计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190700146

融合自注意力机制和多路金字塔卷积的软件需求聚类算法

康 雁 崔国荣 李 浩 杨其越 李晋源 王沛尧
云南大学软件学院 昆明 650091
(kangyan@ynu.edu.cn)

摘 要 随着软件数量的急剧增长以及种类的日益多样化,挖掘软件需求文本特征并对软件需求特征聚类,成为了软件工程领域的一大挑战。软件需求文本的聚类为软件开发过程提供了可靠的保障,同时降低了需求分析阶段的潜在风险和负面影响。然而,软件需求文本存在离散度高、噪声大和数据稀疏等特点,目前有关聚类的工作局限于单一类型的文本,鲜有考虑软件需求的功能语义。文中鉴于需求文本的特点和传统型聚类方法的局限性,提出了融合自注意力机制和多路金字塔卷积的软件需求聚类算法(SA-MPCN&SOM)。该方法通过自注意力机制捕获全局特征,然后基于多路金字塔卷积从不同窗口的通路深度挖掘需求文本特征,使得感知的文本片段逐倍增加,最终融合多路文本特征,利用 SOM 完成聚类。在软件需求数据上的实验表明,所提方法能较好地挖掘需求特征并对其聚类,性能上优于其他特征提取方式和聚类算法。

关键词:需求分析;文本聚类;自注意力机制;金字塔卷积;文本特征
中图法分类号 TP309

Software Requirements Clustering Algorithm Based on Self-attention Mechanism and Multi-channel Pyramid Convolution

KANG Yan, CUI Guo-rong, LI Hao, YANG Qi-yue, LI Jin-yuan and WANG Pei-yao
College of Software, Yunnan University, Kunming 650091, China

Abstract With the rapid increasing in the number of software and the increasing variety of types, how to mine the text characteristics of software requirements and cluster the characteristics of software requirements has become a major challenge in the field of software engineering. The clustering of software requirements texts provides a reliable guarantee for the software development process while reducing the potential risks and negative impacts of the requirements analysis phase. However, the software requirements text has the characteristics of high dispersion, high noise, and sparse data. At present, the work related to clustering is limited to a single type of text, and the functional semantics of software requirements are rarely considered. In view of the characteristics of the demand text and the limitations of the traditional clustering method, this paper proposed a software demand clustering algorithm (SA-MPCN&SOM) combining the self-attention mechanism and multi-channel pyramid convolution. The method captures the global features through the self-attention mechanism, and then extract the required text features from the depth of the different windows based on multi-channel pyramid convolution. Thus, the perceived text fragments are multiplied, and finally the multiplexed text features are clustered using SOM. The experimental results on the software demand data show that the proposed method can better mine the demand features, cluster the demand features, and outperform other feature extraction methods and clustering algorithms.

Keywords Demand analysis, Text clustering, Self-attention, Pyramid convolution, Text feature

1 引言

近年来,大数据和人工智能备受追捧,热度空前。在这期间,软件开发效率和质量得到极大改善,在软件开发过程中挖掘需求文本特征并对其进行聚类对促进软件开发和提升软件质量具有重要意义。众所周知,决定软件质量的关键在于需求分析是否明确。需求分析贯穿整个项目的生命周期,决定软件开发的成败。因此,提取需求文本的特征并对其进行聚类在软件工程领域是一项具有挑战性的工作。需求文本为软件设计提供指导,其主要目的是清楚地了解用户希望的软件功能和性能[1]。研究表明,需求缺陷发现得越晚,移除缺陷或者修复缺陷所需要的成本就会越高[2-4]。现有解决软件缺陷

到稿日期:2019-07-22 返修日期:2019-10-26 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61762092,61762089);云南省软件工程重点实验室开放基金项目(2017SE204)
This work was supported by the National Natural Science Foundation of China (61762092, 61762089) and Yunnan Provincial Key Laboratory of Software Engineering Open Fund Project (2017SE204).
通信作者:崔国荣(506527043@qq.com)

