

## 基于短语的维吾尔文文本分类

阿力木江·艾沙<sup>1,2\*</sup>, 吐尔根·依布拉克<sup>2</sup>, 库尔班·吾布力<sup>2</sup>, 李哲<sup>1</sup>

(1. 新疆大学 现代教育技术中心, 乌鲁木齐 830046; 2. 新疆大学 信息科学与工程学院, 乌鲁木齐 830046)

(\* 通信作者电子邮箱 alim@xju.edu.cn)

**摘要:** 文本特征表示是在文本自动分类中最重要的一环。在基于向量空间模型(VSM)的文本表示中特征单元粒度的选择直接影响到文本分类的效果。对于基于词袋模型(BOW)的维吾尔文文本分类效果不理想的问题,提出了一种基于统计方法的维吾尔语短语抽取算法并将抽取到的短语作为文本特征项,采用支持向量机(SVM)算法对维吾尔文文本进行了分类实验。实验结果表明,与以词为特征的文本分类相比,短语作为文本特征能够提高维吾尔文文本分类的准确率和召回率。

**关键词:** 文本分类; 短语抽取; 支持向量机; 维吾尔语; 互信息

**中图分类号:** TP391.1 **文献标志码:** A

### Phrase based Uyghur language text categorization

ALIMJAN Aysa<sup>1,2\*</sup>, TURGUN Ibrahim<sup>2</sup>, KURBAN Obul<sup>2</sup>, LI Zhe<sup>1</sup>

(1. Modern Education Technology Center, Xinjiang University, Urumqi Xinjiang 830046, China;

2. School of Information Science and Engineering, Xinjiang University, Urumqi Xinjiang 830046, China)

**Abstract:** Text representation is the most important phase in automatic text categorization. In the Vector Space Model (VSM) based text representation, the selection of feature granularity has the direct impact on the text categorization performance. The statistical approach based Uyghur phrase extraction algorithm was proposed and the Uyghur text categorization experiments was conducted using Support Vector Machine (SVM) algorithm based on the extracted phrases as text features. The experimental results show that the phrase based Uyghur text categorization achieves higher classification precision and recall compared to the word based categorization.

**Key words:** text categorization; phrase extraction; Support Vector Machine (SVM); Uyghur language; Mutual Information (MI)

## 0 引言

在文本分类过程中,文本的表示模型是一个既基本又重要的问题。只有先将文本从无结构或者半结构化的原始形式转化为计算机能够理解的表示模型后,计算机才能对文本内容进行分析与处理。向量空间模型(Vector Space Model, VSM)仍是文本特征表示的主要方法,相关研究仍然集中在以什么特征单元(词,短语,n-Gram)作为特征项这个问题上<sup>[1]</sup>。大部分系统仍是单词作为特征项,也就是基于词袋模型(Bag Of Words, BOW)的。虽然BOW具有直观且易于实现等优点,但是存在一个很大的缺陷,就是它没有考虑文本上下文间的语义关系和潜在的概念结构,特征项之间是独立的,不能充分反映出文本总体面貌。正是因为从根本上难以克服基于词的BOW的先天缺陷,基于其之上的很多分类算法准确率都不是很高。最基本最有效的改进应该是从向量空间模型的文本表示方法入手,选择文本表达能力较强的特征单元作为文本特征项,以提高对文本的表达能力<sup>[2]</sup>。特征单元的选择是文本向量的基础,特征单元不同则特征空间不同,其中文本向量的分布也会完全不同。可以说特征单元的选择从根本上影响着整个文本分类的效果。

