基于K-最近距离方法的哈萨克语报纸分类初探

玛依来·哈帕尔[1]，古丽拉·阿东别克[2]

(新疆大学信息科学与工程学院[1]，新疆大学[2]，乌鲁木齐 830046)

Email:maira109@qq.com

摘要：本文利用K-最近距离的方法对哈萨克语报纸进行分类，初步实现了利用统计词频信息和语言信息相结合的方法选择特征词，且计算特征的权重值时不仅考虑词频，还利用了特征的集中度、分散度，经过训练和统计对哈萨克文文本形成特征的权重向量，之后根据K-最近距离判断测试文本的所属类别，从而实现了本文提出的哈萨克语报纸分类的目标。

关键词：文本分类；哈萨克语；K-最近距离；词频;集中度；分散度

Text Categorization for newspaper of Kazakh Based on K-Nearest Neighbor

Mayra·Hapar[1] ,Gulila·Altenbek[2]

(Information science and engineering college, Xin Jiang University, Urumqi 830046,China)

Email:maira109@qq.com

Abstract: The classification of the Kazakh newspaper is decided by K-nearest-neighbor, a method that integrates language information and statistical information from the training corpus is basically realized. The weight of these characters is computed from three parameters: word frequency, centralized degree, decentralized degree. After training, we get the vector space model of the text categorization, so as to realize the objective of Kazakh newspaper classification which is issued in the article.

Keywords: text categorization; Kazakh; K-nearest-neighbor; word frequency; centralized degree; decentralized degree

1 引言

文本的自动分类是自然语言处理的一个十分重要的问题，是对大量的自然语言文本按照一定的主题类别进行自动分类。文本分类就是在给定的分类体系下，由计算机根据待分类文本的内容自动确定文本类别的过程[1]。文本分类处理的研究是计算机、信息处理领域的重要内容。文本分类主要应用于信息检索，机器翻译，自动文摘，信息过滤，邮件分类等。

词是哈萨克语中最小的能独立运用的语言单位，词与词之间用空格分隔开，不存在像汉语中的分词问题。因此用词、词组、词串为特征单位，基于词的标引是最适合文本分类的。词干是一个词中体现词汇意义的部分。一个词除去词尾所剩部分就是词干。如：“وقۇشى-لار-ىمىز”（我们的学生）中的وقۇشى（学生）是词干，词干=词根+词缀。

在哈萨克语词干提取过程中，除了词干提取以外还要进行构形附加成分的切分。这是因为构形附加成分与词干互相黏连，并且构形附加成分之间也互相黏连。构形附加成分往往可以表示一

定意义[4,5]。所以，如果不将这些黏连在一起的构形附加成分切分开，不能准确的领会整个单词的含义。因此在哈萨克语报纸自动分类研究中词干提取有利于特征词提取的效率。

到目前为止，在自然语言处理领域中，中文信息处理在文本分类方面已经取得了显著的成果。而在哈萨克语文本处理中，哈萨克语文本分类方面的研究还处于起步阶段，报纸分类更是一个新的研究领域。本文着重研究了哈萨克语文本自动分类问题。以及，本文研究开发的系统还实现了对语料的词频统计，词性统计等一定的词法分析功能。

2 K 近邻分类算法

2.1 特征的选择和提取

特征选择和提取的目的在于寻找一个好的特征子集，从而有利于类之间的区别。对于文本分类来说，每一类的特征抽取实际上就是抽取那些能够反映和区分此类文本与它类文本的特征项，这是一切分类问题的关键因素。特征词大多是名词，动词和少数形容词，一般来说，介词、副词、感叹词、冠词、限定词等不可能作为特征词[3,6]。

特征选择依赖于多个指标，常用的指标是词频、集中度、分散度等。在语料中抽取反映某一类特征词时使语言信息和统计信息相结合，过滤掉不可能作为特征词的词性，利用专业词典从中选择能够反映这一类文本的特征词，通过以上两步为每一类选出了侯选词。

