基于多策略的维吾尔文网页识别方法*

阿力木·木拉提[1,2,3]，艾孜尔古丽[4]，杨雅婷[1,2]，李晓[1,2]

（1. 中国科学院新疆理化技术研究所，乌鲁木齐 830011；2. 新疆民族语音语言信息处理重点实验室，乌鲁木齐 830011；3. 中国科学院大学， 北京 100049；4. 新疆师范大学计算机科学技术学院， 乌鲁木齐 830054）

摘要： 经过对大量维吾尔文网站的调查与分析，本文从多语种混合网页中针对维吾尔文网页识别进行了研究。这对维吾尔语信息处理工作起着关键的作用。首先本文探讨了维吾尔文不规范网页的字符编码转换规则及原理，以此对不规范维吾尔文字符进行了相应的处理。之后介绍了基于修改的 N-Gram 方法和基于维吾尔语常用词特征向量的两种方法，其中后者融合了维吾尔文常用候选词语料库及向量空间模型（Vector Space Model）。使用三种不同类型的维吾尔文网页文本作为本研究的数据集，在此基础上验证了本文提出的网页识别方法，以及采用不同的方法进行了网页识别的实验。实验结果表明，基于 N-Gram 的方法对正文较长的新闻或论坛网页的识别性能最佳，反而基于常用词特征向量的方法对短文本的网页识别性能优越 N-Gram。所提方法对维吾尔文网页识别的整体性能达到 90%以上，并验证了这两种方法的有效性。

关键词：维吾尔文；网页识别；N-Gram 方法；常用词；向量空间模型

中图分类号：TP391　　**文献标识码：**A

An Approach to Uyghur Webpage Recognition Based on Multi-strategy

Alim Murat[1, 2, 3], Azragul[4], Yating Yang[1, 2] and Xiao Li[1, 2]

(1.Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Science ,Urumqi 830011, China ; 2. Xinjiang Key Laboratory of Minority Speech and Language Information Processing, Urumqi 830011, China; 3. University of Chinese Academy of Science, Beijing 100049, China; 4. Xinjiang Normal University, Urumqi 830054, China)

Abstract: this paper studied the web-page identification task for Uyghur, after the investigation and analysis on a large number of Uyghur websites. In the first, paper discussed the character encoding conversion rules and principles over non-standard Uyghur characters, stored on the web and conducted corresponding treatment to the issue. Then, two major approach to task is presented, one is the modified N-Gram method (MNG) and the other is that a feature vector of Uyghur common word, fusing Uyghur common word corpus with space vector model. The datasets constitute of three different types of Uyghur web-page documents are used, in order to validate the proposed method in this task. The experiment results showed that N-Gram based approach substantially performed well in identifying web-pages of a longer text in news site and forum. Instead, a feature vector of Uyghur common word approach performed superiorly in web-pages of short text and outperform the N-Gram based approach. Overall, the proposed method gained more than 90% F1 score in the task.

Key words: Uyghur; Webpage Recognition; N-Gram Method; Common Word; Vector Space Model

1 引言

互联网的迅速发展带动了新疆的少数民族语言信息处理技术的发展，使用互联网进行查找信息的少数民族用户对各类信息的需求日益增加，同样各个少数民族语言文字的网页数量

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金（61662081）;新疆维吾尔自治区青年科技创新人才培养工程项目-面向维汉机器翻译的维吾尔语命名实体识别研究（2014711006）；新疆维吾尔自治区青年科技创新人才培养工程项目-维汉机器翻译模型关键技术研究（2014721032）；新疆维吾尔自治区自然科学基金-基于多特征融合的复杂形态语言建模研究（2015211B034）;中科院战略性先导科技专项-新疆少数民族信息处理（XDA06030400）；

作者简介：阿力木·木拉提（1988——），男，博士研究生，主要研究方向为自然语言处理、机器翻译；通讯作者：李晓（1957——），男，研究员，博士生导师，主要研究方向多语种信息处理、人工智能；