对中英文文本分类中的特征单元的选择,国内外研究人

员做了大量的研究工作。在文献[3]中,作者指出短语是指在文本中连续出现的具有句法意义的或统计意义的多个词并提出了一种基于统计方法的英文短语抽取算法,通过分类实验验证了短语作为文本特征的有效性。文献[4]对英文文本分类中的特征单元粒度选择进行了研究,指出短语作为文本特征有助于提高英文文本分类的效率。在文献[5]中,作者将汉字和汉语单词作为特征项,在三种不同的中文语料库上进行了分类实验,指出直接使用汉字进行分类,也可以得到和单词一样的分类效果。文献[2]指出,短语特征更有利于表达中文文本的内容特征。此观点在一些应用系统中也得到了证实<sup>[6]</sup>。

我们前期的研究工作<sup>[7]</sup>显示,在中英文文本分类中表现良好的基于词的VSM表示模型(也就是词袋模型,BOW)对维吾尔文来讲效果并不好。在中英文中的一个单词在维吾尔文中不一定就是一个单词,而可能是一个短语。例如,单词“软件(Software)”在维文中是“Yumxaq detal”,是由两个词组成的短语(词组)。对于一个基于BOW表示的、出现“Yumxaq(软的)”和“detal(组件)”的文本,很难判断其所属类别。但是,对出现短语“Yumxaq detal”的文本,我们很容易判断它是关于计算机方面的文本。因此,如果在“Yumxaq detal”中的这两个词分开,被看成独立的两个特征项,那么它们的类别区分能力就

收稿日期:2012-05-02;修回日期:2010-06-08。 基金项目:国家自然科学基金资助项目(61063026;61163028)。

作者简介:阿力木江·艾沙(1973-),男(维吾尔族),新疆喀什人,副教授,博士研究生,主要研究方向:自然语言处理、信息安全; 吐尔根·依布拉克(1958-),男(维吾尔族),新疆乌鲁木齐人,教授,博士生导师,主要研究方向:信息处理技术、人工智能; 库尔班·吾布力(1974-),男(维吾尔族),新疆喀什人,副教授,主要研究方向:模式识别; 李哲(1977-),女,新疆乌鲁木齐人,讲师,主要研究方向:软件工程。

会下降,直接影响分类效果。在维吾尔语中这种在中英文中的一个单词对应于维文中的一个短语的情况是非常常见的。因此,在维吾尔语中如果只把单词作为特征项,很难达到和中文一样的分类效果。本文研究了短语作为文本特征项对维吾尔语文本分类效果的影响。要用短语作为特征项,问题的关键是识别出一个个类别区分能力强的短语。

在维吾尔语中,由两个字组成的短语非常常见,而且有结构稳定、语义完整、统计意义较强等特点,更有利于表达文本内容特征。本文考虑到维吾尔语的以上特点,提出了维吾尔语短语抽取算法,抽取类别区分能力较强的短语并通过实验验证了本文算法抽取的短语作为文本特征能够有效提高维吾尔语文本分类的准确率和召回率。

## 1 维吾尔语短语抽取算法

本文提出了基于互信息的维吾尔语短语抽取算法。首先从每个类别中选取区分度较高的预定数量的单词,对每个类别提取的单词进行合并构造单词集合  $U$ 。然后根据  $U$  和训练集中的每个文本构造候选短语集合  $P$ 。最后根据每个候选短语和每个类别之间的互信息量选择类别区分能力较高的短语集合  $B$ 。对于那些作为短语被选进来的噪声信息如,“Xundaq qilip (于是)”,“Omumen äytqanda (总之)”等,本文根据互信息的定义以及维吾尔语短语具有结构稳定、统计意义较强等特点,采用如下方法进行过滤:要保证被选短语和某类别间的互信息量比起构成此短语的两个单词分别和该类别的互信息量都要大。

### 1.1 符号及相关定义

方便起见,定义如下符号:

$D$ : 包含  $N_d$  个文本的维吾尔语分类训练语料库。

$W$ :  $D$  中所有单词集合。

$C$ : 包含  $N_c$  个类别的类别集合。

$N_w$ : 单词阈值(预定的单词数)。

$N_p$ : 短语阈值(预定的短语数)。

$|A|$ : 集合  $A$  中元素个数。

$MI(t, c_i)$ :  $t$  和  $c_i$  之间的互信息值。

互信息<sup>[8]</sup> 根据特征和类别共同出现的概率,度量特征和类别的相关性。特征  $t$  和类别  $c_i$  互信息值计算公式如下:

$$MI(t, c_i) = \log \frac{p(t, c_i)}{p(t) \times p(c_i)} = \log \frac{p(t|c_i)}{p(t)}$$

其中:  $p(t, c_i)$  表示训练集中既包含特征  $t$  又属于类别  $c_i$  的文本出现的概率,  $p(t)$  表示包含特征  $t$  的文本在训练集中出现的概率,  $p(c_i)$  表示训练集中属于类别  $c_i$  的文本的概率。特征  $t$  在类别  $c_i$  中出现概率高,而在其他类别中出现概率低,即特征  $t$  和类别  $c_i$  相关性大,将获得较高的互信息值  $MI(t, c_i)$ 。反之,将获得较低的互信息值  $MI(t, c_i)$ 。

### 1.2 算法

输入:  $D, W, C, N_w, N_p$ 。

输出: 具有高类别区分能力的短语集合  $B$ 。

For all  $c_i \in C$  do

For all  $w_j \in W$  do

计算  $MI(w_j, c_i)$ , 并按降序排序

$U_i \leftarrow$  前  $N_w$  个互信息值较高的单词

$U \leftarrow \bigcup_{i=1}^{N_c} U_i$  // 具有高区分度的单词集合

For all  $d_i \in D$  do

将在  $d_i$  中出现并属于  $U$  的单词依次加入到  $LU_i$  中

$P_i = \text{Null}$

For  $j = 2$  to  $|LU_i|$  do

将  $b_j = (LU_i[j-1], LU_i[j])$  作为第  $j$  个候选短语加入到  $P_i$ ,  
即  $P_i \leftarrow P_i \cup b_j$

$N_d$

$P \leftarrow \bigcup_{i=1}^{N_d} P_i$

// 候选短语集合

For all  $c_i \in C$  do

For all  $b_j \in P$  do

计算  $MI(b_j, c_i)$ , 并按降序排序

$F_i \leftarrow$  互信息值非 0 的候选短语

$LP_i = \text{Null}$

For  $j = 1$  to  $|F_i|$  do

对于  $F_i$  中第  $j$  个候选短语  $b_j = (w_{j1}, w_{j2})$

If  $MI(b_j, c_i) > \max(MI(w_{j1}, c_i), MI(w_{j2}, c_i))$  then

将  $b_j$  加入到  $LP_i$

$B_i \leftarrow$  选  $LP_i$  中前  $N_p$  个短语

$N_c$

$B \leftarrow \bigcup_{i=1}^{N_c} B_i$

// 具有高类别区分能力的短语集合

## 2 SVM 分类算法

Vapnik 提出的支持向量机 (Support Vector Machine, SVM) 理论<sup>[9]</sup> 是一种基于统计学习理论的二值分类学习方法,因具有较好的泛化性能而受重视。文献[10]最先用 SVM 算法做文本分类,并同其他机器学习算法比较,发现 SVM 泛化性能好,能处理很高维的分类问题,并且无需进行特征选择。SVM 算法基本思想是寻找一个能够对样本进行正确分类的最优分类面,所谓最优分类面就是要求分类面不但能将两类样本点无错误地分开,而且要使两类的分类空隙最大。 $d$  维空间中线性判别函数的一般形式为  $g(x) = w^T x + b$ , 分类面方程是  $w^T x + b = 0$ , 将判别函数进行归一化,使两类所有样本都满足  $|g(x)| \geq 1$ , 此时离分类面最近的样本  $|g(x)| = 1$ , 而要求分类面对所有样本都能正确分类,就是要求它满足:

