文章编号：2096-1472(2020)-04-07-04
DOI:10.19644/j.cnki.issn2096-1472.2020.04.002

基于卫星装配工艺的短文本聚类研究

崔晴洋[1]，梁小峰[2]，倪 静[1]，李 帅[2]，张 生[1]，仲梁维[1]

(1.上海理工大学，上海 200093；
2.航天东方红卫星有限公司，北京 100094)
✉840661306@qq.com;15866585@qq.com;nijing501@126.com;
LS0387@163.com;zhangsheng@usst.edu.cn;zlv@usst.edu.cn

摘 要：为了实现机械手对卫星的自动装配，保证在装配过程中机械手能明确每一步的操作类型。本文主要基于对人工作业的卫星装配工艺规程文件进行文本挖掘，以装配工步内容作为短文本进行操作类型的分类。利用自然语言处理中常用的TF-IDF算法与TextRank算法提取关键字，结合基于装配工艺术语的分级加权方法，构建三种不同的词向量模型与词袋空间。最后使用K-means聚类算法，分别对上述三种方案下的聚类结果进行比较与评估。结果表明，基于装配技术术语的分级加权方案表现最好，平均准确率、召回率、*F*值分别为88.67%、88.71%、88.66%。基于装配技术术语的短文本聚类方法不仅能自动对复杂的操作类型进行自动分类，大大减少了人工干预，而且极大地提升了分类的准确率。

关键词：操作类型；TF-IDF；TextRank；分级加权；K-means
中图分类号：TP391.1 **文献标识码**：A

Research on Short Text Clustering Based on Satellite Assembly Process

CUI Qingyang[1],LIANG Xiaofeng[2],NI Jing[1],LI Shuai[2],ZHANG Sheng[1],ZHONG Liangwei[1]

(1.*School of Mechanical Engineering,University of Shanghai for Science and Technology,Shanghai* 200093,*China*;
2.*Aerospace Dongfanghong Satellite co.,LTD.,Beijing* 100094,*China*)
✉840661306@qq.com;15866585@qq.com;nijing501@126.com;
LS0387@163.com;zhangsheng@usst.edu.cn;zlv@usst.edu.cn

Abstract: In order to realize the automatic assembly of the manipulator to the satellite,the manipulator can specify the operation type of each step in the assembly process.This paper is mainly based on the text mining of manual satellite assembly process documents and classifies the operation types with the assembly step content as the short text.Keywords were extracted by TF-IDF and TextRank algorithms commonly used in natural language processing.Three different word vector models and word pocket spaces were constructed by combining the hierarchical weighting method based on assembly technology terms.Finally,the K-means clustering algorithm is used to compare and evaluate the clustering results under the above three schemes.The results showed that the grade-weighted scheme based on assembly technical terms had the best performance,with average accuracy,recall rate,and *F* value of 88.67%,88.71%,and 88.66%,respectively.The method based on assembly technical terms can automatically classify complex operation types,reducing manual intervention,and significantly improve the classification accuracy.

Keywords:operation type;TF-IDF;textrank;hierarchical weighting;K–means

1 引言(Introduction)

随着计算机技术与物联网的飞速发展，人工智能在每个领域的地位也显得越来越有分量，各行各业的人士对人工智能的探索也在不断加深[1]。自然语言处理作为人工智能的一个分支，也正在很多的地方发光发热。它主要是以电子计算机，编程语言作为工具对人类特有的书面和口头形式的自然语言信息进行各种类型的处理和加工的技术，是一门涉及语言学、计算机科学和数学的交叉性学科。

卫星的零部件装配都是在狭小空间中进行，装配存在着视野受限、目标位置不可见等问题，因此采用一种基于双目视觉定位的机器人辅助装配路径规划方法，通过机械手实现自动装配。机械手需要在原有的装配工艺规程文件中获取工艺信息来完成不同的装配动作，因此基于对原工艺规程文件的装配工步操作类型的划分至关重要。