在极速增长。因此，研究网页识别技术，将极大提高少数民族用户使用互联网同其他网民交流、沟通，能够对地区文化、经济的发展以及信息化建设起到关键的推动作用。

目前国内外众多学者投入了网页识别相关技术的研究。Janitima Polpinij 等[1]采用 SVM（Support Vector Machine）分类器和朴素贝叶斯分类器对泰语和英语网页进行识别，结果表明，朴素贝叶斯分类器得到较高的准确率，与此同时导致显然的极度过滤问题。Kriegel 等[2]以主题频次向量（Topic Frequency Vector）作为网站的主题特征，依据网站所包含每个主题的文档数来相应特征项的权值，从而进行网站分类。通过网页类别进一步标记，将网站定义为一种有标记的树结构，采用 Markov 模型来识别商业网站[3]。文献[4]采用两步分类算法，采用优化的互信息特征抽取方法以及朴素贝叶斯，构建了基于中文网页的高性能文本分类方法。

随着计算机的普及和网络的覆盖，特别是在智能端使用维吾尔语来进行传播信息的渠道越来越多，使得更进一步促进了维吾尔文信息化的发展，以此大量的维吾尔文网站应运而生。由于维吾尔语的网页自动发现与内容采集技术相对落后，维吾尔文网站的受众有限，且有关语言网络资源不稳定等因素的影响，一些维吾尔文网站经常出现故障、知名度小和难以发展的生存危机。因此，如何在庞大的网络资源中及时、准确地发现维吾尔文网络资源、并对其进行采集和存储并加以利用，是维吾尔文信息处理中紧迫解决的基础性研究。为此现代维吾尔语的网页识别方法作为本文的宗旨，同时，对维吾尔文网页字符的编码进行了优化的研究。

维吾尔文的书写体系在一定程度上受过其他语言的影响，使得同一个字符在不同页面中有多个编码。特别是维吾尔文网页中编码不统一问题相当严重并不规范，而且维吾尔文网页的全文检索造成了一定的困难。对维吾尔文网页识别而言，网页文本的分类与识别是同一个问题，因此本文以维吾尔文网页识别为目标，研究维吾尔文网页文本识别方法。本研究采用网页文本节点特征与基于 VSM 的维吾尔语常用词统计学方法，识别维吾尔文网页。因此在本研究中，需要消除维吾尔文网页文本导致的编码混乱，使用统一的编码来表示维吾尔文字符，是维吾尔文网页识别问题的前提条件。以此，本文使用基于改进的 N-Gram 方法和维吾尔语常用词及向量空间模型相结合的方法提高网页识别率。

2 维吾尔语网页文本编码转换技术研究

一般来说，在维吾尔文网页文本的识别中会出现多编码、编码范围交叉重叠、HTML 页面 Meta 标签属性无符合标准等问题。因此，考虑到以上众多编码不规范的问题，对维吾尔文网页文本编码进行了相应的转换和调整。

维吾尔语书写规则，在基于阿拉伯文字的基础上建立的，所以维吾尔文字母所属的 Unicode 编码区域定位在阿拉伯文字编码区域。 维吾尔语中 32 个字母因位置不同有 126 个书写形体[5],然而 ISO 没有为维吾尔语字母分配自己的编码区域，故使得维吾尔文字母包含在阿拉伯编码区域。阿拉伯文在 Unicode 中分为两个区域，基本标准编码区域（0060—06FF）和扩展区域（FE70--FEFF）两种格式。为此，维吾尔文在计算机的信息处理、传送、存储和管理等过程中，普遍应用两种编码区域的维吾尔文字符。

目前许多维吾尔文网站网页上挂载 Unicode 标准基本编码字符的压缩字体库 EOT 文件，但也有少数维吾尔文网页采用自己研发的 TTF 字体库，这些字体库所使用的输入法具有字符不规范或者字体库字符归为在扩展编码区域，将为维吾尔文网页识别以及后期采集工作增加难度。因此筛选待测网页和采集之前，使用统一的编码区域，是本研究中及其重要的一个环节。这样有效地避免采集存储的网页数据库中会出现乱码和字体不显示等现象。于是本文借鉴扩展编码区域维吾尔文字符相应的编码值，对这些字符进行加以规范化的处理。相应的