$$y_i(w^T x_i + b) - 1 \geq 0 \quad (1)$$

其中  $i = 1, 2, \dots, n$ 。

式(1)中使等号成立的那些样本叫作支持向量(Support Vector)。两类样本的分类空隙(Margin)的间隔大小:

$$\text{Margin} = 2 / \|w\| \quad (2)$$

因此,最优分类面问题可以表示成如下的约束优化问题,即在条件(1)的约束下,求函数:

$$\varphi(w) = \frac{1}{2} \|w\|^2 = \frac{1}{2} (w^T w) \quad (3)$$

的最小值。为此,可以定义如下的 Lagrange 函数:

$$L(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^n \alpha_i [y_i (w^T x_i + b) - 1] \quad (4)$$

其中,  $\alpha_i \geq 0$  为 Lagrange 系数,我们的问题是对  $w$  和  $b$  求 Lagrange 函数的最小值。把式(4)分别对  $w, b, \alpha_i$  求偏微分并令它们等于 0,得:

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^n \alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^n \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \alpha_i} = 0 \Rightarrow \alpha_i [y_i (w^T x_i + b) - 1] = 0$$

以上三式加上原约束条件可以把原问题转化为如下凸二次规划的对偶问题:

$$\max \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (x_i^T x_j)$$

s. t.  $\alpha_i \geq 0, i = 1, \dots, n$

$$\sum_{i=1}^n \alpha_i y_i = 0 \quad (5)$$

这是一个不等式约束下二次函数机制问题,存在唯一最优解。若  $\alpha_i^*$  为最优解,则:

$$w^* = \sum_{i=1}^n \alpha_i^* y_i x_i \quad (6)$$

$\alpha_i^*$  不为零的样本即为支持向量,因此,最优分类面的权重向量是支持向量的线性组合。

$b^*$  可由约束条件  $\alpha_i [y_i (w^T x_i + b) - 1] = 0$  求解,由此求得的最优分类函数是:

$$f(x) = \text{sgn}((w^*)^T x + b^*) = \text{sgn}\left(\sum_{i=1}^n \alpha_i^* y_i x_i^T x + b^*\right) \quad (7)$$

其中  $\text{sgn}()$  为符号函数。

上面介绍的是线性可分的二值分类器,对于线性不可分的问题,通过非线性变换将输入空间映射到一个高维特征空间,然后在这个新空间中求取最优分类面,而该非线性变换是通过定义适当的核函数来实现的<sup>[11]</sup>。而基于 SVM 的多值分类器的构造是可以通过组合多个二值分类器来实现,具体的构造方法有 one-versus-one 和 one-versus-rest 两种<sup>[12]</sup>。

### 3 实验及分析

#### 3.1 数据集

对于中英文的文本分类研究,国内外已经有相对标准的、开放的文本分类语料库,这样就可以在共同的文本集上比较不同的特征表示和分类方法的性能。而就维吾尔文文本分类而言,目前还没有标准、开放的分类型文本集可供使用。所以,

从人民网维吾尔文版(<http://uyghur.people.com.cn/>)和天山网(<http://www.xjtsnews.com/>)等主流维吾尔文网站上收集了 2467 篇文本,通过人工方式将其分为政治、经济、体育、旅游、教育、文化共 6 个类。训练集包括 1660 篇文本,测试集包括 807 篇文本,具体分布情况如表 1 所示。

表 1 分类语料库

类别	训练文本数	测试文本数
政治	252	126
经济	267	114
体育	283	146
旅游	287	127
教育	310	151
文化	261	143

#### 3.2 短语抽取

采用本文算法,从维吾尔文训练语料中识别出类别区分能力较高的短语。

首先对数据集进行预处理,即把数据集中的所有的文本转换成 UTF-8 编码格式,对文本中的标点符号、数字、非维吾尔语字符以及停用词进行过滤,识别出一个个维吾尔文单词,建立了原始特征项集合。通过实验确定  $N_w$  为 2000,  $N_p$  为 250。每个类别提取 2000 个词,把 6 个类别提取的高区分度单词合并后发现单词总数并不是  $6 \times 2000 = 12000$ ,而是 7726。这说明各类别中重复的单词较多。而从 6 个类别中抽取的短语数总共为 1457,接近于  $6 \times 250 = 1500$ ,说明各类别中抽取的短语几乎不重复,抽取到的短语类别区分能力较强。表 2 给出了本文算法抽取的部分短语。