的方式大部分是规避风险，例如，利用 BP 神经网络的软件需求分析风险评估模型[5]，通过确定风险级别来减小负面影响。大多数软件需求风险评估技术本质上都是基于主观的，基于灰色聚类评估的软件需求风险模型[6]为定量、准确评价软件需求风险提供了有效的方法。目前，大部分需求分析都是基于风险评估的，鲜有对需求文本的特征提取进行研究，拥有一个好的语料库才可以达到好的聚类效果，才能为软件开发指明设计方向。

需求是一个确定和理解不同用户类的需要和限制的过程。软件需求源于实际，更具有客观性和易变性，需求信息会随着时间的推移而不断改变或拓展，用户在此期间难以给出非常准确和完美的需求描述；而且用户对需求的描述中通常包含一些模糊和不全面的信息，不能清晰描述系统边界，导致开发人员无法清晰了解用户的需求。软件需求数据存在维度高、噪声大和数据稀疏等缺点，国内外对需求文本的聚类研究举步维艰。本文主要采用无监督的聚类方法进行需求获取，所提算法可以处理海量的、零散的需求信息，提供客观的分类结果，缩短需求分析时间，减少设计开发过程中的障碍，方便后续的需求验证和需求变更处理。本文提出的需求获取方法不仅适用于企业内部系统，更适用于用户群体众多、用户量巨大的应用系统。

本文提出了一种基于自注意力机制的多路金字塔聚类算法，规避了开发者在软件设计过程中所面临的风险，为开发指导提供了很大的帮助。本文的主要贡献如下：

(1)国内外对软件需求文本的聚类研究尚少，本文将对软件需求领域的文本聚类展开研究；

(2)使用 BERT[7]词向量预训练软件需求文本，考虑上下文关系，学习词语之间的语义信息；

(3)使用融合 Self-Attention[8]的多路金字塔神经网络进行文本特征提取，提高对全局文本特征的提取能力；

(4)将多路金字塔网络模型与 SOM[9]组合，实现对软件需求文本的相似性聚类；

(5)通过多种不同的聚类评价指标来证明本文算法的可行性和有效性。

2　相关研究

在软件开发过程中，几乎所有的项目都需要有需求文本，软件需求文本有助于开发人员高效、快速地进行设计，但从复杂的需求文本中提取有价值的信息较为困难。一个软件拥有多种不同的需求，如何将相似性需求聚类对于诸多学者来说都是不小的挑战。

2.1　需求文本聚类研究

Martin 等[10]的研究成果表明，软件产品的大部分故障来源于需求阶段。根据需求文本将功能相似性语句聚类，得到不同的聚类结果。在软件开发过程的后期融合新需求，是十分困难的。文献[11]利用软件需求描述中的语义信息获取问题域的高层逻辑，通过对系统代码进行动态分析来辅助程序聚类，有利于程序复用。文献[12]采用自然语言处理技术提取服务需求中的所有有用功能信息集，根据服务功能信息集度量服务的功能语义相似度，使用 K-means 算法实现服务聚类。文献[13]提出了基于深度学习卷积神经网络的短文本聚类算法，对短文本进行神经网络特征提取，再由传统聚类算法对文本特征进行聚类。

需求文本聚类已由算法改进逐渐改变为特征提取的改进，只有好的语料才能达到好的聚类目标。聚类的结果不仅便于需求分析人员进行需求审查，也便于软件开发人员进行软件编程，同时便于维护人员对软件系统进行修改和变更[14]。

2.2　文本聚类方法的研究

一般的文本聚类算法都是先通过 Word2vec[15]预训练词向量，然后直接由传统聚类算法进行划分。但是，在特征提取过程中，我们应该考虑序列的内部关系和语义信息，直接由 Word2vec 训练的方法未考虑上下文关系。传统的聚类算法及其改进都是基于划分方法和层级方法的，如 K-means[16]，Agglomerative Clustering[17]等。尽管这些算法都在不同领域数据上取得了显著的效果，但这并不意味着这些聚类算法具有普适性，能满足所有的领域知识聚类；同时，在实际应用中，文本和词汇的海量性，以及单个文本在整个向量空间上分布的稀疏性，造成现有算法很难准确找到基于子空间的文本类别[1]。为解决维度灾难问题，PCA[18]和 LDA[19]将高维空间向量映射成低维空间向量，然后基于低维空间向量聚类，但该降维方法适用于线性映射，结果缺乏合理的解释和描述。