编码转换规范如表 1 所示。

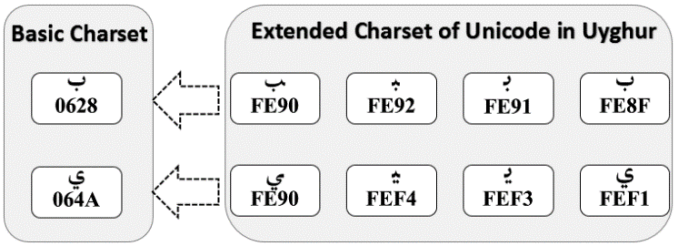

图 1 维吾尔文 Unicode 编码转换

通过根据这两个编码区域之间相互转换规则和原理，其凭借扩展编码区域的相同字符拥有不同编码值得特征， 从而依次判断网页文本的每一个维吾尔文字符，同时进行 Unicode 扩展区域字符的转换,最终匹配对应的基本区域编码值。据统计，目前维吾尔文网页所采用的字体文件大约有 264 个，如“UKIJ Tuz Tom”、“Alp Basma Aq”和“Alkatip Tuz Tom”等

3 维吾尔文网页识别模型

鉴于互联网的海量信息具有多样化、多语种的特点，如何从众多混杂各种语言的网页中，判定以及筛选内容为准确。本文综合采用以下 2 种方法对维吾尔文网页进行判断及识别。

3．1 基于改进的 N-Gram 方法

N-Gram 是指 N-1 阶马尔可夫语言模型(Markov Model)的表示。该模型使用这样的假设：随机变量$S_1,S_2,\ldots,S_m$ 中，如果其中任何一个变量 Si 出现的概率只与前面 N-1 个变量$S_{i-1},S_{i-2},\ldots,S_{i-n+1}$有关。以此序列 S 的概率:

$$P(S_i|S_{i-n+1}S_{i-n+2},\ldots,S_{i-2}S_{i-1}) = P(S_i|S_1S_2,\ldots,S_{i-2}S_{i-1})$$

N-gram 方法的具体原理是将给定文本的内容根据 N 的取值范围进行操作，形成多个长度均为 N 的文本词汇序列，每个序列称为 gram，即作为该文本的一个特征；其对所有 gram 的出现频率进行统计，并按照预设定好的阂值对其进行筛选、统计出现次数较高的 grams，以此形成该文本的 gram 特征列表。列表中的每个 gram 均为一个特征向量维度[6]。

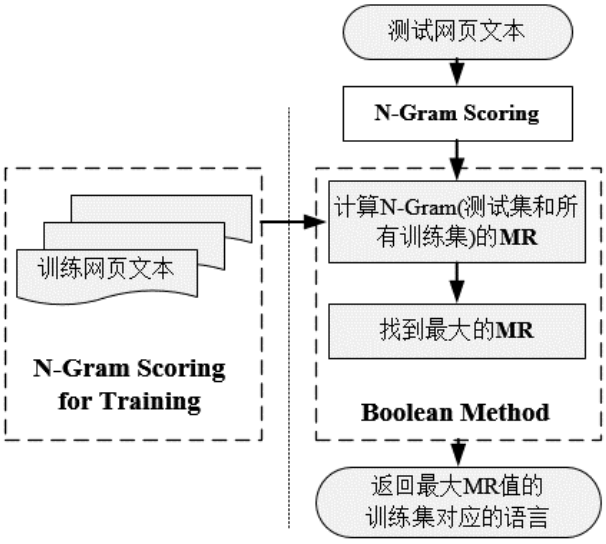

图 2 MNG 方法处理流程

如图 2 所示，本文采用的修正的 N-Gram 方法,简称 MNG 方法（Modified N-Gram Approach），由 Choong[7]通过改进原始 N-Gram 方法的基础上提出的。该方法将对某种语言

类型的训练文本进行 N-Gram 打分，并生成训练集，由此创建的训练集中不存在 N-Gram 出现频率，只有完全不同的 N-Gram 序列。以相同的方式，将对测试文本进行相应的转换，以及生成测试集。测试集是主要由 N-Gram 频率文件组成。之后，对所有训练集和测试集 N-Gram 之间的匹配率（Matching Rate，MR）进行计算。