表 2 抽取到的部分维吾尔文短语

类别	维吾尔文短语
政治	Dölat ra'isi, Merkezî komit, Helq qurultiyi, Da'imi komit, Diplomatic munasiwet, Diplomatiye ministiri, Amërika zongtungi, Partiye texklatliri, Islahat ëchirëtix, Partiye quruluxi, Siyasi qurulux, Ilmi terekkiyat
经济	Helq puli, Amërika dolliri, Pul muamile, Import, eksport, Helq bankisi, Omumi sommisi, ëxix süriti, Pul pahalliqi, Mal bahasi, Iqtisadi qurulux, Ichki ihtiyaj, Iqtisadi hemkarliq, Paychik baziri, Taxqi përiwut, Taxqi soda, Meblegh sëlax
体育	Putbol musabiqisi, Tenherket yighini, Olimpik Tenherket, Ayal tenherketchi, Asiya longqisi, Ammiwi tenterbiye, Dölet komandisi, Tenterbiye ixliri, Tenterbiye hizmiti, Tenterbiye musabiqisi, Tenherket musabiqisi
旅游	Menzire rayuni, Sayahet noqtisi, Sayahet yëtekchisi, Qanas köli, Boghda köli, A'ile sayahetchiligi, ëkologiyilik sayahet, Dëhqanlar xatliqi, Helqaraliq sayahet, Sayahat lëniyisi, Sayahet kopiratipi
教育	Maarip idarisi, Ottura mektep, Sinip mes'uli, Ali mektep, Ali Maarip, Ottura tëhnikom, Oqutquchilar qoxuni, Oqutquchi oqughuchi, Oqutux usuli, Sapa maaripi, Baxlanghuq mektep, Qoxtil ma'aripi, Oqutux matëriyali
文化	Medeniyet quruluxi, Oghlaq tartixix, Usul ansambili, Usul medeniyiti, Unwërsal sen'et, Uyghur muqamliri, Uyghur xi'ëriyiti, Edebiyat sen'et, Medeni miras, Darwazliq san'iti, Muqam san'iti, Helq sen'etkari

#### 3.3 分类实验

本文采用 SVM 算法进行分类实验。SVM 分类器采用支持向量机算法库 LIBSVM 的 Java 包<sup>[13]</sup>来构造并采用 one-versus-rest 方法来实现多值分类,采用的核函数是线性核函数。实验是在配置为 Pentium(R) dual-core CPU 2.10 GHz 处理器、2 GB 内存,操作系统为 Windows 7 的 PC 机上进行的。微软的 Windows 7 操作系统已经全面支持维吾尔语。分类性能评价采用常用的评价指标<sup>[14]</sup>准确率(Precision)、召回率(Recall)和 F1 值等。

$P(\text{准确率}) = \text{分类正确的文本数} / \text{实际分类的文本数}$

$R(\text{召回率}) = \text{分类正确的文本数} / \text{应有的文本数}$

$F1 = 2PR / (P + R)$

分别进行两个实验,在实验中用归一化的 TF-IDF 公

式<sup>[15]</sup>来计算特征项权重。

实验 1 用 7726 个单词作为特征项集合。首先在单词空间中对训练集和测试集中的文本进行向量化,在训练集上训练 SVM 分类器,构造分类模型。然后对测试文本进行了分类实验。