2.2 特征的权重计算

上面讨论的选择特征词的方法还没有最后确定特征词，因为没有计算每一侯选词在这一类文本中的权重。一个词相对于一类文本的权重主要包括 3 个指标：词频度，对这一类文本的集中度，对于这一类文本的分散度[1,8]。计算方法如下：

(1)词频度

$$\alpha_{ij} = n_{ij} \tag{2-1}$$

其中，设词 w_i 在类别 c_j 中的出现频度为 n_{ij}。

(2)集中度

$$\beta_{ij} = \left(n_{ij} - m_{ij}\right)^2 / m_{ij}, \quad m_{ij} = \sum_{j=1}^{l} n_{ij} \sum_{i=1}^{k} n_{ij} \Big/ \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} \tag{2-2}$$

其中，l 表示文本的类别数目；k 为 c_j 类的训练集中单词个数，$1 \leqslant i \leqslant k$ ，$1 \leqslant j \leqslant l$。

(3)分散度

$$\gamma_{ij} = \sum_{i=1}^{l} \left(x_{ij}^r\right)^2, \left(x_{ij}^r\right)^2 = \left(y_{ij}^r - m_{ij}^r\right)^2 / m_{ij}^r, \quad m_{ij}^r = \frac{\sum_{i=1}^{k} y_{ij}^r \sum_{r=1}^{t} y_{ij}^r}{\sum_{i=1}^{k} \sum_{j=1}^{t} y_{ij}^r} \tag{2-3}$$

其中，y_{ij}^{r}表示单词i在j类文本的第r篇中出现的频度；k 表示c_j类的训练集中单词个数；t 表示第j类文本的训练篇数，$1 \leqslant r \leqslant t$，$1 \leqslant i \leqslant k$。

(4)每一单词相对于某一类的权重

$$f(w_i, c_j) = p_1\alpha_{ij} + p_2\beta_{ij} + p_3\gamma_{ij} \quad , \quad p_1 + p_2 + p_3 = 1 \tag{2-4}$$

其中，p_1、p_2、p_3分别表示词频、集中度、分散度的系数。本文试验中，词频、集中度、分散度的系数取值都为 1/3。通过以上方法确定了每一类的特征词，并且计算出每一类的特征词其在这一类中的权重，可很容易的构造特征向量空间。

2.3 相关性计算

设类别c_i的特征词向量为$c_i = (t_{i1}, t_{i2}, \cdots, t_{ik}, \cdots, t_{is})$，文本$d_j$的特征词向量为$d_j = (t_{j1}, t_{j2}, \cdots, t_{jk}, \cdots t_{js})$，相似性用特征词向量空间中向量$d_j$和向量$c_i$之间夹角的余弦表示如下[2]：

$$Sim(C_i, d_j) = \frac{\sum_{k=1}^{l_i} t_{ik} \bullet t_{jk}}{\sqrt{\sum_{k=1}^{l_i} t_{ik}^2 \bullet \sum_{k=1}^{l_i} t_{jk}^2}} \tag{2-5}$$

通过相关性计算，可确定文本d_j的类别为在特征词向量空间中与文本向量的相似度最大的那个类别c_m，即：

$$c_m = \arg \max Sim(C_i, d_j) \tag{2-6}$$

3 哈萨克语报纸分类

3.1 哈萨克语报纸类别

本文以《新疆日报》电子版为生语料，从中整理出部分为训练文本。针对“新疆日报”里的文章，本文有以下典型的类别：

（1）新闻类 حابار-وشار

（2）农业类 اؤنل-شارۋاشىلىق

（3）文学类 ادەبیەت

（4）经济类 دکونومیکا

（5）体育类 دەنە-تاربیە

3.2 哈萨克语报纸分类特征词提取

本文按新疆日报哈文版的内容，把语料分为训练集和测试集，训练语料每一类文本各8篇，并且做了停用词表，把训练集中的词和停用词表进行匹配，把训练集中出现的停用词表中的词去除，进而得到了侯选词，之后在此基础上计算出了它们的词频度、集中度和分散度。最后计算出了权重，并根据权重来确定了每一类的特征词。