3　SA-MPCN&SOM 软件需求聚类

3.1　SA-MPCN&SOM 算法模型

文本聚类结果的好坏不仅取决于文本特征的提取，还依赖于聚类算法对数据的适用性。为提高需求文本特征提取和聚类的效率，本文提出了融合自注意力机制和多路金字塔卷积的软件需求聚类算法模型——SA-MPCN&SOM(Self-Attention Multi-Channel Pyramid Convolution Network and Self Organization Map)，其通过改善特征提取方式，自组织自适应聚类。本文首先对软件需求文本进行分词、去停用词和去标点符号的处理，然后通过 BERT 词向量工具将其训练成词向量形式，利用 Self-Attention 提取注意力特征图，再用多路金字塔卷积网络(Multi-Channel Pyramid Convolution Network)完成文本特征的提取，最终由 SOM(Self Organization Map)完成聚类。本文算法的模型如图 1 所示。

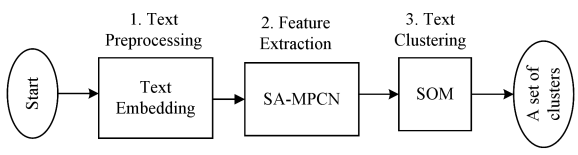


图 1　SA-MPCN&SOM 模型

Fig. 1　SA-MPCN&SOM model

3.2　文本预处理

本文将获得的软件需求语料进行 NLTK 分词、去停用词处理后，利用词向量工具将原始样本映射到向量空间上。鉴于 Word2vec 的训练未考虑语序对句子的影响，句子不能拥有一个整体含义，本文使用 BERT[7]来完成对需求文本的向量转换。文本预处理模型如图 2 所示。

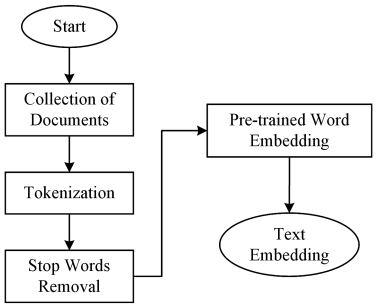


图 2 文本预处理

Fig. 2 Text preprocessing

BERT 是第一个用在预训练 NLP 上的无监督的深度双向系统。由于 Bi-LSTM 结构随着模型深度的增加，整个句子形成前后语义自我包含的循环，BERT 模型采用 Masked LM 打破标准语言模型的单向性局限，随机覆盖输入句子的单词，并基于覆盖词周围的语境预测覆盖词[7]。

在需求文本的预处理中获取语料中所有短文本的最大长度和不重复单词的数量，然后补充每个短文本至最大长度，以便于神经网络的批处理操作。BERT 对词表进行编码，将每个单词转换成 k 维向量，则将每个需求文本表示成 $n*k$ 维向量。这一向量充分保留了短文本中的语义信息和位置信息，因此本文用这一向量作为短文本的原始特征向量。

3.3 基于 SA-MPCN 的特征提取

无论是分类还是聚类，只有拥有好的语料才能达到理想的效果。针对软件需求文本具有隐含性、数据稀疏、噪声大、新词频繁出现等特点，传统聚类算法仅通过预训练的词向量无法达到理想的聚类效果。对此，本文提出了一种新的特征提取方式——SA-MPCN。序列经过 SA-MPCN 完成文本特征提取后，相邻词语间的信息得到了更好的提取和融合。SA-MPCN 模型如图 3 所示。