2 文本预处理(Text preprocessing)

针对原有的卫星工艺规程文件，由于其文件格式复杂，内容繁多，而我们只需要对其工步内容进行分类，因此选择对工步内容进行单独提取，并将提取得到的工步内容分行存储进文本中，此时的工步内容杂乱无序，很难进行归类，此时需要进行文本预处理，这里的操作包括了对文本的分词，去除停用词，利用TF-IDF算法与TextRank算法提取关键词。

2.1 分词处理

对于存储进文本中的工步内容进行分词处理，由于属于中文短文本，这里选用了Python中的jieba分词组件对文本内容进行分词。该组之间具有三种分词模式：精确模式、全模式和搜索引擎模式，这里由于做的是对装配操作的分类，只需要对已有的内容进行最精准的切分，不需要扩充词语，因此选用精确模式来切词。

2.2 去除停用词

在分词完成之后，所有的工步内容都被划分为一个个的词组，词组是可以表达中文文本语义的最好的形式。但在每一条由词组组成的工步中，存在着很多对语义的表达没有任何代表性的词，如“！”“，”“【】”这一类的符号，又如“的”“了”“啊”之类的助词，还有一些连词，这都属于停用词，这里引用了中文停用词词典，去除了分词完的文本中存在于停用词词典中的词，重新储存为文本格式，这样就使得每一个工步中的词组更加具有代表性，之后就需要提取关键词赋予一定的权值。

2.3 TF-IDF提取关键词

TF-IDF算法是关键词提取算法中的一种十分有效且简单的算法，并且效果较为显著，它的主要原理是用$tf_{i,j}$，即给定的词语i在第j文档中出现的频率乘以idf_i，即词语i的逆文档频率，用总文档数目$|D|$除以包含指定词语的文档数目$|\{j:t_i\in d_j\}|$，再将得到的商取对数实现[2]，计算公式如式(1)：

$$TF-IDF=TF_{ij}\bullet IDF_i=\frac{n_{ij}}{\sum_k n_{kj}}\bullet\log\frac{|D|}{|\{j:t_i\in d_j\}|} \tag{1}$$

通过TF-IDF算法提取关键词权值最大的前五个，详见表1。

表1 TF-IDF算法提取关键词及权值

Tab.1 TF-IDF algorithm extracts keywords and weights

关键词名	权值
安装	0.548
螺钉	0.501
支架	0.282
设备	0.240
绝缘	0.206

2.4 TextRank提取关键词

TextRank是一种基于图排序的算法，主要是通过把文本分割成若干词组并建立图模型，利用投票机制对文本中的重要成分进行排序，仅仅利用单篇文档本身的信息就可以实现关键词的提取，做文摘[3]。由于这里估计的是工步中每个词组的重要性，因此我们假设每个词的连接权重都为1，则可得如公式(2)：

$$WS(V_i)=(1-d)+d\bullet\sum_{V_j\in\ln(V_j)}\frac{1}{|Out(V_j)|}WS(V_j) \tag{2}$$

其中，d表示阻尼系数，一般为0.85，V_i表示图中的任一节点，$\ln(V_i)$表示指向顶点V_i的所有顶点集合。$|Out(V_j)|$表示由顶点V_j连接出去的所有顶点的集合个数。$WS(V_i)$表示顶点V_i的最终排序权重。

通过TextRank算法提取关键词权值最大的前五个，详见表2。

表2 TextRank算法提取关键词及权值

Tab.2 TextRank algorithm extracts keywords and weights

关键词名	权值
安装	1.0
螺钉	0.621
设备	0.573
支架	0.398
绝缘	0.365

3 构建向量空间(Construct vector space)

在利用TF-IDF和TextRank分别取得所有装配工步内的前30个关键词和权重值之后，我们就需要将文本进行向量化了，向量化的每一个工步内容所组合而成的向量空间就可以作为文本聚类算法的输入项参与到分类的工作中。

