基于TextRank算法和互信息相似度的维吾尔文关键词提取及文本分类

阿力甫·阿不都克里木[1,2,3] 李 晓[1,2]
(中国科学院新疆理化技术研究所 乌鲁木齐 830011)[1]
(中国科学院大学 北京 100039)[2] (新疆多语种信息技术重点实验室 乌鲁木齐 830046)[3]

摘 要 针对维吾尔语文本的分类问题,提出一种基于TextRank算法和互信息相似度的维吾尔文关键词提取及文本分类方法。首先,对输入文本进行预处理,滤除非维吾尔语的字符和停用词;然后,利用词语语义相似度、词语位置和词频重要性加权的TextRank算法提取文本关键词集合;最后,根据互信息相似度度量,计算输入文本关键词集和各类关键词集的相似度,最终实现文本的分类。实验结果表明,该方案能够提取出具有较高识别度的关键词,当关键词集大小为1250时,平均分类率达到了91.2%。

关键词 维吾尔语,文本分类,关键词提取,TextRank算法,互信息相似度
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.12.006

Uyghur Keyword Extraction and Text Classification Based on TextRank Algorithm and Mutual Information Similarity

Ghalip ABDUKERIM[1,2,3] LI Xiao[1,2]
(Xinjiang Technical Institute of Physical and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China)[1]
(University of Chinese Academy of Sciences, Beijing 100039, China)[2]
(Xinjiang Key Laboratory of Multi-language Information Technology, Urumqi 830046, China)[3]

Abstract This paper proposed Uyghur keyword extraction and text classification scheme based on TextRank algorithm and mutual information similarity for the issues of classification in Uyghur language text. Firstly, the input document is pre-processed to filter out non-Uyghur characters and stop words. Then, keywords set in the text is extracted through using the TextRank algorithm which is weighted by semantic similarity of words, position of words and importance of frequency. Finally, the similarity between keyword sets in the input text and a variety of keyword sets is measured according to the mutual information similarity, and the text classification is realized. The experimental results show that this scheme can efficiently extract the keywords, and the average classification rate reaches 91.2% when the set size is 1250.

Keywords Uyghur language, Text categorization, Keyword extraction, TextRank algorithm, Mutual information similarity

1 引言

随着信息和存储技术的飞速发展,互联网上的文本数量呈爆炸式增长,需要一种技术自动地对互联网中的文本进行组织和分类,文本分类是按照某一种算法把文本分到某一个预先制定好的类别中的过程[1]。随着国家对新疆地区的大力投入,其信息化建设得到了快速发展,维吾尔语等少数民族语种的大量文字信息开始以数字化形式呈现。对维吾尔语书写的大量文本数据进行文本分类,对新疆地区的文化发展具有一定的意义[2]。目前,对英文和中文等大语种的文本分类技术已经得到大量研究,并趋于成熟。然而,对维吾尔语表述的数字文本的文本分类的相关方面的研究还处于起步阶段。对于维吾尔语文本的分类,关键词的提取是至关重要的。维吾尔语是一种黏着性语言,其时态变化较为复杂,且具有丰富的形态结构[3]。维吾尔语表达一个具体含义的最小单元经常不是一个独立单词,而是多个单词的组合,这为关键词提取造成了很大的困难[4]。

现有的常见关键词提取方法有文本频率(Document Frequency,DF)、CHI统计量以及信息增益(Information Gain,IG)等。其中,最具代表性的为词频-逆向文件频率(Term Frequency-Inverse Document Freq,TF-IDF)算法[5],其利用词频和逆文本频率的乘积对单词重要性进行加权。目前,学者提出了多种维吾尔语文本关键词的提取方法,例如:文献[6]提出一种基于语义词特征提取的维吾尔语文本的分类方法,

到稿日期:2016-03-23 返修日期:2016-05-21 本文受新疆多语种信息技术重点实验室开放课题(XJDX0905-2013-06)资助。
阿力甫·阿不都克里木(1980—),男,博士生,高级工程师,主要研究方向为维吾尔文信息处理、互联网技术等;李 晓(1957—),男,硕士,研究员,博士生导师,主要研究方向为多语种信息处理、信息系统研究与开发等,E-mail:Ghalip@ms.xjb.ac.cn。