实验 2 用 7726 个单词和 1457 个短语共同作为特征项集合,首先在单词+短语空间中对训练集和测试集中的文本进行向量化,在训练集上训练 SVM 分类器,构造分类模型。然后对测试文本进行了分类实验。实验结果如表 3 所示。

从表 3 中可以看出,单词和短语共同作为文本特征, SVM 分类器在准确率、召回率和 F1 值上均有明显提高。与以单词为特征的文本分类相比,单词和短语共同作为文本特征,分类器在各类别上的 F1 值提高幅度为 6.07% ~ 13.88%,有了明



显的改进。其中“旅游”这一类提高幅度最大,达到了13.88%，“教育”类提高幅度最低,有6.07%。分析原因是因为很多旅游相关名称如“Boghda köli(天池)”、“Dêhقانlar xatliqi(农家乐)”、“Qanas köli(喀纳斯湖)”、“Menzire noqtisi(景点)”等都是两个词组成的短语,如果将两个词分开处理会降低其类别区分能力。只有两个词组合在一起,它就携带较高的类别信息量,进而对分类的贡献也就比较大。而在教育类中出现频率较高的“Ali mektap(大学)”,“Oqutquchi oqughuchi(师生)”,“Oqutux matëriyali(教材)”,“Oqutqushilar qoxuni(师资)”等各短语中至少有一个单词本来就携带较高的类别信息量,因此对该类而言,短语作为特征项分类的 $F1$ 值提高幅度比起其他类别要低一点。

表3 分类实验结果对比

类别	单词			单词+短语		
	$P$	$R$	$F1$ 值	$P$	$R$	$F1$ 值
政治	0.7170	0.7600	0.7379	0.9286	0.7800	0.8478
经济	0.6630	0.8133	0.7305	0.7528	0.8933	0.8171
体育	0.7547	0.8000	0.7767	0.9245	0.7656	0.8376
旅游	0.7037	0.7600	0.7308	0.8000	0.9524	0.8696
教育	0.7969	0.9444	0.8644	0.8831	0.9714	0.9251
文化	0.7692	0.8000	0.7843	0.8065	0.9524	0.8734

#### 4 结语

文本分类的进一步改进除了算法方面,应该还立足于影响文本分类最底层、最根本的因素:文本表示中的特征项的选择,提高特征项的完整独立程度。在维吾尔语中短语具有较强文本表示功能,在表示文本时,能将文本的内容特征(如主题类别)鲜明地表示出来。本文提出了基于统计方法的维吾尔语短语抽取算法,将抽取到的短语作为文本特征进行了分类实验。实验结果表明,短语作为文本特征明显提高维吾尔文文本分类的准确率和召回率。

##### 参考文献:

- [1] 苏金树,张博峰,徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报,2006,17(9):1848-1859.
- [2] 刘华. 基于关键短语的文本分类研究[J]. 中文信息学报,2007,

21(4):34-41.

- [3] CAROPRESO M F, MATWIN S, SEBASTIANI F. Statistical phrases in automated text categorization, Statistical Phrases in Automated Text Categorization [R]. Paris: Centre National de la Recherche Scientifique, 2000: 78-102.
- [4] KOSTER C, BENEY J. Phrase-based document categorization revisited[C]// PaIR'09: Proceedings of the 2nd International Workshop on Patent Information Retrieval. New York: ACM, 2009: 49-55.
- [5] 张爱华,荆继武,向继. 中文文本分类中的文本表示因素比较[J]. 中国科学院研究生院学报,2009,26(3):400-407.
- [6] 李钝,曹付元,曹元大. 基于短语的文本情感分类研究[J]. 计算机科学,2008,35(4):132-134.
- [7] 阿力木江·艾沙,吐尔根·依布拉音,艾山·吾买尔. 基于机器学习的维吾尔文文本分类研究[J]. 计算机工程与应用,2012,48(5):110-112.
- [8] 张震,胡学钢. 基于互信息量的分类模型[J]. 计算机应用,2011,36(6):1678-1680.
- [9] VAPNIK V. The nature of statistical learning theory[M]. New York: Springer-Verlag, 1995.
- [10] JOACHIMS T. Text categorization with support vector machines: Learning with many relevant features[C]// European Conference on Machine Learning. Berlin: Springer-Verlag, 1998: 137-142.
- [11] 孙建涛,郭崇慧,陆玉昌,等. 多项式核支持向量机文本分类器泛化性能分析[J]. 计算机研究与发展,2004,41(8):1321-1326.
- [12] HSU C-W, LIN C-J. A comparison of methods for multiclass support vector machines[J]. IEEE Transactions on Neural Networks, 2002, 13(2):415-440.
- [13] CHANG C-C, LIN C-J. LIBSVM: A library for support vector machines [EB/OL]. [2011-09-10] <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.
- [14] SEBASTIANI F. Machine learning in automated text categorization[J]. ACM Computing Surveys, 2002,34(1):1-47.
- [15] 施聪莺,徐朝军,杨晓江. TFIDF 算法研究综述[J]. 计算机应用,2009,29(S1):167-170.

