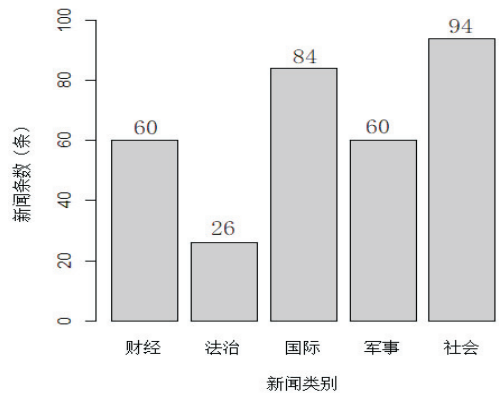

图 1 五类新闻条数

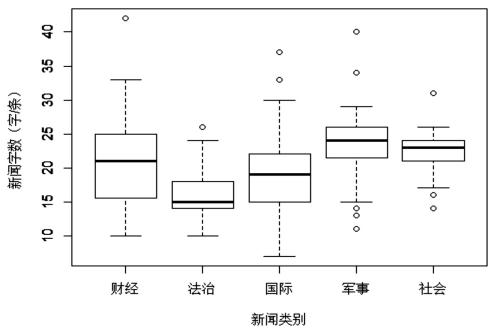

图 2　五类新闻标题字数

图 2 是五类新闻标题字数的箱型图。可以发现,这五类新闻中,法治类新闻的平均字数最少,军事类的平均字数最多,但总的来说差异不大,都在 15 到 20 字之间,符合对汉语新闻短文本研究的目标。财经类和国际类新闻字数的变化区间较大,与它们的新闻类型有关;社会类新闻的字数变化最小。对五种新闻标题做分词处理,画出词频在前 10 的直方图,如图 3 所示。

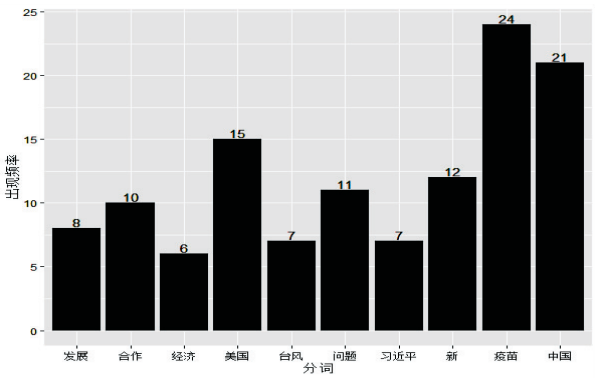

图 3　词频直方图

4　聚类结果

4.1　特征选择结果

聚类处理过程中,为了能够更加有效地节省存储的空间及提高检索的效率,文中使用哈工大停用词表删去常用停词。鉴于分析对象的是新闻标题,没有强烈的情感倾向,未特殊保留情绪功能词。根据 1.1.3 节对文本进行自动分词并去停词后,得到了 1 510 个词条,即 TF-IDF 矩阵 $\boldsymbol{A}$ 的维度为:251×1 510。现利用式(6)对矩阵 $\boldsymbol{A}$ 进行 SVD 分解,并进行主成分分析。

选取式(8)的前三个主成分,通过观察数据发现第二主成分的数值大部分趋于 0,只有 1 条标题的第二主成分值为-15,2 条在-5 左右,而这 3 条异常值都属于军事新闻,内容如下:“薛晓峰出席第十四届澳门青年学生军事夏令营暨第二届澳门大学生军事生活体验营开营典礼”、“特朗普回国后翻脸不认人 俄军用两大军事动作回应”、“百余名学子走进酒泉卫星发射基地体验航天梦”。可以看出这三条新闻的字数、特征词都有较大差异,共同点仅仅是它们同属于军事新闻,所以无法判断抛开极端值后,第二主成分主要包含的是文本的哪些信息。而第一主成分在前四类新闻中都不存在较大差异,可能是点过于密集的原因,但却把第五类社会类新闻与前四类区分开了,说明第一主成分主要包含的是文本的语义信息。图 4 和图 5 分别是去掉三个离群值后第一主成分和第二主成分的关系以及前三个主成分的三维图:

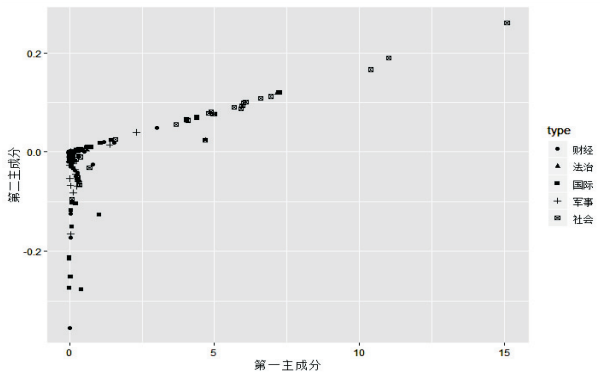

图 4　五类新闻标题前两个主成分关系

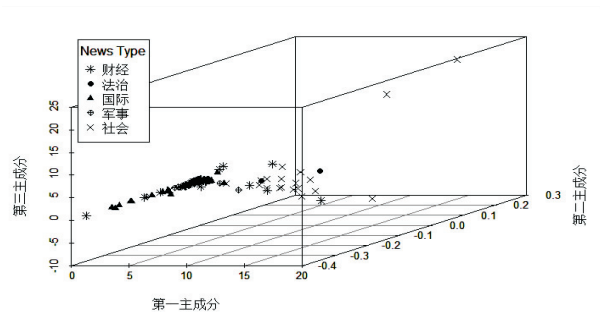

图 5　五类新闻标题前三个主成分的关系

4.2　聚类结果

选定前 15 个主成分,并将 k 均值聚类中的 K 值设为 5 进行聚类,可以把聚类后每一类的词云图与原始分类下的词云图进行比较,看它们在主要内容上有没有发生变化。以社会类新闻为例,对比结果见图 6。