利用一种组合统计量(DME)来衡量文本中相邻单词之间的关联程度,以此来提取关键词;文献[7]利用 χ^2 统计量来提取关键词,并利用支持向量机(Support Vector Machine,SVM)算法来构造维吾尔语文本分类器;文献[8]提出一种新的统计量(CHIMI),将 χ^2 统计量和互信息(Mutual Information,MI)进行结合组成 CHIMI,抽取 Bigram 作为文本特征。然而,这些统计方法只是较为简单地对单词出现的频率等信息进行统计,不能很好地反映出单词的语义信息和单词位置对单词重要性的影响。

TextRank 算法[9]是一种基于图的排序算法,能够通过投票机制获得重要性较大的关键词。目前,也有学者将 TextRank算法应用到维吾尔语文本关键词提取中。例如,文献[10]利用 TextRank 算法来提取关键词集合,TextRank 算法利用单词位置和单词频率来加权,其证明了 TextRank 所提取的关键词优于传统 TF-IDF 算法。但是,该方法中的Text-Rank 加权过程没有考虑单词的语义信息,具有一定的局限性。

本文对传统的 TextRank 进行了改进,在考虑词语位置和词频重要性的同时,添加了词语语义相似度来加权Text-Rank 关键词提取算法,以此提出一种基于 TextRank 算法和互信息相似度的维文关键词提取及文本分类方法。在对输入文本进行预处理后,利用词语语义相似度、词语位置和词频重要性加权的 TextRank 算法提取本文关键词集合。然后,计算输入文本关键词集和各类关键词集的互信息相似度度量,根据判断阈值实现文本的分类。实验结果表明,本方案能够提取出具有较高识别度的关键词,当关键词集大小为 1250 时分类效果最佳,平均正确分类率达到了 91.2%。

2 提出的文本分类方案

本文提出一种维吾尔语文本信息过滤方案,首先需要收集一些已分类的文本作为训练文本集。本文过滤方案主要包括 3 个部分:1)维吾尔语文本预处理阶段,过滤掉停留词,并提取维吾尔语单词的词干,以降低文本维度;2)特征提取阶段,使用 TextRank 算法根据语义相似度、词语位置和词频提取关键词集;3)文本分类阶段,通过分布式互信息相似度度量,计算输入文本关键词集与训练文本获得的各类关键词集的相似度,来判断输入文本的类别。另外,在文本分类之后,通过人为判断文本分类的正确性,并将其作为反馈来调整相似度比较的判断阈值,以获得最佳分类性能。本文文本分类方案框架如图 1 所示。

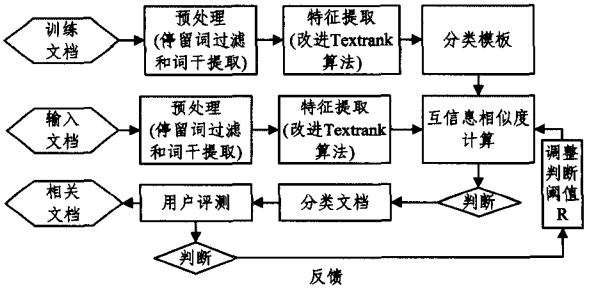

图 1 本文文本分类方案框架

3 文本预处理

维吾尔语文本的预处理主要包括两个部分:文本过滤和词干提取。其中,文本过滤主要过滤掉文本中非维吾尔语文字和停用词;词干提取是提取文本中具有真正含义的词汇。经过文本预处理,可将文本原始特征维度降低约一半。

在文本去噪过程中,首先对文本进行过滤,获得维吾尔语单词集,然后通过与事先准备好的停用词表进行比对,过滤掉停用词。停用词为对文本主题没有贡献、不包含文章类别信息的词,例如介词、副词、代词等。去掉停留词能够实现特征降维,提高分类精度[11]。