3.1 文本向量化

对于每一条工步内容，可以视作一个由若干个具有语义代

表性的词组组合而成的短文本，这里采用了基于词频的计数向量构造方法来初始化这一个文本向量，简而言之，对于一个工艺短文本，例如：“1.检查设备表及热敏电阻表面状态是否完好。”，通过分词与去除停用词的处理之后产生的词组包括了“检查/设备/ 表面/热敏电阻/表面/状态/完好”，此时就可以构建该文本向量的初始化状态，详见表3。

表3 文本计数向量

Tab.3 Text count vector

词组	检查	设备	表面	热敏电阻	状态	完好
频次	1	1	2	1	1	1

向量[1,1,2,1,1,1]就可以用来表示该文本的初始化状态。同样的按照这样的方式创建每一个文本的计数向量。

3.2 生成词袋模型

根据上述生成计数向量的方法，将所有的向量累积起来，创建一个包含了以所有计数向量的词组作为特征值的初始化词袋空间模型[4]，如式(3)：

$$VSM=\begin{pmatrix} v_1 & \cdots & v_n \\ a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \tag{3}$$

在该模型中，v_i代表了特征，即划分出来的词组，而a_{ij}代表的是在语料库中，第i个短文本第j个位置的词组的频次，如果不存在这个词组，则a_{ij}为0。这样构造出的词袋模型没有增加权值，因此存在的问题是每个词相较于其他词都具有一样的代表性。这时我们就需要发挥TF-IDF和TextRank提取的关键词的作用，对前30个关键词进行加权处理，全面提升这30个关键词在空间向量中所占有的地位，构建加权矩阵，将词袋模型VSM乘以加权后矩阵W，所得到的加权后的词袋模型VSMend就是最终聚类的输入项，如下式：

$$W=\begin{pmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_n \end{pmatrix} \tag{4}$$

$$VSM_{end}=VSM*W=\begin{pmatrix} v_1 & \cdots & v_n \\ a_{11}\cdot w_1 & \cdots & a_{1n}\cdot w_n \\ \vdots & \ddots & \vdots \\ a_{m1}\cdot w_1 & \cdots & a_{mn}\cdot w_n \end{pmatrix} \tag{5}$$

3.3 分级加权法

上述通过两种关键词提取的方法来对词袋模型进行加权，但其实都存在缺陷。对TF-IDF来讲，短文本的词频通常来讲不会太高，并且文档数目较少，这会导致大多数情况下提取关键词的表现不是很好[5]。TextRank则会对文本中多次出现的词赋予更大的权重，这会导致一些可能没有被停用词去除的连词具有较大的权值，如果无法做到非常准确的词性过滤。也会导致该算法的表现力下降[6]。

这里探究一种新的加权方法，既可以将关键词的代表性突显出来，又可以不让大频次的无用词出现。结合机械工艺装配的专业知识，了解到关于机械装配方面和其他专业一样存在专业术语，在工艺规程文件中，往往装配工艺术语可以最好地代表短文本的内容。进一步按照研究的目的将术语分为动词，名词两种。显然由于要对操作类型进行分类，动词的权重肯定大于名词。

按照上述思路，从网上下载机械装配工艺术语，并将其分成动词术语库、名词术语库，将加权矩阵分为三级，若特征属于动词术语库则赋予一级因子权值，若特征属于名词术语库则赋予第二级因子权值，其他则赋予三级因子权值[7]。例如：“紧固”，为一级因子，赋予最大权值，“螺钉”为二级因子，赋予第二权值，“人员”为三级因子，赋予最小权值，此时加权矩阵W由三级因子权值组成，如图1所示。

