山东农业大学学报(自然科学版),2016,47(6):949-952 VOL.47 NO.6 2016
Journal of Shandong Agricultural University (Natural Science Edition) doi:10.3969/j.issn.1000-2324.2016.06.028

基于特征提取算法的辅助机器翻译系统设计与开发

林 杨
内蒙古大学外国语学院, 内蒙古 呼和浩特 010021

摘 要: 随着电子信息技术的发展，如何准确、高效、快捷的将数据分类，已成为当前的热点问题。本文设计了一种基于x^2统计算法和规则判断方法相结合的多特征提取方法，利用该多特征提取算法生成特征词集，采用TF-IDF频率算法生成文本特征向量，使用支持向量机(SVM)分类器模型进行文本分类。并且为分类系统设计了相应的调用接口，保证了该分类模块的可用性。同时还设计了分类词库，保存各个类别的独有特征词，用于优先判断待分类文件的类别。

关键词: 特征提取; 翻译系统; 设计
中图法分类号: TP311.1 **文献标识码**: A **文章编号**: 1000-2324(2016)06-0949-04

Design and Development of the Auxiliary Machine Translation System Based on Feature Extraction Algorithm

LIN Yang
College of Foreign Languages/Inner Mongolia University, Hohhot 010021*, China*

Abstract: With the development of electronic information technology, how to classify the data accurately, efficiently and quickly has become a hot issue. In this paper, a multi feature extraction method based on x^2 statistical algorithm and rule judgment method was designed and implemented the text classification system through using the multi feature extraction algorithm to generate a set of feature words, taking TF-IDF frequency algorithm to generate text feature vectors and using support vector machine (SVM) classifier model to classify text And the corresponding call interface was designed for the classification system, which ensured the engineering availability of the classification module. At the same time, the classified lexicon was designed to save unique feature words in each category for the judgment of priority to be classified document categories.

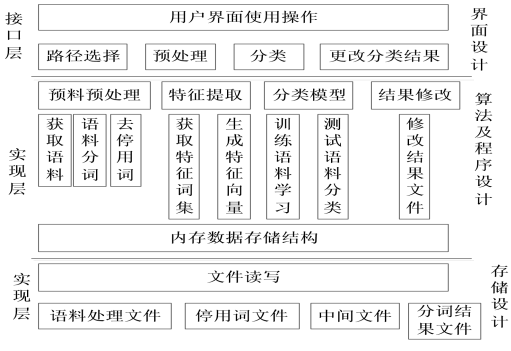
Keywords: Feature extraction; translation system; design

随着电子信息技术的发展，越来越多的人开始接触网络，从网上获取和交流各种信息。随之而来的，就是对海量数据的处理。如何准确高效的从大量数据中找到我们关心的信息已成为当前自然语言处理领域的一大机遇和挑战[1]。对于文本信息，传统的人工分类方法不但耗费大量的人力和时间，而且不同人的标准不同，分类结果一致性低。自动文本分类技术可以帮助人们更高效的实现文本分类，提高了文本分类的实用性[2]。因此，对于文本分类技术的研究，具有重要意义。

1 多特征文本分类方案设计

1.1 架构设计

整个文本分类系统的架构设计如下图 1 所示，对于接口层，用户可通过图形界面进行文本分类；实现层完成了整个文本分类系统的各个流程的算法及程序设计，以及内存数据存储结构设计，并



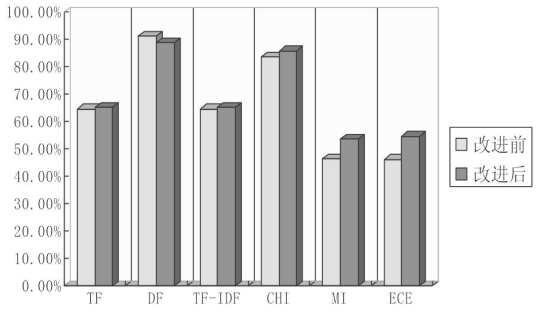


图 1 文本分类系统架构设计图 **图 2 改进后的各特征词集获取算法分类准确率**

Fig.1 Text classification system architecture Fig.2 Improved classification accuracy of each feature word set