第一步：首先把 40 篇文章作为训练集，分别计算词频，集中度，分散度。在计算词频时，先读入其中的一篇文章，并且提取单词，之后进行词干切分并计算词频，依次类推计算，我们就会得到 40 个有词干和词频共有的表，其流程图如图 1 所示：

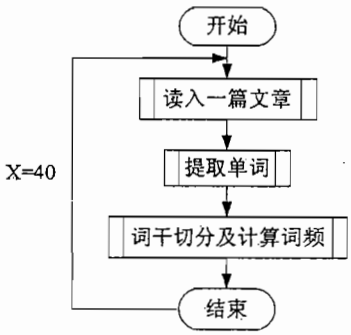

图 1 计算词频α值

第二步：根据所得到的词干切分表，再去计算它们的集中度。先输入前面所得到的每篇文章词干切分α值词表，经过计算可得到同样的单词在所有类中的词频之和，乘一个类中的所有词的词频之和，除以所有类中的所有词的词频之和就得到m_{ij}表，之后把每个单词的词频减去其在m_{ij}表中的值再平方，最后再除于m_{ij}表中的值就可得到词的集中度β_{ij}表。

第三步：计算分散度γ_{ij}时，先计算第j类文本在第r篇中的所有单词的词频之和乘以第i个单词在j类文本所有r篇中的词频之和，除以第j类所有r篇中的所有单词的词频之和就可得到m_{ij}^{r}值表，之后把单词i在j类文本第r篇中出现的频度减去单词i在m_{ij}^{r}表中的值再平方，除以单词i在m_{ij}^{r}表中的值可得到x_{ij}^{r}的平方值表，最后计算所有r篇中的x_{ij}^{r}平方值可得到γ_{ij}值表。

4 实验结果与分析

本文训练集使用语料每类各 8 篇，共 40 篇文章，如表 1 所示：

表 1 实验训练集语料

类型	篇数	词数
经济类	8	3780
农业类	8	2789
体育类	8	2565
文学类	8	4766
新闻类	8	2593

每一类特征词的词频α、集中度β、分散度γ和权重计算结果如下列表所示：

表 2 经济类特征词的词频、集中度、分散度和权重统计表

	词	α	β	γ	权重
经济类	كاسپورتن	2	3.05990882320115	2.7690778670815	36.0966223009422
	شاروا	18	44.3686779364166	108.38923169994	56.9193032121189
	بۇقارا	24	36.7189058784138	26.087022153083	38.9353093438326
	قالاشىق	19	29.0691338204109	24.936833462967	34.3353224277929
	… …	…	…	…	…

表 3 农业类特征词的词频、集中度、分散度和权重统计表

	词	α	β	γ	权重
农业类	ئۇل-شاروا شىنىق	1	8.91737875288683	1.5954641350211	38.3761429596931
	دامۇ	9	668.803406466513	27.7315956029314	235.178334023148
	ئۇل-قىستاق	15	285.356120092379	41.7483677548301	114.034829282403
	توقماشىنىق	2	17.8347575057737	9.4704641350211	97.6840721359827
	… …	…	…	…	…

表 4 体育类特征词的词频、集中度、分散度和权重统计表

	词	α	β	γ	权重
体育类	سپورت	47	126.332992071617	137.395197828859	103.576063300159
	جاريسى	32	86.0139520487607	102.52965279944	73.5145349494002
	دەنە-تاربيە	1	2.68793600152377	5.28804452120514	39.919935075763
	الاك	21	56.4466560319992	69.9592349233228	49.1352969851073
	… …	…	… …	… …	… …

表 5 文学类特征词的词频、集中度、分散度和权重统计表

	词	α	β	γ	权重
文学类	جۇرنال	1	4.8116539349795	32.5289855072464	37.802131474087
	ادەبىيەت	22	211.71277313910	0.182194616977226	155.86924696506
	ئەسەر	1	9.6233078699591	0.00828157349896481	70.8496577113908
	جازۇشى	1	9.6233078699591	0.00828157349896481	70.8496577113908
	……	…	……	……	……