在词干提取过程中,首先根据维吾尔语单词与单词之间的空格符来进行分词。由于维吾尔语单词是由字母拼写而成的,通过将不同的词缀粘贴到单词的头部来实现语法功能,因此提取文本中能够代表真实含义的词汇相当困难。维吾尔语中,同一词干可以演变为很多不同含义的词语,虽然这些词语的词形不同,但词义却不会有很大区别,其中一个典型例子如表 1 所列。为了提取单词的词义,并考虑特征的数量,本文以词干(医院)作为特征项,以此从文本中提取出词干集。

表 1 维吾尔词语变体

变形(汉文)	变形(维吾尔文)	词缀	词干
从医院	دوختۇرخانىدىن	دىن	
在我们的医院	دوختۇرخانىمىزدا	مىزدا	
把医院	دوختۇرخانىنى	نى	
在医院	دوختۇرخانىدا	دا	
到医院	دوختۇرخانىغا	غا	دوختۇرخانا
我们的医院	دوختۇرخانىمىز	مىز	(医院)
你的医院	دوختۇرخانىڭىز	ڭىز	
你们的医院	دوختۇرخانىڭلار	ڭلار	
在医院的	دوختۇرخانىدىكى	دىكى	

4 基于 TextRank 算法的关键词提取

TextRank 算法的基本思想是将一个文本文档进行分割,从而形成多个子单元。然后,将这些单元构建成一个图模型,文本中的单词以结点的形式表示,单词间的关联以边的形式表示,并通过投票机制对文本中具有代表性的关键词进行排序[12]。

本文利用 TextRank 算法将预处理后获得的初始关键词构建成一个有向图 $G=(V,E)$,其中,V 表示节点集合,E 表示有向边集合。在图 G 中,以初始关键词集为节点集合,以 WordWeigh 为节点权重。TextRank 算法的基本流程如图 2 所示[9]。

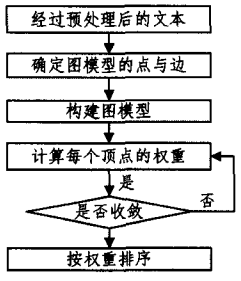

图 2 TextRank 算法的基本流程

TextRank 算法中,通常利用滑动窗口方法获取词与词之间的关系,为了加强词与词之间的联系,一般选用较大的滑动窗口[13]。通过窗口获取关键词后,将窗内左边词用于出度,将右边词用于入度,并构建一个初始 TextRank 模型。图中节点 V_i 的权重计算表达式如下:

$$WS(V_i)=(1-d)+d\times\sum_{v_j\in \ln(V_i)}\frac{W_{ji}}{\sum_{v_k\in Out(v_j)}W_{jk}}WS(V_j) \quad (1)$$

其中，d 表示取值范围为[0,1]的阻尼系数，一般取值为 0.85。W_{ji} 表示 V_j 的影响力到 V_i 的权重比重。对于一个给定的点 V_i，$ln(V_i)$为指向 V_i 的点集合，$Out(V_i)$为点 V_i 指向的点集合。根据式(1)和上一步获取的 W_{ji} 进行多次迭代计算直至所有 $WS(V_j)$收敛后停止运算，从而得到候选关键词集。

对于 W_{ji} 的计算，传统 TextRank 算法主要将其分为 2 个部分加权获得，即位置重要性和词频重要性[14]。然而，对于维吾尔语，语义信息不能忽略，所以本文对 TextRank 算法进行改进，在 W_{ji} 计算中融入词语语义重要性。

令 W 表示结点的整体影响力权重，a,b,c 分别表示词语语义相似度、词语位置和词频重要性对整体权重的影响因子，且 $a+b+c=1$。对于任意两个结点 V_i 和 V_j，$e=\langle V_i,V_j\rangle$表示这两个节点之间的有向边，通过该有向边来传递结点 V_i 对 V_j 的影响力，该影响力的大小由有向边的权重决定。词语语义相似度、词语位置和词频重要性的影响力权重表达式如下。

(1)令 $W_a(v_i,v_j)$表示结点 V_i 的语义相似度通过有向边传递到结点 V_j 时对其造成的影响力权重，表达式为：

$$W_a(v_i,v_j)=\frac{sim(v_i,v_j)}{\sum\limits_{v_i\in Out(v_j)} sim(v_i,v_k)} \tag{2}$$