收稿日期: 2016-06-19 **修回日期**: 2016-08-18

作者简介: 林 杨(1980-),女,广东新会人,硕士,讲师.主要研究方向为翻译理论与实践. E-mail:linyangimu@163.com

数字优先出版:2016-10-31 http://www.cnki.net

为接口层提供了各个功能的调用接口；数据层为实现层和接口层提供数据支持，通过文件的读写对数据层数据进行操作。

1.2　特征提取设计及改进

1.2.1 *构造数据结构*　对于语料的处理，不但要获取到各个类别中的文件信息，还要保存词的信息。因此通过链表进行存储[3]。对于每个类别中词信息的存储，采用二维链表。通过定义 Head 头结点，然后通过读取文件，依次保存语料各类别中的词信息，便于后续计算，对于每个类别中文件信息的存储，同样采用二维链表。通过定义 File head 头结点，然后通过读取文件，依次保存语料各类别中的文件信息，便于后续计算。

1.2.2 *结果分析*　实验得到各个特征提取算法的结果文件后，统计得到各个特征词集获取算法的判断结果中属于各类别的文章数，通过对准确率、召回率、精确率[4]的计算，得到各算法的文本分类效果如下表 1 所示。

表 1 各特征词集获取算法的准确率、召回率、精确率

Table 1 Accuracy rate, recall rate and precision rate of each feature word set

特征集获取算法 Algorithm	各类别分类结果(正确文章数/错误文章数) Classification result (Right/wrong texts)							准确率 Accuracy	召回率 Recall	精确率 Precision
	8	10	13	14	16	20	24			
TF	10/0	0/10	5/5	6/4	10/0	1/9	10/0	64.62%	60.00%	60.00%
DF	6/4	9/1	9/1	10/0	8/2	10/0	10/0	91.09%	88.57%	88.57%
TF-IDF	10/0	0/10	5/5	6/4	10/0	1/9	10/0	64.62%	60.00%	60.00%
x^2(CHI)	5/5	8/2	9/1	8/2	3/7	10/0	10/0	83.49%	75.71%	75.71%
MI	6/4	1/9	9/1	5/5	3/7	0/10	8/2	46.46%	45.71%	45.71%
ECE	6/4	1/9	9/1	5/5	3/7	0/10	8/2	46.31%	45.71%	45.71%

通过表中各个特征集获取算法的分类结果可以看出，DF 和 x^2 统计算法的分类准确率和精确率相对较高，但分析算法可知，DF 算法不能有效的去除不同类别中的共用词，这往往会导致一些不能用来区分类别的“常用词”被选为特征词，使得分类结果不准确。并且 DF 算法对小语料集的分类效果好，但对大语料集的分类效果较差。因此，对于该系统而言，决定使用 x^2 统计算法作为最终的特征词集获取算法。

1.2.3 *算法改进*　分析系统生成的特征词集文件可以看出，得到的特征词集中包含了大量的无用词，同时，经查看发现，除汉字外，其他字符都需要考虑半码和全码两种形态。在使用 x^2 统计算法得到每个词的开方值并排序后，在函数 Get FE()中添加了规则判断，通过规则，对这些词进行筛选，然后再取每个类别的最多前 1000 个词，生成相应的特征词集[5]。分析改进的特征词集获取方法获得的结果文件，并计算各个算法的分类准确率，得到分类效果如下图 2 所示。通过上述实验，拟采用基于 x^2 统计算法和规则判断方法相结合的多特征提取方法，来获取特征词集。

1.2.4 *本系统的特征提取方法*　基于 x^2 统计算法和规则判断方法相结合的多特征提取方法，能够最大程度的提高系统的分类准确率，因此采用此种方法实现本系统的特征集获取方法。文本特征向量的生成则采用 TF-IDF 频率算法实现[6]。同时，设计了分类词库，保存各个类别的独有特征词，用于优先判断测试文件的类别。

1.3　分类器模型设计

在分类过程中，发现了一个现象：每个类别都有其独有的一些特征词。比如对于测试语料，当“联赛”、“射门”等词在文档中出现时，该文档很有可能属于“体育”类别，而当“航空母舰”、“战斗机”等词反复出现时，该文档则很有可能属于“军事”类别。

针对这种现象，设计了分类词库，保存各个类别的独有特征词。通过各个类别的独有特征词在测试文件中出现的频率，优先判断该文件的类别，作为 SVM 分类方法[7]的补充。