MNG 方法在语言识别任务中与其他方法不同。一般而言，该方法对生成的 N-Gram 频率没有依赖性，其通过布尔值（Boolean Value）来决定输出结果。本文以此使用该方法对维吾尔文网页进行识别。大致思路如下：如果所有训练集的 N-Gram 特征项当中存在网页测试文本的 N-gram 特征项，该布尔值为 1；如果训练集和网页测试文本的 N-Gram 特征项之间无匹配，则布尔值为 0。这样将测试文本和训练集当中的所有 N-Grams 进行比较，通过总的布尔值除以测试网页文本中不同 N-Gram 的总数计算相应的匹配率。计算公式如下：

$$MR(Uy) = \sum_{x-1}^{n} \frac{B(x)}{n}$$

经过 N-Gram 的匹配率计算可以得出，用维吾尔文网页文本训练的 N-Gram 模型中，测试网页文本得出的匹配率越高，该待测页面是维吾尔文网页的概率以此增加。

3．2 基于常用词的方法

基于常用词的方法采用每种语言最常用候选词的词库。使用常用词方法的先出优势是，算法效率快、容易实现。然而，该方法需要构建一个常用词的频率词库，从而正确的判定测试文本的语种。

鉴于此，本文将借鉴国家语言资源监测中心少数民族分中心“维吾尔语文研究基地”、新疆师范大学“网络信息安全与舆情分析重点实验室”构建的现代维吾尔语常用词语料库。该语料库主要包含四大媒体语料：平面媒体、有声媒体、网络媒体和教材媒体等。根据常用词语料库中每个词的频度，文本进行统计分析研究，赋予每个词一个相关的值 [8]。文献[9][10][11]中，作者从词语的使用频率角度对词语进行基本考察，并其维吾尔语词语的“词种数、频次、文本书、词长”作为构建常用词库的依据。该语料具体情况如表 1 所示：

表 1 维吾尔语常用词语料统计结果

词次	词汇种数	词干种数	总文档数
43,529,435	703,669	147,054	96,025

从语料统计分析结果得出，学者研制的维吾尔语常用候选词，仅在全四大媒体语料中的覆盖度为 95.23%。数据以此表明，这些常用候选词在语料的覆盖度几乎接近于所有四大媒体语料包含的词语。终归一言，该语料库完全地描述常用词具有的特性，其能够完全地代表维吾尔语常用词。因此，本文引入向量空间模型，通过语料中的每一篇文档用向量来表示，从此有效地整合维吾尔语常用词语料库与向量空间模型，因而验证该方法在维吾尔文网页识别的有效性及可行性。

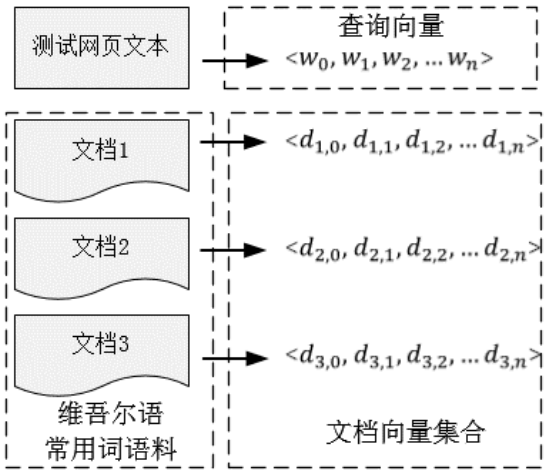

图 3 维吾尔文常用词语料在 VSM 中的应用

向量空间模型（Vector Space Model）[12]：通过向量的方式来计算相似度，其中由一个向量来表示一篇文档，而测试文档也同样用一个向量来表示[13]。该模型基于以下思想：使用语料库中每一篇文本（本文将常用词语料看作该模型的参照文档）定义一个文档向量；每个文档向量都有 n 个分量。文档向量中的分量是指整个语料文本中计算出来的每个独立词项的权值。每篇文档中，词项权值以基于词项在所有语料中出现的频率及词项在某一个文档中出现频率自动赋值。一般来讲，权值可以用分量的出现频率来近似表示。文中提到常用词语料文档集中，一个文档的权重向量表示为：