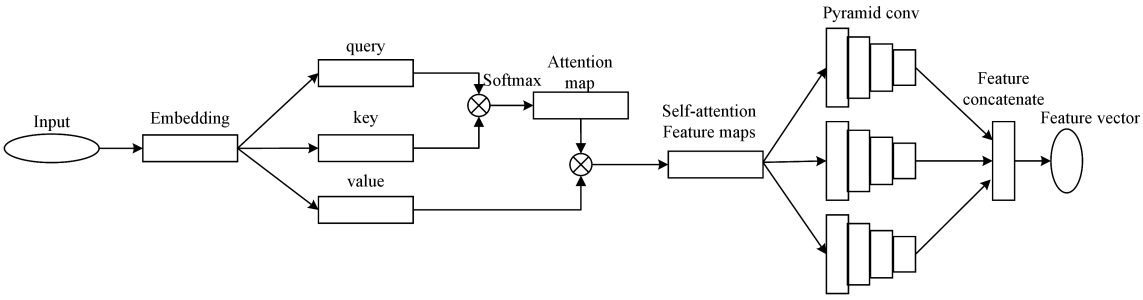


图 3 SA-MPCN 模型

Fig. 3 SA-MPCN model

在序列中，每个单词对上下文单词的影响程度不同，对整个序列的语义信息贡献也不同。本文在输入层融合 Self-Attention，旨在学习句子内部的词依赖关系，捕获句子的内部结构。假设输入序列为 $x=[\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\cdots,\boldsymbol{x}_m]$，每个序列的长度为 n，每个词的维度为 d，Self-Attention 的输出维度为 d'，其查询向量序列、键向量序列和值向量序列的输入如下：

$$Q_i,K_i,V_i=x_i \tag{1}$$

其中，查询向量序列、键向量序列和值向量序列都来自同一个序列输入。对这 3 个向量做线性转换：

$$Q_i'=Q_i*W_Q \tag{2}$$

$$K_i'=K_i*W_K \tag{3}$$

$$V_i'=V_i*W_V \tag{4}$$

查询向量、键向量和值向量经过式(2)－式(4)的线性转换后，序列长度保持不变，单词维度由 d 转换成 d'，其中 W_Q，W_K，W_V 是依据 Q_i，K_i，V_i 随机生成的权值。接下来计算 query 和 key 的相似度，为防止结果过大，将 Q_i' 与 $K_i'^T$ 点乘后除以一个尺度标度 $\sqrt{d_k}$，d_k 代表 query 和 key 向量的维度，再利用 softmax 操作将其结果归一化为概率分布：

$$\alpha_i=softmax\left(\frac{Q_i'*K_i'^T}{\sqrt{d_k}}\right) \tag{5}$$

其中，α_i 就是 x_i 序列对应的 attention 权重，将其乘以矩阵 $\boldsymbol{V}_i'$ 就得到注意力特征图。

$$attention(query,key,value)=\alpha_i*V_i' \tag{6}$$

经过式(6)的处理后，我们用获得的注意力特征图来进行特征提取。MPCN(Multi Pyramid Convolution Network)由三通路动态卷积网络构成，对应图 3 中的 Pyramid conv，每路卷积采用尺寸不同的卷积窗口提取文本特征，且每路卷积由多个金字塔模块构成。金字塔模块如图 4 所示。

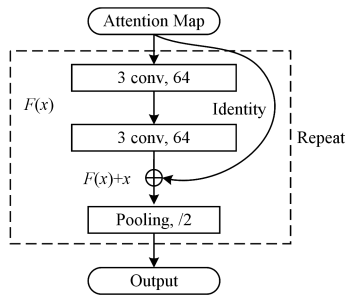


图 4 金字塔卷积模块

Fig. 4 Pyramid convolution module

每个卷积模块由两个等长卷积和一个残差模块构成，等长卷积会让每个词位的 embedding 描述语义描述得更加丰富和准确，残差模块避免了网络深度增加所带来的梯度弥散。文本特征在经过两个卷积层和恒等映射后，会使用尺寸为 3、步长为 2 的池化层进行下采样，每做一次下采样，下一个卷积模块的序列长度减半，但其能够感知到的文本片段比之前长了一倍。通过卷积模块与池化层的交错进行，最终序列长度变为 1。对 3 个通路进行 MaxPooling 后，通过 Concatenate 进行特征拼接，并将拼接结果作为聚类算法的特征输入。