对于一篇输入文本，经过预处理后，得到去停用词后的文件。这时，依次统计各个类别分类词库中的特征词在该文档中出现的次数。当某一类别满足公式 1 时，则直接将该文档设为该类别，否则按照 SVM 结果进行分类。

$$\frac{Count(i)}{Total(N)} > 80\% \tag{1}$$

式 1 中，*Count* (*i*)表示当前类别分类词库中的特征词在该文档中出现的次数，*Total* (*N*)表示所有类别分类词库中的特征词在该文档中出现的总次数。

1.4　界面设计

为了方便用户使用操作，使用 MFC 的 Dialog 对话框为系统设计了简单的用户界面，方便分类操作。界面内容应包含四个模块，分别是路径选择模块、预处理模块、分类模块，以及更改分类结果模块。为了方便用户操作，适应分类的操作流程，故从上到下根据操作流程将界面分成四部分。

当点击生成文件目录模块的“选择目录”、训练模块的“选择语料”以及测试模块多测试文件的“选择语料”时，会弹出文件夹选择窗口。窗口只有选择正确的路径后，确定按钮才会可点。当点击生成文件目录模块的“选择目录”、训练模块的“选择语料”以及测试模块多测试文件的“选择语料”时，会弹出文件夹选择窗口，选择图中只有选择正确的路径后，确定按钮才会可点。在测试模块中选择单测试文件，然后点击“选择语料”时，会弹出文本文档选择窗口，系统设定一次只能选择一个文件。

其中，当依次完成对训练语料的预处理和生成模型，以及对测试语料的预处理和分类后，系统在生成文件目录下生成中间文件，如下图 3 所示。

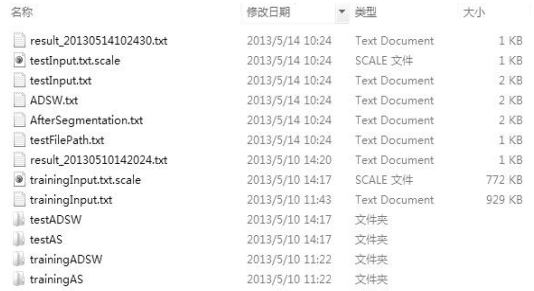


图 3 生成的中间文件

Fig.3 The generated mediate files

图 4 分类结果展示

Fig.4 The classification results

通过该系统实现对文件的分类后，分类结果保存在结果文件中，同时会在界面上实时显示出来，如图 4 所示。分类结果会按照文件名和类别相对应的形式展示出来，方便用户的对比查看。如果想更改分类结果，点击更改分类，在弹出的“修改分类结果”对话框中选择文件和类别，进行结果更改，如图 5 所示。

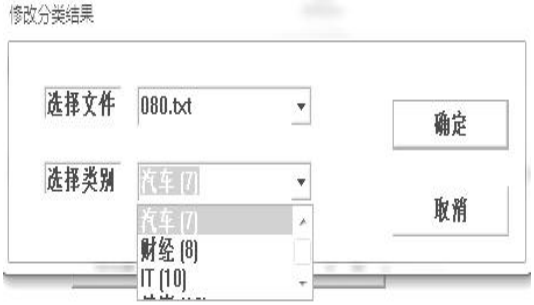


图 5 修改分类结果对话框

Fig.5 The dialog box to modify the classification results

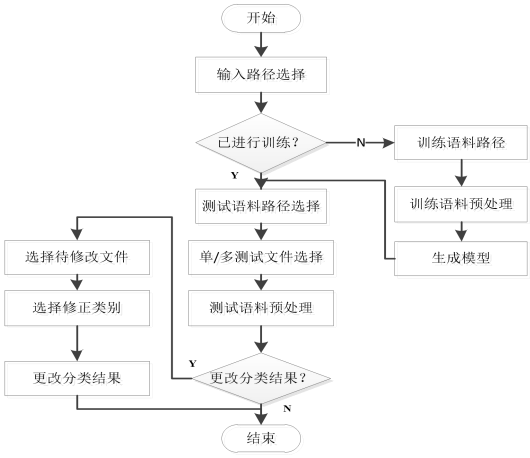


图 6 系统分类流程图

Fig.6 The process of system classification

“选择文件”下拉列表中包括了当前进行分类的所有文件名，“选择类别”下拉列表中包含了所有的预置类别。系统在对测试语料预处理时，将语料中的文件名和路径统一保存在结果目录下的 Test File Path.txt 文件中，将类别预置文件 Category.txt 保存在程序运行目录下，通过文件的读取操作获得