$$M_d = [w_{1,d}, w_{2,d}, \ldots, w_{N,d}]^T$$

当维吾尔文网页识别开始时，首先对测试网页文本进行相应的预处理,并将其表示成一个查询向量$Q = (q_1, q_2, \ldots, q_n)$，共有 n 个独立词项。测试网页文本向量 Q 和文档向量 M 的相似度可以简单地定义为两个向量的内积。这种策略经常被用来两篇文档的相似度。 以此，依据相似度公式来计算向量 M 和向量 Q 的相似度[10]。相似度计算公式如下：

$$Sim(M,Q) = \frac{M * Q}{|M| * |Q|}$$

$$= \frac{\sum_{i=1}^{n} w_i * q_i}{\sqrt{\sum_{i=1}^{n} w_i^2 * \sum_{i=1}^{n} q_i^2}}$$

上述提出过，在向量空间模型中，每篇文档向量中有 n 个分量，每一个分量表示该词语在该文档中的权值，用以描述该词语在表示此文档内容时所起作用的重要程度。计算词项权值唯一的原则是要尽最大限度的区分不同文本。因此，网页测试文本向量（称为查询向量）中的每个词语在语料文本向量中出现的频率来表示该词项的权权值。其具体计算公式如下所示：

$$W(t,d) = \frac{t_f(t,d) * log(\frac{N}{n} + a)}{\sqrt{\sum_{i=1}^{m} [t_f(t,d) * log(\frac{N}{n} + a)]^2}}$$

其中,$W(t,d)$表示常用词文本语料文档集中，查询向量中的词项 t 在文本 d 中的权值。

$1 \le i \le m$，m 为其中的文本 d 中词项 t 的频次。$t_f(t,d)$表示 t 在常用词语料文本 d 中的覆盖率，N 为常用词语料文本总数，n 为包含查询向量中的词项 t 的常用词语料文档的个数。

4 实验与分析

本文提出的维吾尔文网页识别研究，分为基于 N-Gram 的方法和基于常用词特征向量的方法来进行实验。现有的维吾尔文网页识别方法主要是以表达式规则和人工辅助，并没有给出标准的识别结果，以此与本文提出的方法无法进行比较和分析。为了验证本文所提出的方法在不同的网页数据的性能和效率，采用了三种不同类型的维吾尔文网页数据：新闻类、论坛以及博客。

数据来源：目前维吾尔文还没有标准的、开放的语种识别语料库。因此，本文实验所使用的数据来自于新疆最大的维吾尔文网址导航（www.ulinix.cn），其该网站提供的网址列表中随机抽取 210 个站点，且对这些网站针对性地提取网页文档，以 txt 文档格式进行存储，由此构建 N-Gram 特征库，提供有效的数据。详细语料的类型和规模以表 2 所示。考虑到采用 SVM 进行网页识别，文档需要用向量来表示，故用维吾尔语常用词语料文本作为向量空间模型的参照文档，从而与测试文档向量进行计算相似度，以及判定维吾尔文网页。维吾尔文常用词语料库文档统计结果，以表 1 所示。

表 2 用于构建 N-Gram 特征库的网页文档分布统计

站点类型	#站点数	#网页文档数	
		训练集	测试集
新闻	65	100,000	5000
论坛	65	60,000	3000
博客	80	20,000	1000

另外，在 N-Gram 模型中阶数 N 的确定是维吾尔文网页识别的关键所在。将用所有 N-Gram 特征项作为网页文本的特征，导致特征维数非常高，这会对识别效率和速度有极其影响。由此，从庞大的 N-Gram 特征项集合中筛选出对网页识别贡献较大的 N-gram 特征项，将保留能够描述训练文本中维吾尔文的语言现象和特点的 N-Gram 特征项。为此本文按照[14]提出的维吾尔文 N-Gram 模型的参数 N 的选取问题，训练文本用 5-Gram 来表示，并按出现频率对特征项进行降序排列，选取前 1000 个 5-gram 特征项，保存在训练特征库中。