(a)聚类后的社会类新闻标题

(b)初始的社会类新闻标题

图 6　聚类与手动分类结果词云图对比

4.3 对聚类效果的评价

通过计算聚类的平均准确率,可以对聚类效果有一个大致的认识,聚类的平均准确率采用文献[2]的定义,具体计算公式如下:

平均准确率:

$$AA = \frac{PA + NA}{2}$$

积极准确率:

$$PA = \frac{a}{a + c}$$

消极准确率:

$$NA = \frac{d}{b + d}$$

其中,a、b、c、d 的取值如表 1 所示。

表 1 准确率的评价指标

自动聚类中属于同一类	手工分类中属于同一类	标识
是	是	a
是	否	b
否	是	c
否	否	d

可以列出取 15 个主成分时的每种新闻聚类准确率,如表 2 所示。

表 2 不同主成分选取个数的聚类准确率 %

新闻类别	积极准确率	消极准确率	平均准确率
财经	100.0	99.0	99.5
法治	90.5	86.1	88.3
国际	72.5	84.0	78.3
军事	65.9	89.0	77.5
社会	68.1	100.0	84.1

在聚类的主成分个数为 15 的情况下,k 均值聚类的平均准确率为 85.5%,其中积极准确率的均值为 79.4%,消极准确率均值约为 91.6%,显著高于积极准确率。

5 结束语

文中旨在通过对词条矩阵进行奇异值分解来降低矩阵的维度,达到更精确的分类目的。通过对新闻标题的聚类可以发现,聚类的结果并不是非常理想,认为可能有如下几个原因:由于新闻的时效性和多样性,不能从标题中提取出极具代表性的特征词;样本量较少,每条标题的字数过少;聚类方法不够好,可以尝试采用基于语义而非特征词的聚类方法。

在对文本聚类的研究过程中也进行了难点总结:首先,在将文本转换为词条矩阵时,若是保留去掉停词之后的所有词频大于等于 1 的词语,会使最后进行计算的矩阵过大,导致分类效率过低,而如果只提取词频较大的词语,则会损失一些信息,也会导致分类结果变差;其次,在生成词条矩阵时,已有的算法包默认只保留大于等于三个字的字符,这对处理英文文本比较实用,但处理汉语文本时会损失文本信息。

参考文献:

[1] FURNAS G W, DEERWESTER S, DUMAIS S T, et al. Information retrieval using singular value decomposition model of latent semantic structure[C]//Proceedings of SIGIR′88. [s.l.]:ACM,1988.

[2] 姜 宁,史忠植. 文本聚类中的贝叶斯后验模型选择方法[J]. 计算机研究与发展,2002,39(5):580-587.

[3] 徐建锁,王正欧. 基于 LSI 和自组织神经网络的高效文本聚类方法[J]. 天津大学学报,2004,37(11):1026-1030.

[4] 王国勇,徐建锁. TCBLSA:一种中文文本聚类新方法[J]. 计算机工程,2004,30(5):21-22.

[5] 姚清耘,刘功申,李 翔. 基于向量空间模型的文本聚类算法[J]. 计算机工程,2008,34(18):39-41.

[6] 宋 涛,施水才,房 祥,等. 基于改进的潜在语义分析的文本聚类[J]. 北京信息科技大学学报:自然科学版,2012,27(3):21-25.

[7] 黄昌宁,赵 海. 由字构词——中文分词新方法[C]//中国中文信息学会二十五周年学术会议. 北京:中国中文信息学会,2007:53-63.

[8] 刘 源,谭 强,沈旭昆. 信息处理用现代汉语分词规范及自动分词方法[M]. 北京:清华大学出版社,1994.

[9] 宗成庆. 统计自然语言处理[M]. 第 2 版. 北京:清华大学出版社,2008.

[10] 吴金学. 基于概率潜在语义分析的文本聚类研究[J]. 青岛理工大学学报,2008,29(2):95-99.

[11] DEERWESTER S, DUMAIS S T, FURNAS G, et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science,1990,41(5):391-407.

[12] 范云鹏. 矩阵低秩逼近在图像压缩中的应用[D]. 西安:西安电子科技大学,2012.

[13] CHUI C K, KAO Ben, HUNG E. Mining frequent itemsets from uncertain data[C]//Proceedings of the 11th Pacific-Asia conference on advances in know-ledge discovery and data mining. Nanjing:Springer,2007:47-58.

[14] ALAHAKOON D, HALGAMUGE S K. Dynamic self-organizing maps with controlled growth for knowledge discovery[J]. IEEE Transactions on Neural Networks, 2000, 11(3):601-614.

[15] 高 岭,申 元,高 妮,等. 基于文本挖掘的漏洞信息聚类分析[J]. 东南大学学报:自然科学版,2015,45(5):845-850.