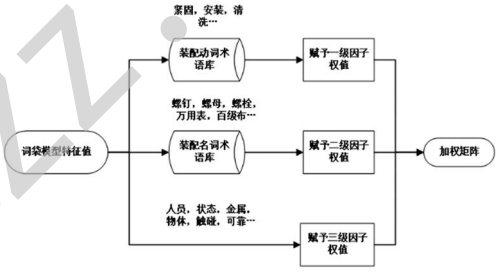

图1 运用分级加权法生成加权矩阵

Fig.1 Generate a weighting matrix using hierarchical weighting

在规划完分级加权大致的流程之后，下一步需要考虑加权规则等细节问题，由于初始化的词袋模型为每一个计数向量累加而得到的，对于一些代表性不强的而在每一个文本中出现次数频繁的词组，即三级因子，可能会导致它的权值在一开始就可能是一级因子或二级因子的n倍，如果直接进行加权，可能会导致一级因子与二级因子的加权效果变得不够显著甚至小于三级因子，从而无法达到突出关键词的效果。在这里，我们引入了一个约束来限制三级因子权值可能过大的问题。每一级因子的权重系数原理如式(6)：

$$w_{mn}=\frac{(w_1,w_2,w_3)\bullet\ln(c_{mn}+1)}{c_{mn}} \tag{6}$$

这里的w_{mn}是加权矩阵W上的第m行n列的权重系数，c_{nm}代表的是每一行计数向量对应该特征值的频次，w_1、w_2、w_3则是通过名词装配术语库，动词装配术语库筛选所得因子的权重值，这里从一级到三级赋值5、3、1，该系数主要是先将计数向量中的频次乘以倒数从而抹除所有向量受频次影响的权重比例，之后乘以$\ln(c_{mn}+1)$重新提升频次的影响力，但由于在短文本的环境下，不会出现频次大的出奇的情况，因此相对于w_1、

w_2、w_3的权值而言，大大减小了频次的影响力，这也大大提高了分级加权的影响力。此时，加权矩阵W不再是一个对角矩阵，构建词袋模型也由原来的矩阵相乘变为矩阵点乘的形式，如式(7)：

$$VSM_{end} = VSM.*W = \begin{pmatrix} v_1 & \cdots & v_n \\ a_{11}\cdot w_{11} & \cdots & a_{1n}\cdot w_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1}\cdot w_{m1} & \cdots & a_{mn}\cdot w_{mn} \end{pmatrix} \tag{7}$$

4 文本聚类(Text clustering)

4.1 K-means聚类分析

K−means聚类算法是划分法中比较经典的算法，可以高效准确地对庞大的数据进行聚类。K−means算法的逻辑主要是确定k各初始的点作为质心，然后将数据集中的每个点分配到一个簇中，为每个点找距离最近的质心，并将其分配给该质心对应的簇。完成之后，每个簇的质心更新为该簇所有点的平均值[8]。迭代上述过程至质心不再发生变动。

将上述TF−IDF算法、TextRank算法，以及分级加权法加权得到的词袋模型分别输入到K−means聚类算法中进行聚类。这里K值得选择根据两种方法获取。①基于平均离差得肘部方法选择，②基于轮廓系数的分数评价[9]，如图2所示。

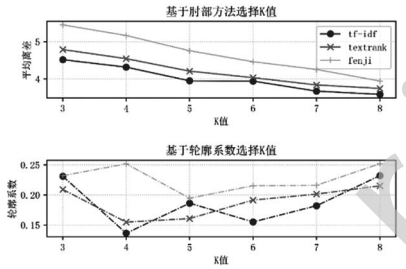

图2 两种确定K值的方法综合分析折线图

Fig.2 Two methods to determine the value of K were used to analyze the line graph comprehensively

根据工艺规程文件内容，K的选取应该在3到8种，即存在3到8种的操作类型，分类太少肯定达不到分类的效果，分类太多可能效果显著，但很多相同的操作类型可能会因为主内容不相似而被分成了几类，也不符合实情。如图2所示，当K值为8时，三种方法的轮廓系数最高同时平均离差最小，所以我们选择K值为8。在这之后，通过人工分类的方法将装配工艺内容正确分出八类。这样就可以用这三种K−means的聚类结果与实际结果进行比较。