其中，$sim(v_i,v_j)$表示结点 V_i 和 V_j 间的词语语义相似度，词语语义相似度越高，说明有向边传递 V_i 对 V_j 造成的影响力越大。

(2)令 $W_b(v_i,v_j)$表示结点 V_i 的词语位置通过有向边传递到结点 V_j 时对其造成的影响力权重，表达式为：

$$W_b(v_i,v_j)=\frac{Loca(v_j)}{\sum\limits_{v_i\in Out(v_j)} Loca(v_k)} \tag{3}$$

其中，$Loca(v_j)$表示结点 V_j 的位置重要性。通常情况下，文本标题中出现的词为全文主题关键词的概率较大。故当 V_j 出现在标题中时，$Loca(v_j)$设为 20；当 V_j 出现在正文中时，$Loca(v_j)$设为 1。

(3)令 $W_c(v_i,v_j)$表示结点 V_i 的词频通过有向边传递到结点 V_j 时对其造成的影响力权重，表达式为：

$$W_c(v_i,v_j)=\frac{Freq(v_j)}{\sum\limits_{v_i\in Out(v_j)} Freq(v_k)} \tag{4}$$

其中，$Freq(v_j)$表示结点 V_j 在文本中出现的次数，体现出高词频的词语将从连接点获得更高的影响力权重。则 W_{ji} 可通过下式计算得到：

$$w_{ij}=a\cdot W_a(v_i,v_j)+b\cdot W_b(v_i,v_j)+c\cdot W_c(v_i,v_j) \tag{5}$$

5 基于互信息相似度的文本分类

相似度计算中通常有 5 种相似度度量，即匹配度、重叠度、Jaccard 系数、夹角余弦和互信息。文献[15]证明了利用互信息可以获取比其它相似度计算方法更高的精确度。

互信息表示一个随机变量与其它变量之间的信息量，即表示两个概率分布之间的距离度量。设定 k_p 为本文TextRank算法从文本 A 中提取的第 p 个关键词，k_q 为从文本 B 中提取的第 q 个关键词。$p(k_p)$和 $p(k_q)$分别表示文本 A 和文本 B 中生成 k_p 和 k_q 的概率，利用式(6)和式(7)可以分别计算出其结果值，其中 $PA_1(k_p)$和 $PA_2(k_q)$分别表示文本 A 和文本 B 的“聚合投影结果”。$N(A)$和 $N(B)$分别表示文本 A 和文本 B 中关键字的个数，$p(k_p,k_q)$表示 k_p 和 k_q 的联合概率，可以式(8)计算。

$$p(k_p)=\frac{PA_1(k_p)}{\sum\limits_{j=1}^{N(A)} PA_1(k_j)} \tag{6}$$

$$p(k_q)=\frac{PA_2(k_q)}{\sum\limits_{j=1}^{N(B)} PA_2(k_j)} \tag{7}$$

$$p(k_p,k_q)=\frac{PA_1(k_p)*PA_2(k_q)}{\sum\limits_{j=1}^{N(A)} PA_1(k_j)+\sum\limits_{j=1}^{N(B)} PA_2(k_j)} \tag{8}$$

用 $\sigma(A_i)$表示文本 A 和文本 B 的关键词集合之间的互信息相似度。如果 $\sigma(A_i)$的结果值较高，表示文本 A 的关键词与文本 B 的关键词类似，即文本 A 与文本 B 的属性类似。$\sigma(A_i)$的表达式如下：

$$\sigma(A_i)=\sum_{k_p\in N(A)}\sum_{t_q\in N(B)} p(k_p,k_q)\log\frac{p(k_p,k_q)}{p(k_p)p(k_q)} \tag{9}$$

表 2 列出了文本 A 和文本 B 中关键词的词频、聚合投影和分布式投影的结果。以表 2 中的“聚合投影结果”这一栏为例，如果文本 A 中的 k_p 为“大学”，文本 B 中的 k_q 为“教师”，则 $p(k_p)=\frac{1}{5}$，$p(k_q)=\frac{1}{4}$，$p(k_p,k_q)=\frac{1*1}{1+1}$。当获取所有的 $\{\frac{1}{5},\frac{1}{5},\frac{1}{5},\frac{1}{5},\frac{1}{5},0\}$和$\{\frac{1}{4},\frac{1}{4},\frac{1}{4},0,0,\frac{1}{4}\}$之后，文本 A 和文本 B 之间的$\sigma(A_i)$值为 5。