3.4 基于 SOM 的文本聚类

SOM 网络是通过模拟人类大脑皮层对信号的自组织映射特性得到的。SOM 不仅降低了文本特征的维度，而且利用

本身的自组织映射特性对文本特征及其邻域特征进行了权值调整。SOM 网络结构如图 5 所示。

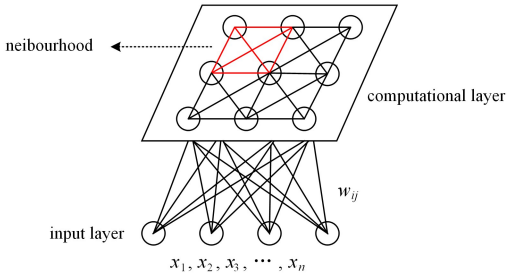


图 5　SOM 结构图

Fig. 5　SOM structure diagram

SOM 算法的核心在于竞争学习和邻域权值的调整，计算公式如下：

$$winner=\text{argmax}\parallel x_i * w_j \parallel \tag{7}$$

$$r=C_1\left(1-\frac{t}{iteration}\right) \tag{8}$$

$$w_{ij}(t+1)=w_{ij}(t)+\sigma(t,N)[x_i-w_{ij}(t)] \tag{9}$$

式(7)代表文本 x_i 与神经元 w_j 的内积，内积最大的元素下标就是获胜神经元；式(8)是优胜邻域的领域半径；式(9)更新获胜神经元权值以及优胜邻域的权值。

3.5　SA-MPCN&SOM 算法

SA-MPCN&SOM 算法的流程如算法 1 所示。

算法 1　SA-MPCN&SOM

输入：软件需求语料库

输出：相似性需求聚类

Step1　数据预处理。首先用 NLTK 对软件需求数据进行分词处理；然后根据停用词表去除无关紧要的词，获取整个数据集的文本最大长度，并将所有文本补全至最大长度；最后使用 BERT 训练词向量。

Step2　注意力特征图提取。将词向量按照式(1)输入，设置 Self-Attention 输出维度，按照式(2)－式(4)完成向量的线性转换，根据式(5)和式(6)提取注意力机制特征图。

Step3　MPCN 文本特征抽取。将 Step2 提取的特征图放入 MPCN 网络，经过深度挖掘后输出最终融合的特征向量。

Step4　SOM 聚类。将 MPCN 网络的输出作为 SOM 的特征输入，通过式(7)计算获胜神经元，然后根据式(8)计算获胜神经元的优胜邻域，采用矩形表示，接着根据式(9)对邻域权值进行调整，调整完成后再进行下一轮的迭代。

Step5　迭代结束。输出最终的聚类结果，算法终止。

4　实验及结果分析

4.1　实验数据

本文主要使用软件的功能性需求文本作为实验数据，数据来自一个英文网站：Softpedia。Softpedia 平台会为用户提供各式各样的软件工具，包括 windows，games，drivers 和 news&reviews 等。软件多达几十类，包括杀毒软件、编辑工具、CD/DND 制作工具、压缩工具、文件管理、地图 GPS 工具等。同时，该网站也会提供最新的软件新闻资讯。截至北京时间 2019 年 5 月 27 日，该网站总共收录应用程序13 016 个，应用程序共被下载 3 320 687 342 次。

本文主要挑选 windows 类下的软件需求数据进行爬取，一共爬取了 11 类软件数据，共 15 598 条。爬取到的软件类型和数量如表 1 所列。

表 1　软件需求数据表

Table 1　Software requirements data sheet

软件类别	数量
Antivirus	625
Authoring-Tools	676
CD-DVD-Blu-ray-Tools	1545
Compression-tools	466
Desktop-Enhancements	7612
File-managers	546
Gaming-Related	114
iPod-Tools	174
Maps&GPS	25
Mobile-Phone-Tools	433
Network-Tools	3382