下拉列表中的数据。

2　系统实现及结果分析

2.1　系统流程图

整个文本分类系统得到了完整的实现，用户可通过界面对文本进行分类操作，也可以通过引用链接库，通过分类接口进行相关分类操作。最后，给出整个系统的分类运行流程图，如图 6 所示。

2.2　结果分析

利用系统对语料进行学习并分类，得到分类结果后，通过对准确率、召回率、精确率的计算，得到该系统的文本分类效果，结果如表 2 所示。从表 2 中可以看出，最终系统分类的准确率达到了 89.58%，精确率为 83.19%。从表中可以明显的发现，分类结果中类别 16（旅游）的分类准确率非常低，三次的平均准确率只有 9.23%，从而使得整个系统的分类准确率和精确率相对较低。其次，类别 8（财经）的分类准确率也相对较低，平均只有 70.68%。

分析分类结果文件发现，造成旅游类别和财经类别分类结果不理想的原因是由于训练语料文件中类别重叠造成的。由于语料来源于网页文件，多数语料中的内容都与 IT 类别有交集，涉及到了“网页”、“互联网”等一些 IT 类别中的高频词。因此，如果能够进一步保证语料的质量，使得各类别不出现特征词交叉，采用该系统进行分类的精确率还能得到进一步提升。

表 2 系统分类的准确率、召回率、精确率

Table 2 Accuracy rate, recall rate and precision rate of system classification

测试集 Test	各类别分类结果(正确文章数/错误文章数) Classification result (Right/wrong texts)							准确率 Accuracy	召回率 Recall	精确率 Precision
	8	10	13	14	16	20	24			
1	14/11	25/2	26/6	25/3	3/17	36/0	39/2	88.20%	75.61%	80.38%
2	19/7	26/2	26/6	25/3	1/17	36/0	39/2	89.04%	76.73%	82.30%
3	15/4	22/0	23/2	21/1	1/13	27/0	30/1	91.50%	81.47%	86.88%
合计	48/22	73/4	75/14	71/7	5/47	99/0	108/5	89.58%	77.94%	83.19%

3　结 论

本文通过对文本分类各环节所需理论的探讨和技术算法的研究，主要完成了以下工作：算法设计通过实验对比，设计了基于x^2统计算法和规则判断方法相结合的多特征提取方法，优化了特征词集，有效提高了原单一特征提取算法下文本分类的精确率；分类词库分类设计了分类词库，保存各个类别的独有特征词；设计并实现的文本分类系统，综合分类准确率可以达到83.19%，是一种稳定性强，效率高，且准确率相对较高的实用型文本分类技术；辅助机器翻译系统分类接口设计该系统提供了调用分类功能的接口，作为辅助翻译系统的文本分类模块，保证了程序可应用于辅助机器翻译系统中；分类结果修改功能程序能够对分错类别的文本进行手工修正，通过相应接口，可以更改结果文件中某一文档的分类；分类修正功能程序能够将已分类的测试语料加入训练语料中，进行重新学习，对分类模型进行修正。

参考文献

[1]　奉国和.文本分类性能评价研究[J].情报杂志,2011,30(8):66-70

[2]　宋淑彩,庞　慧,丁学钧.GA-SVM 算法在文本分类中的应用[J].计算机仿真,2011,28(1):222-225

[3]　王法波,许信顺.文本分类中一种新的特征选择方法[J].山东大学学报:工学版,2010,40(4):8-11

[4]　李新福,赵蕾蕾,何海斌,等.使用 Logistic 回归模型进行中文文本分类[J].计算机工程与应用,2009,45(14):152-

[5]　王雪松,程玉虎,郝名林.一种支持向量机参数选择的改进分布估计算法[J].山东大学学报:工学版,2009,39(3):7-10

[6]　张玉芳,彭时名,吕　佳.基于文本分类 TFIDF 方法的改进与应用[J].计算机工程,2006,32(19):76-78

[7]　王元珍,钱铁云,冯小年.基于关联规则挖掘的中文文本自动分类[J].小型微型计算机系统,2005,26(8):1380-1383

[8]　余俊英,王明文,盛　俊.文本分类中的类别信息特征选择方法[J].山东大学学报:理学版,2006,41(3):10-13