为了更好地表示两个关键词集的相似度，本文在互信息相似度度量的基础上提出一种分布式互信息相似度度量，即根据“分布式聚合结果”来计算概率。$p(k_p)$和 $p(k_q)$的值可利用式(10)和式(11)获得，其中用 $DA_1(k_p)$和 $DA_2(k_p)$分别表示文本 A 和文本 B 中的关键词 k_p 的词频。可以通过式(11)计算出 $p(k_p,k_q)$的值。

$$p(k_p)=\frac{DA_1(k_p)+DA_2(k_p)}{\sum\limits_{r=1}^{N(A)} DA_1(k_r)+\sum\limits_{s=1}^{N(B)} DA_2(k_s)} \tag{10}$$

$$p(k_p,k_q)=\frac{\min(DA_1(k_p)+DA_2(k_q))}{\sum\limits_{r=1}^{N(A)} DA_1(k_r)+\sum\limits_{s=1}^{N(B)} DA_2(k_s)} \tag{11}$$

以表 2 中的“分布式聚合结果”这一栏为例，如果文本 A 中的 k_p 为“大学”，文本 B 中的 k_q 为“教师”，则 $p(k_p)=\frac{10}{27}$，$p(k_q)=\frac{8}{27}$，$p(k_p,k_q)=\frac{\min(6,3)}{16+11}=\frac{3}{27}$。当获取$\{\frac{10}{27},\frac{8}{27}\frac{5}{27},\frac{2}{27},\frac{1}{27},\frac{1}{27}\}$和$\{\frac{10}{27},\frac{8}{27}\frac{5}{27},\frac{2}{27},\frac{1}{27},\frac{1}{27}\}$之后，文本 A 和文本 B 之间的相似度$\sigma(A_i)$为 0.476。

表 2 文本 A 和文本 B 中关键词的词频、聚合投影和分布式投影的结果

关键词	文本 A			文本 B		
	词频	聚合投影结果(PA_1)	分布式聚合结果(DA_1)	词频	聚合投影结果(PA_1)	分布式聚合结果(DA_1)
大学	6	1	6	4	1	4
教师	5	1	5	3	1	3
图书馆	2	1	2	3	1	3
考试	2	1	2	—	0	0
数学	1	1	1	—	0	0
竞赛	—	0	0	1	1	1
总计	16	5	16	11	4	11

本文为了分类维吾尔语文本，将输入文本提取的关键词集合和训练集获得的各个类的关键词集合进行相似度计算，将相似度 σ 大于设定阈值的文本归到该类中。其中，阈值大小的设定是通过多次实验获得的。本文通过大量实验证明，在采用分布式互信息相似度度量时，设定相似度阈值为 0.55 时的分类准确度最高，所以在第 6 节的实验过程中都设定阈值 $\sigma=0.55$。

6 实验及分析

6.1 实验环境

为了评估本文方案的性能，构建一个计算平台，以 Intel 酷睿 i5 作为 CPU，主频为 2.4GHz，操作系统为 Windows7，利用 Matlab2011 进行实验。

对于维吾尔语的文本分类应用，目前还没有可使用的标准文本集。本文从人民网(维吾尔语版)和天山网等主流维吾尔语网站上收集了 2500 篇文本，通过人工方式将其分为 6 类：政治、经济、体育、旅游、教育和文化。其中，1600 篇文本作为训练集，900 篇作为测试集。各类的训练和测试样本数如表 3 所列。