评估方法：实际上维吾尔文网页识别是一个分类问题，由此采用分类系统的三个评价指标：准确率 P (Precision)、召回率 R（Recall）和 F1 值，对本文所提出的方法进行整体评估。本文首先使用基于 MNG 方法进行了维吾尔文网页识别实验，具体的识别结果如下所示。

表 3 使用 MNG 和常用词特征向量方法得出的维吾尔文网页识别结果（%）

方法	MNG			基于常用词		
网页类型	P	R	F1	P	R	F1
新闻	99.8	100	99.9	98.2	99	100
论坛	88.6	95	91.6	82.4	97.8	89.4
博客	95	80.1	87	88.5	98	93

使用 MNG 方法进行网页识别的结果表明，通过不同类型的网页文本进行测试，整体来看，基于 N-Gram 模型的方法对维吾尔文网页的识别性能相对较高。其中，对维吾尔文的新

闻类网页的识别达到了 99.9%的 F1 值，说明了此类网页上识别性能最好；当论坛类网页上进行测试时，识别效果明显变低；对博客类的网页的识别效果明显地有所下降，说明该方法在此网页上性能相对较差，F1 值从 99.9%下降至 87%。

使用常用词特征向量进行网页识别结果表明，以同样的测试数据，该方法整体识别和基于 N-Gram 模型之间的相差不大，对识别效果的整体影响并不明显。与另外一种方法不同，该方法分别在新闻类和博客类网页的识别性能达到了最高 F1 值 100%和 93%。尽管对两类（新闻、博客）网页的识别性能有所提高，但对论坛类网页的识别性能显得较弱。

通过综合分析本文提出的两种方法对维吾尔文网页的识别结果，以及观察结果对比中的 F1 值，为此验证融合的方法在本研究的可行性，本文采用了融合方法进行网页识别。具体的实验结果如表 3 所示。

表 4 融合方法得出的维吾尔文网页识别结果（%）

网页类型	P	R	F1
新闻	100	100	100
论坛	83.6	98	90.2
博客	95	98	96.4

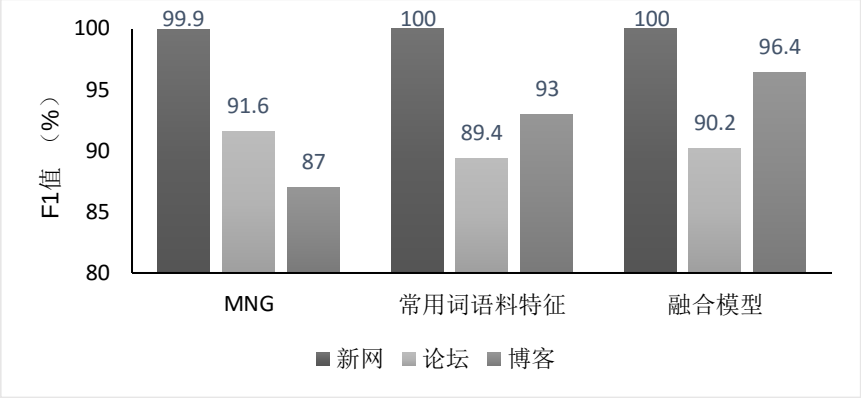

图 4 MNG、常用词特征向量及融合模型所得的 F1 值对比图

图 4 中所示的识别结果中可以得出，当采用以两种方法结合的方式进行维吾尔文网页识别时，总体识别效果优越前两种方法。特别是融合模型对博客类型的网页的识别性能有了显著的提高，以此融合模型相比前两种方法能够互补前两种方法的识别性能较差的问题。