4.2　实验设置

由于 BERT 训练数据的耗时较长，本文以 BERT-Base，Cased 作为预训练模型来对软件需求文本进行训练。多路金字塔卷积网络存在大量的参数，为了使模型在训练过程中快速达到最好的状态，本文对模型参数设置如下：

(1)对于所有的数据集，使用修正线性单元激活函数。

(2)三路卷积使用的滤波器窗口大小 $size=[3,4,5]$，每种滤波器的通道为 64。

(3)池化层的 $size=3$，$stride=2$。

(4)SOM 的权值竞争矩阵的大小为 $1*11$，迭代轮数为 30。

4.3　结果分析

本文将不同长度的文本补齐到最大长度，然后使用 BERT 词向量模型对文本的单词进行编码，最后使用本文构建的 SA-MPCN 网络提取特征。为验证本文采取的特征提取模型优于其他深度学习模型，本文模型与对比深度学习模型在完成特征提取后，由 SOM 完成特征向量的聚类。本文算法与现有代表性深度学习模型 TextCNN，DPCNN，TextRNN，TextBiRNN，TextAttBiRNN，ESIM 进行特征提取准确率的对比结果如图 6 所示。

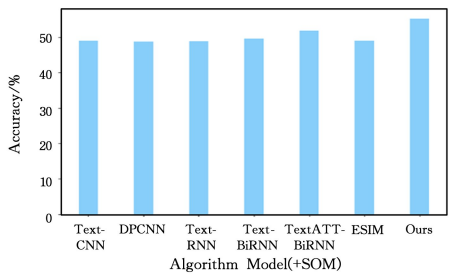


图 6　特征提取准确率的对比

Fig. 6　Comparison of feature extraction accuracy

由图 6 可知，不同深度学习模型在特征提取时有不同的准确率，但大部分模型的准确率为 48%～52%。这是由于大部分深度学习模型采用单一尺寸的卷积核进行特征提取，或者特征提取模型只是一味增加深度，忽略了模型深度增加所带来的文本信息的丢失，因此再次提升需求文本聚类的准确

率将变得十分困难。然而，本文的 SA-MPCN 模型在软件需求文本聚类的准确率上可以达到 55%左右。经过分析，需求文本特征经过自注意力机制来捕获序列的上下文关系，并根据每个词语对序列的重要程度分配不同的权重，权重大的词语对整个序列具有表征意义，充分体现了词语间的差异性和相关性，有利于卷积网络提取文本特征，也使后期的聚类具有有效性和可靠性；将注意力机制提取的 Attention Map 作为多路金字塔卷积网络的输入，不同路径的卷积模块采用不同尺寸的卷积核来提取文本特征，增加文本特征的多样性。卷积模块随着卷积深度的增加，采用恒等映射的方法将原始输入融入卷积之后的特征，这既深度挖掘了需求文本特征，又消除了卷积深度带来的信息缺失的影响。序列长度会随着金字塔卷积缩短，但文本特征的感知区域却在增加，有效提高了文本特征挖掘的性能，其挖掘的文本特征也为聚类算法的向量输入奠定了基础，使聚类算法的输入是基于较好的语料而划分的。实验结果表明，本文的特征提取模型能进一步改善聚类算法的特征输入。