表 6 新闻类特征词的词频、集中度、分散度和权重统计表

	词	α	β	γ	权重
新闻类	ۋىن جيابائو	10	27.0991037346384	35.865267744524	34.321457159721
	لارىا	22	59.6180282162046	91.644513243894	57.754180486699
	زۇشىيى	3	8.12973112039153	12.352195849099	78.273089898304
	شىنخۇا	4	10.8396414938554	19.269544776975	44.36972875694
	……	…	……	……	……

通过计算词频、集中度、分散度后，得到的特征词库如下表所示：

表 7 特征词库表

经济类	农业类	体育类	文学类	新闻类
كاسپورىن	اۇل-شارۋاشىلىق	سپورت	جۇرنال	ۋىن جيابائو
شارۋا	داۋ	جارىسى	ادەبىيەت	لارىا
بۇقارا	اۇل-قىستاق	دەنە-تاربيە	ئەسەر	زۇشىيى
قالاشىق	توقماشىلىق	الاك	جازۇشى	شىنخۇا
……	……	……	……	……

构造特征向量空间时，通过用以上的计算方法得到了每个特征词在其它类中的权重，为了描述方便列出了经济类中的三个特征词在其它类中的词频和权重，如表 8 所示：

表 8 经济类中的词在其他类中的词频和权重表

经济类词	农业类		体育类		文学类		新闻类	
	词频	权重	词频	权重	词频	权重	词频	权重
شارۋا	11	95.98567541662	0	0	0	0	0	0
داۋ	9	235.1783340231	0	0	0	0	13	85.26603668
اۇل-قىستاق	15	114.0348292824	0	0	0	0	0	0

构造特征词向量空间后，进行文本与类别的相关性计算，从而确定文本的类别。评估文本分类系统的两个指标准确率和召回率。准确率是所有判断的文本中与人工分类结果吻合的文本所占

的比率。其数学公式为：

$$准确率(precision)=\frac{分类的正确文本数}{实际分类的文本数} \tag{4-1}$$

召回率是人工分类结果应有的文本中分类系统吻合的文本所占的比率。其数学公式为：

$$召回率(recall)=\frac{分类的正确文本数}{应有文本数} \tag{4-2}$$

用以上的方法对哈萨克语文本进行了训练和测试，训练语料每一类文本各8篇共40篇，253.4KB，共有16493个单词“例”（表示语料库中所有单词的数目），和4719个单词“型”（表示不相同的单词数目），利用训练出来的数据对测试集中的30篇文章进行了分类，正确分类的是22篇。测试结果如表9所示：

表9 测试结果

测试文本篇数	正确分类	准确率	召回率
30篇	22篇	73.3%	73.3%

结束语

本文提出并实现了利用统计词频信息和语言信息相结合的方法来选择哈萨克文特征，计算特征的权重值时不仅考虑词频，还利用了特征的集中度、分散度。经过训练和统计对每一类文本形成特征和权重向量，这样所有类的训练集文本就形成了一个多维的向量空间，然后利用 K-最近距离的方法对测试集进行分类。

参考文献

[1] 孙健，基于K-最近距离的自动文本分类的研究[J]，北京邮电大学学报，2001年3月，p12-14.

[2] 李国臣，文本分类中基于对数似然比测试的特征词选择方法[J]，中文信息学报，1997年.

[3] 刘开瑛等，中文文本中抽取特征信息的区域与技术[J]，中文信息学报，1998年，p1-7.

[4] 冯志伟，孙乐 译，自然语言处理综合[M]，电子工业出版社，2005年6月，p36-54.

[5] 苑春法，李庆中，王昀，李伟，曹德芳等译，统计自然语言处理基础[M]，电子工业出版社，2005年1月，p355-373.

[6] 张晓辉，李莹，王华勇，应用特征聚合进行中文文本分类的改进KNN算法[J]，东北大学学报，2003年 24(3)：p229-232.

[7] 张若峰，基于实例的文本自动分类技术的研究与实现[硕士学位论文]，吉林大学，2005年.

[8] 丁均彦，文本分类系统的研究与实现[硕士学位论文] 北京：清华大学，1998年.

[9] 王丁，运海红，张辉，文本自动分类系统的研究与实现[J]，信息技术 2005年第3期.