4.2 K-means聚类评估

对于通过三种不同的加权方式聚类得到的八个类别，我们分别使用准确率、召回率、F值来来进行对比评价，在这里准确率即为每一类中预测正确的操作类型数量N_t与全部文本数量N_T的比值[10]，如式(8)：

$$precision = N_t / N_T \tag{8}$$

召回率为预测正确的操作类型数量N_t与原类样本总数N_m的比值，如式(9)：

$$recall = N_t / N_m \tag{9}$$

F值的计算公式为：

$$F = \frac{2\bullet precision\bullet \mathrm{recall}}{precision + recall} \tag{10}$$

如图3所示为八个类别的准确率、召回率、F值的折线图，如图4所示为三类标准平均值的柱状图。

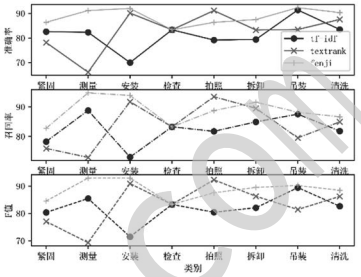

图3 基于准确率、召回率、F值的聚类评估折线图

Fig.3 Clustering evaluation line graph based on accuracy, recall and F value

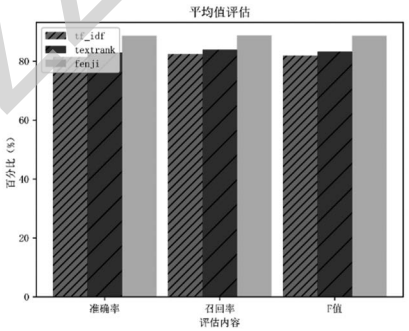

图4 基于准确率、召回率、F值的平均值柱状图

Fig.4 Bar chart based on the average of accuracy rate, recall rate,and the value of F

由图4可以明显地看出基于分级加权法的K−means聚类大体上相较于其他两种关键词提取方法，在准确率、召回率上面都有所提升，分别为88.67%、88.71%，同时在综合评定的F值上也有较大的提高，为88.66%。这意味着通过这种方法可以更好地对装配工艺的操作类型进行自动分类，提升了分类的精确度，为后续自动装配的工作做出了贡献。

5 结论(Conclusion)

本文基于卫星装配工艺规程文件，按照常规的短文本聚类步骤，采用TF−IDF、TextRank关键词提取加权和基于机械装配术语库的分级加权法，三种方法确定特征的权重系数，生成词袋模型[11]，之后采用K−means聚类进行对机械手操作类型的分类，评估结果发现基于专业术语库的加权方法对于这种专业性较强的短文本聚类效果更佳。本研究着力在加权方法上进行研究，目的就是增强关键词的代表性，实现高效聚类。

参考文献(References)

[1] 张国锋,吴国文.基于核函数的改进k-means文本聚类[J].计算机应用与软件,2019,36(9):281-284;301.

[2] 王露瑶,张涛,陈才,等.基于卡方统计改进的TF-IDF的文本分类的研究[J].电子世界,2019,(6):24-25;28.

[3] 周锦章,崔晓晖.基于词向量与TextRank的关键词提取方法[J].计算机应用研究,2019,36(4):1051-1054.

[4] 薛苏琴,牛永洁.基于向量空间模型的中文文本相似度的研究[J].电子设计工程,2016,24(10):28-31.

[5] 张莉婧,李业丽,曾庆涛,等.基于改进TextRank的关键词抽取算法[J].北京印刷学院学报,2016,24(4):51-55.

[6] 徐馨韬,柴小丽,谢彬,等.基于改进TextRank算法的中文文本摘要提取[J].计算机工程,2019,45(3):273-277.