表 3 分类文本库

类别	训练文本数	测试文本数
政治	254	147
经济	247	159
体育	271	144
旅游	261	147
教育	286	137
文化	281	166

6.2 性能指标

本文采用分类中常用的性能指标 F_1 来评估方案的性能，其由分准确率(Precision)和召回率(Recall)计算获得。

$$准确率(P)=\frac{a}{a+b}$$

$$召回率(R)=\frac{a}{a+c}$$

其中，a 表示正确分类的文本数，b 表示分类为该类但不属于该类的文本数，c 表示属于该类但未被分类到该类的文本数。通常将准确率和召回率进行综合，得到评估文本分类质量的 F_1 值，其表达式如下：

$$F_1=\frac{2RP}{R+P} \tag{12}$$

通常情况下，方案的 F_1 值越高，分类效果越好。实验中，本文将各个类别的 F_1 值求平均，得到最终性能指标：F_1 平均值。

6.3 分类实验

实验中，首先对维吾尔语文本集进行预处理，为了方便后续处理，把文本转换成 UTF-8 二进制编码格式。然后，过滤掉文本中的非维吾尔语字符和停用词。预处理结束后，获得一个具有 24420 个关键词的初始关键词集，然后进行词干提取，将同一词根演变而来的特征进行聚合，使初始关键词集项降维到 13826 个。

利用提出的 TextRank 关键词提取算法提取出与类别具有高互信息(高区分度)的词干作为最终关键词。设定每个类别提取 500～1500 个关键词。表 4 描述了本文关键词提取方法在政治、经济、体育、旅游、教育和文化类别中提取的前 5 名的关键词，这些特征词具有最强的区别能力。

表 4 本文 TextRank 关键词提取方法获得的各类别中前 5 名的关键词

政治	经济	体育	旅游	教育	文化
پارتىيە	پۇل-مۇئامىلە	مۇسابىقە	ساياھەت	مەكتەپ	ئادەت
(党)	(金融)	(比赛)	(旅游)	(学校)	(习俗)
مەركەز	ئىقتىساد-سودا	تەنتەربىيە	مەنزىرە	ئوتتۇرا	ناخشا-ئوسسۇل
(中央)	(经贸)	(体育)	(风景)	(中学)	(歌舞)
مەملىكەت	شىمبورت-ئېكسپورت	كوماندا	تەڭرىتاغ	ماتارىپ	ئەدەبىيات-سەنئەت
(国家)	(进出口)	(球队)	(天山)	(教育)	(文艺)
مۇقىملىق	ئەرەققىيات	ئاتلېتىكا	قارلىق تاغ	تىل	يېزىق
(稳定)	(发展)	(田径)	(雪山)	(语言)	(文字)
ھۆكۈمەت	كىرىم	ۋاسكېتبول	مەنزىرە رايون	ئىمتىھان	ئۇرۇپ-ئادەت
(政府)	(收入)	(篮球)	(景区)	(考试)	(民俗)

首先，对本文方法进行验证性实验，设置提取的关键词集大小分别为 500，750，1000，1250 和 1500。在这些情况下分别进行分类实验，并计算分类性能指标 F_1，各种场景下的准确率 P、召回率 R 和 F_1 值的结果如表 5 所列。图 3 描述了本文方案在不同关键词集大小下各类别分类的 F_1 值(对应于表 5)。

表 5 本文方法在各种关键词集大小的情况下的分类结果(%)

类别	500			750			100			1250			1500		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
政治	82.4	47.2	59.8	72.7	72.0	72.4	81.2	69.7	74.6	83.7	81.6	82.4	81.1	79.3	80.0
经济	61.5	74.5	67.1	81.3	76.1	78.6	85.8	93.4	89.3	95.4	97.3	96.3	83.6	97.5	89.8
体育	65.2	67.1	66.1	73.7	75.5	74.4	84.3	86.3	85.2	89.2	95.2	92.1	86.1	87.2	86.5
旅游	60.7	65.7	62.8	74.6	66.2	70.1	87.1	71.6	78.3	86.1	82.7	84.1	90.3	73.4	80.6
教育	77.7	71.8	74.2	76.6	95.1	84.9	86.2	96.1	90.9	92.1	97.3	94.6	93.0	95.7	94.0
文化	73.7	72.3	72.4	87.8	74.7	80.4	77.9	89.9	83.1	93.9	96.5	95.0	87.9	94.2	90.8