通过分析以上得出的识别结果，本文所提出的两种方法有以下几个特点：

1） 网页中只有少量维吾尔文文本，因而无法构建语言模型并不能有效地描述文本，将会导致一定程度的错误。

2） 无论任何类型的维吾尔文网页，对正文部分较长的文档输出较高的准确率。

3） 维吾尔语常用词语料库特征向量对正文较短的网页识别效果相对较高，适合作为维吾尔文网页特点。

5 总结与下一步工作

本文提出了基于修正的 N-Gram 模型和维吾尔语常用词向量特征方法，以此进行了维吾尔文网页识别研究。同时，针对维吾尔文网页中常见的页面编码混乱问题做了分析及预处理，以便快速、准确的识别维吾尔文网页。同时使用不同类型的网页数据基础上，构建了 N-Gram

模型的特征库，以同类型的测试数据上进行了网页识别的实验，获取了较高的识别效果。另外，考虑到测试网页中出现的词语频率和文档数在维吾尔文网页识别中起重要的作用，统计及分析维吾尔语常用候选词，并与向量空间模型进行融合，从而提高了识别维吾尔文网页的概率，在实际系统的应用中得到了较好的性能效果，其系统综合性能提高到 90%以上。

本文在后期的工作中，将会进一步地扩展维吾尔语常用候选词，以便增加常用词在训练文本的覆盖度，同时构建更多维数的 N-Gram 特征项，从而更加地提升维吾尔文网页识别的整体性能。

参考文献

[1] Polpinij J, Chotthanom A, Sibunruang C. Content-based text classifiers for pornographic web filtering [C]. IEEE International Conference on System, Man and Cybernetics.Taipei,Taiwan,2006: 1481-1485

[2] Kriegel H P, Schubert M. Classification of Websites as Sets of Feature Vectors[C]. International Conference on Datebases and Applications (DBA 2004) , Innsbruck,Austria,2004:127-132

[3] Ester M, Kriegel H P, Schubert M. Web site mining: a new way to spot competitors, customers and suppliers in the World Wide Web[C]. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002). ACM, New York, NY, USA,2002, 249-258

[4] 樊兴华，孙茂松.一种高性能的两类中文文本分类方法[J].计算机学报,2006,29(1):124-131

[5] 哈力克·尼亚孜，吾买尔·阿皮孜。基础维吾尔语[M]。新疆大学，1995:1-2

[6] 庞景安.Web 文本特征提取方法的研究与发展[J]。情报理论与实践，2006 29(3): 338-340

[7] Choong C, Mikami Y, Marasinghe C A, et al. Optimizing n-gram Order of an n-gram Based Language Identification Algorithm for 68 Written Languages[J]. International Journal on Advances in ICT for Emerging Regions (ICTer), 2009, 2(2).

[8] 艾孜尔古丽.现代维吾尔语常用词计量研究[D].新疆师范大学,2013

[9] 艾孜尔古丽,齐向卫.基于网站用词调查的现代维吾尔语词干提取和应用研究[J].计算机应用与软件,2012,29(3):32-35

[10] 艾孜尔古丽,努尔艾合买提.现代维吾尔语常用词统计关键技术研究[J].中文信息学报,2014,28(5):192-197

[11] 艾孜尔古丽,艾山江·阿不力孜.现代维吾尔文网络媒体用词研究[J].计算机应用与软件,2012,29(2):67-69

[12] Salton G, Wong A, Yang C S, et al. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.

[13] (美)格罗斯曼(Grossman, D. A.), (美) 弗里德 (Frieder,) 等. 信息检索: 算法与启发式方法: 第 2 版[M]. 人民邮电出版社, 2009.

[14] 图尔妮萨古丽·赛麦提. 基于 N-gram 的维吾尔文文本分类研究与系统实现[D].新疆大学,2014.

阿力木·木拉提（1988--），男，博士研究生，主要研究领域为机器翻译、自然语言处理。 Email: alim.murat@ms.xjb.ac.cn

艾孜尔古丽（1987--），女，讲师，主要研究领域为计算语言学、自然语言处理。 Email:Azragul2010@126.com

杨雅婷（1985--），女，副研究员，主要研究领域为机器翻译、自然语言处理。Email: yangyt@ms.xjb.ac.cn

通讯作者：李晓（1957--），男，研究员，博士生导师，主要研究方向为多语种信息处理、人工智能。Email: xiaoli@ms.xjb.ac.cn