[7] 蔡志川,李运怀.基于分级加权法的城镇垃圾填埋场选址评价[J].地质学刊,2019,43(2):341-348.

[8] (美)哈林顿(Harington,P.).李锐,译.机器学习实战[M].北京:人民邮电出版社,2013.

[9] (美)加文·海克(Gavin Hackeling).scikit-learn机器学习[M].张浩然,译.北京:人民邮电出版社,2019.

[10] 曹晓.文本聚类研究综述[J].情报探索,2016(01):131-134.

[11] (美)爱丽丝·郑(Alice Zheng).陈光欣,译.精通特征工程[M].北京:人民邮电出版社,2019.

作者简介：

崔晴洋(1996-)，男，硕士生.研究领域：计算机辅助设计与智能制造.

梁小峰(1981-)，男，硕士，高级工程师.研究领域：机械设计.

倪　静(1972-)，女，博士，副教授.研究领域:信息系统.

李　帅(1985-)，男，博士，工程师.研究领域：机械工程.

张　生(1968-)，男，学士，高级工程师.研究领域：计算机应用.

仲梁维(1962-)，男，硕士，教授.研究领域：计算机辅助设计，企业信息化.本文通讯作者.

(上接第21页)

研究，能够提出相应的切实有效地修补建议，可以切实提高网站的自身安全级，并具有重要的实用价值，促进Web的应用安全，拥有较为可观的市场发展前景。

本作品目前还存在一些改进的空间，比如稳定性还可以进一步加强，可设计为浏览器插件以方便进一步使用等。

参考文献(References)

[1] Micro Focus.动态应用安全测试(DAST):Fortify WebInspectWebInspect[EB/OL].https://www.microfocus.com/zh-cn/products/Webinspect-dynamic-analysis-dast/overview,2020-01.

[2] Acunetix.Web Application Security Scanner[EB/OL].https://www.acunetix.com,2020-01.

[3] 陈禹.Web漏洞扫描器一览[J].计算机与网络,2016,42(20):56-57.

[4] 尹彦涛.Web漏洞扫描系统设计与实现[D].中国海洋大学,2014.

[5] 张烨青.Web应用安全漏洞扫描器爬虫技术的改进与实现[D].北京邮电大学,2014.

[6] Manohar,E.,Shalini Punithavathani,D..Hybrid Data Aggregation Technique to Categorize the Web Users to Discover Knowledge About the Web Users[J].Wireless Personal Communications,2017(4):1-6.

[7] Wenfei Fan,Jeffrey Xu Yu,Jianzhong Li.Query translation from XPath to SQL in the presence of recursive DTDs[J].The VLDB Journal,2009,18(4):2-4.

[8] Jevri Tri Ardiansah,Aji Prasetya Wibawa,Triyanna Widyaningtyas.SQL Logic Error Detection by Using Start End Mid Algorithm[J].Knowledge Engineering and Data Science,2017,1(1):1-11.

[9] Steinhauser,Antonin,Tuma.DjangoChecker:Applying extended taint tracking and server side parsing for detection of context-sensitive XSS flaws[J].Software:Practice and Experience,2019-Wiley Online Library,2019(1):1-6.

[10] 周开东,魏理豪,王甜,等.远程文件包含漏洞分级检测工具研究[J].计算机应用与软件,2014(2):21-23.

作者简介：

徐贵江(1999-)，男，本科生.研究领域：网络安全，WEB系统开发.

黄媛媛(1999-)，女，本科生.研究领域：信息安全，软件开发.

陈子豪(1997-)，男，本科生.研究领域：信息安全，软件开发.

殷旭东(1970-)，男，硕士，工程师，实验师.研究领域：网络安全，移动计算.本文通讯作者.

钱振江(1982-)，男，博士，副教授.研究领域：信息安全，信息物理融合系统和定理证明.