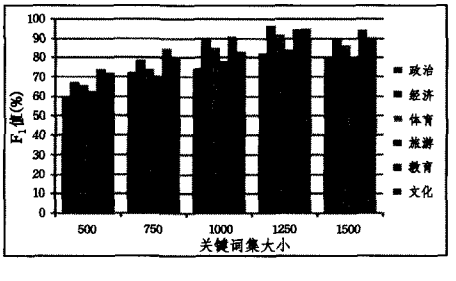

图 3 不同关键词集大小下本文方案中各类别分类的 F_1 值

从图 3 可以看出，在不同关键词集大小下，各个类别的分类精度不等，其中“经济”和“教育”类别的分类精度最高，“旅游”的分类精度相对较低，这可能是因为与其它类相比，“旅游”类中所含的关键词区别度相对较低。另外，随着关键词集大小的不同，整体分类准确性也不同。

为了更好地证明本文方案对维吾尔语文本分类的性能，将本文文本分类算法与基于传统 TF-IDF 的方法以及文献[8]提出的结合 χ^2 统计量和互信息的方法以及文献[10]提出的基于传统 TextRank 的方法进行比较。图 4 描述了各种方法在不同关键词集大小的情况下 F_1 值的比较结果，其为各类别分类的 F_1 值的平均值。

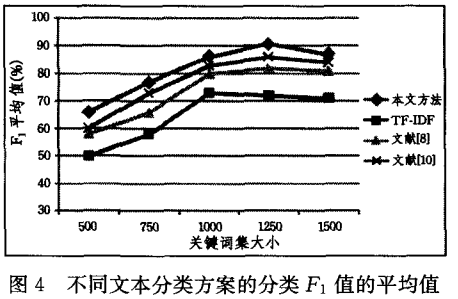

图 4 不同文本分类方案的分类 F_1 值的平均值

从图4可以看出，随着关键词集大小的增加，各种方法的分类精度都随之增加，然后趋于稳定或有所下降。这是因为当关键词集过大时，其选入了有些不重要的“噪声”特征，影响了分类效率。当关键词集大小为1250时，各种方案都获得了最高性能，其中本文方案获得的 F_1 值为0.91，分别比传统TF-IDF、文献[8]和文献[10]高出约23%、11%和6%。这是因为本文TextRank算法提取关键词时，除了考虑单词位置和单词词频影响外，还考虑了单词的语义权重，使其能够更好地提取出识别度高的关键词。另外，本文还利用能够更好表达相似度的分布式互信息相似度度量来分类文本。

结束语 本文针对维吾尔语文本的分类问题，提出了一种基于TextRank算法和互信息相似度的维文关键词提取及文本分类方法。首先，对输入文本进行预处理，滤除非维吾尔语字符和停用词；然后，利用词语语义相似度、词语位置和词频重要性加权的TextRank算法提取文本关键词集合；最后，根据互信息相似度度量，计算输入文本关键词集和各类关键词集的相似度，最终实现文本的分类。在不同关键词集大小下进行实验，结果表明本文方案在特征集大小为1250左右时分类准确性最高。与现有方法进行的比较表明，本文方案能够提取具有较高识别度的关键词，并具有较高的正确分类率。

参 考 文 献

[1] Parhat R, Meng X T, Hamdulla A. Uyghur Text Sentiment Classification Based on Discriminative Keyword Model[J]. Computer Engineering, 2014, 40(10): 132-136(in Chinese)
热依莱木·帕尔哈提，孟祥涛，艾斯卡尔·艾木都拉. 基于区分性关键词模型的维吾尔文本情感分类[J]. 计算机工程，2014，40(10)：132-136

[2] Maimaitiyiming Hasimu, Wushouer Silamu, Weinila Mushajiang, et al. Research N-gram based Uyghur text classification technique[J]. Application Research of Computers, 2015, 32(7): 1986-1988(in Chinese)
买买提依明·哈斯木，吾守尔·斯拉木，维尼拉·木沙江，等. 基于N元模型的维吾尔文文本分类技术研究[J]. 计算机应用研究，2015，32(7)：1986-1988