(上接第2922页)

文下一阶段的研究重点。

#### 4 结语

本文探讨传统分词方法导致的词语义不完整性问题,并提出了一种基于频繁模式挖掘的维吾尔文组词的新思路,并将无先验知识的频繁模式挖掘问题转化为特定模式的匹配问题,设计出了一种简单而高效的智能组词算法。该算法仅需对文本进行一次扫描,也不需产生候选模式集,就挖掘出能作为语义词的完整频繁模式集。实验结果证明本算法的正确性和实用性,本算法也可以直接引用到同语系的哈萨克文和柯尔克孜文的组词中,具有一定的推广意义。

##### 参考文献:

- [1] TOHTI T, HAMDULLA A, MUSAJAN W. Research on Web text representation and the similarity based on improved VSM in Uyghur Web information retrieval[C]// CCPR 2010: Chinese Conference on Pattern Recognition. Chongqing: [s. n.], 2010: 984-988.
- [2] 阿力木江·艾沙,吐尔根·依布拉音,艾山·吾买尔,等. 基于机器学习的维吾尔文文本分类研究[J]. 计算机工程与应用,2011,48(5):110-112.

- [3] 刘晓涛,郭福亮. 一种有趣关联模式挖掘方法[J]. 计算机工程,2010,36(11):36-38.
- [4] 朱琼,施荣华. 一种数据流中的频繁模式挖掘算法[J]. 计算机应用,2008,28(6):1463-1466.
- [5] 刘兵. Web 数据挖掘[M]. 1版. 北京:清华大学出版社,2009.
- [6] 宋余庆,朱玉全,孙志挥,等. 一种基于频繁模式树的约束最大频繁项目集挖掘及其更新算法[J]. 计算机研究与发展,2005,17(5):777-783.
- [7] 张锦,马海兵,胡运发. 一种基于 FP-Tree 的频繁模式挖掘自适应算法[J]. 模式识别与人工智能,2005,18(6):763-768.
- [8] 李也白,唐辉,张淳,等. 基于改进的 FP-tree 的频繁模式挖掘算法[J]. 计算机应用,2011,31(1):101-103.
- [9] 敖富江,杜静,陈彬,等. 一种基于混合搜索的高效 ToP-K 最频繁模式挖掘算法[J]. 国防科技大学学报,2009,31(2):90-93.
- [10] 马青霞,李广水,孙梅. 频繁模式挖掘进展及典型应用[J]. 计算机工程与应用,2011,47(15):138-144.
- [11] 花红娟,张健,陈少华. 基于频繁模式树的约束最大频繁项集挖掘算法[J]. 计算机工程,2011,37(9):78-80.
- [12] 肖波,徐前方,蔺志青,等. 可信关联规则及其基于极大团的挖掘算法[J]. 软件学报,2008,19(10):2597-2610.