[3] Mairehaba·AILI, Jiang Wen-bin, Wang Zhi-yang, et al. Directed Graph Model of Uyghur Morphological Analysis[J]. Journal of Software, 2012, 23(12): 94-100(in Chinese)
麦热哈巴·艾力，姜文斌，王志洋，等. 维吾尔语词法分析的有向图模型[J]. 软件学报，2012，23(12)：94-100

[4] Trstenjak B, Mikac S, Donko D. KNN with TF-IDF based Framework for Text Categorization[J]. Procedia Engineering, 2014, 69(1): 1356-1364

[5] Jayashree R, Srikanta M K, Sunny K. Keyword Extraction Based Summarization of Categorized Kannada Text Documents[J]. International Journal on Soft Computing, 2011, 2(4): 152-164

[6] Alimjan AYSA, Turgun IBRAHIM, Kurban OBUL, et al. Research of Uyghur Language Text Categorization Based on SVM[J]. Computer Engineering and Science, 2012, 34(12): 140-144(in Chinese)
阿力木江·艾沙，吐尔根·依布拉音，库尔班·吾布力，等. 基于SVM的维吾尔文文本分类研究[J]. 计算机工程与科学，2012，34(12)：140-144

[7] Alimjan AYSA, Kurban UBUL, Turgun IBRAHIM. Bigram feature extraction for Uyghur text[J]. Computer Engineering and Applications, 2015, 51(3): 216-221(in Chinese)
阿力木江·艾沙，库尔班·吾布力，吐尔根·依布拉音. 维吾尔文Bigram文本特征提取[J]. 计算机工程与应用，2015，51(3)：216-221

[8] Pawar D D, Bewoor M S, Patil S H. Text Rank: A Novel Concept for Extraction Based Text Summarization[J]. International Journal of Computer Science & Information Technolo, 2014, 34(6): 152-163

[9] Mahpirat Wali, Zhao Meng-yuan, Askar Hamdulla. Keyword based Uyghur single document summarization[J]. Computer Engineering and Applications, 2015, 51(16): 130-135(in Chinese)
买哈铺热提·外力，赵梦原，艾斯卡尔·艾木都拉. 基于关键词的维吾尔单文本自动文摘技术研究[J]. 计算机工程与应用，2015，51(16)：130-135

[10] Turdi TOHTI, Akbar PATTAR, Askar HAMDULLA. Adaptive word grouping algorithm based on mutual information in Uyghur language[J]. Application Research of Computers, 2013, 30(2): 429-431(in Chinese)
吐尔地·托合提，艾克白尔·帕塔尔，艾斯卡尔·艾木都拉. 基于互信息的维吾尔文自适应组词算法[J]. 计算机应用研究，2013，30(2)：429-431

[11] Wang Z, Feng Y. FN-Rank: Domain Keywords Extraction Algorithm[J]. Open Automation & Control Systems Journal, 2015, 7(1): 1347-1351

[12] Li Peng, Wang Bin, Shi Zhi-wei, et al. Tag-TextRank: A Webpage Keyword Extraction Method Based on Tags[J]. Journal of Computer Research and Development, 2012, 49(11): 2344-2351(in Chinese)
李鹏，王斌，石志伟，等. Tag-TextRank：一种基于Tag的网页关键词抽取方法[J]. 计算机研究与发展，2012，49(11)：2344-2351

[13] Litvak M, Last M, Kandel A. DegExt: a language-independent keyphrase extractor[J]. Journal of Ambient Intelligence & Humanized Computing, 2012, 4(3): 377-387

[14] Razlighi Q R, Kehtarnavaz N. Spatial Mutual Information as Similarity Measure for 3-D Brain Image Registration[J]. IEEE Journal of Translational Engineering in Health & Medicine, 2014, 24(2): 27-34

[15] Li Bo, Shi Hui-xia, Wang Yi. A Text Extension Algorithm Based on Synonymy Discovery[J]. Journal of Chongqing University of Technology(Natural Science), 2014, 28(2): 76-81(in Chinese)
李波，石慧霞，王毅. 一种基于同义词发现的文本扩充算法[J]. 重庆理工大学学报(自然科学)，2014，28(2)：76-81