为验证本文算法的组合性能，还将其与 SOM 和 MPCN&SOM 进行聚类准确率的对比。

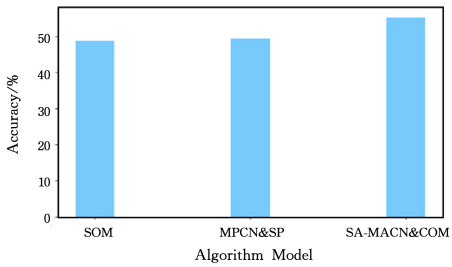


图 7 组合算法准确率的对比

Fig. 7 Comparison of accuracy ratio of combination algorithms

由图 7 可知，经过 BERT 训练的词向量经过 SOM 聚类后，其准确率为 48.80%；而将 BERT 训练的词向量经过 MPCN&SOM 聚类后的准确率为 49.45%，相比 SOM 聚类的准确率提高了 0.65%。因此，多路金字塔卷积网络对软件需求文本的特征提取改善了聚类算法的特征输入，有效提高了聚类算法的准确率。在 MPCN&SOM 的基础上，本文又引入 Self-Attention 对序列内部结构进行捕获，获取了上下文的语义关系，其准确率达到了 55.28%，相对于 MPCN&SOM 的准确率提高了 5.83%，说明 Self-Attention 的融合有效提升了聚类算法的准确率；同时，与序列内部语义关系的进一步融合，对聚类算法也具有重要意义。

本文除了在准确率方面对模型进行对比外，还将在 SC (Silhouette Coefficient)，CH(Calinski Harabasz)和 DB(Davies Bouldin)等聚类评价指标上进行比较。

SC 通过描述类中内聚度和类间分离度来评价聚类算法的好坏，其值越接近 1，聚类越合理。SC 的计算公式如下：

$$SC(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{10}$$

其中，$a(i)$是样本 i 与同簇内其他样本的平均距离，$b(i)$是样本 i 与其他簇中所有样本的最小平均距离。

CH 通过比较类间差距的迹与类内差距的迹来评判算法，其值越大代表类内越紧密而类间越分散。CH 的计算公式如下：

$$CH(k)=\frac{tr(B_k)m-k}{tr(W_k)k-1} \tag{11}$$

其中，m 为训练样本数，k 为类别数，B_k 为类之间的协方差矩阵，W_k 为类别内部数据的协方差矩阵，tr 为矩阵的迹。

DB 衡量的是每个簇的最大相似度的均值，值越小代表聚类结果越好。DB 的计算公式如下：

$$DB=\frac{1}{N}\sum_{i=1}^{N}\max_{j\neq i}\left(\frac{\sigma_i+\sigma_j}{d(c_i,c_j)}\right) \tag{12}$$

其中，N 为聚类簇数，分子代表簇内所有点到该簇质心点的平均距离之和，分母表示两类别质心间的距离。

表 2 多种评价指标的对比

Table 2 Comparison of multiple evaluation indicators

方法	Silhouette Coefficient	Calinski Harabasz	Davies Bouldin
TextCNN	0.008	334.588	4.576
DPCNN	0.153	2963.982	3.067
TextRNN	0.165	12229.717	1.526
Text-BiRNN	0.100	919.392	2.573
TextAttBiRNN	0.127	1300.339	2.627
ESIM	0.068	2627.672	2.915
SA-MPCN	0.195	1054.827	1.736

由表 2 可知，在 Silhouette Coefficient 上，本文算法优于其他算法，比其他模型中效果最好的结果高出 0.030，证明本文算法的内聚度和分离度都相对较优。在 Davies Bouldin 上，本文算法比 TextRNN 高出 0.21，但相差不大，说明本文聚类后同簇内部紧密而不同簇分离较远。而在 Calinski Harabasz 上我们可以看到，本文模型位于中间水平，但由于数据不平衡，最终结果与 TextRNN 一样存在组间分散严重和组内聚集不密集的问题。综合来看，本文除了准确率表现较优外，其他聚类评价指标也能达到不错的效果。

为验证本文聚类算法的性能，将其与现有的代表性聚类算法进行准确率的对比。对比算法包括 DBSCN，Spectral Clustering，Hierarchical clustering，GMM 和 K-means。

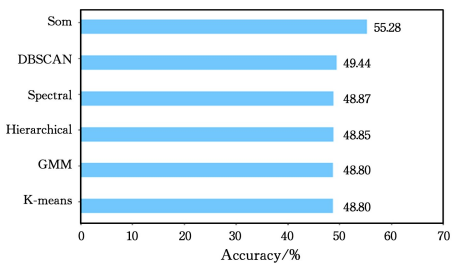


图 8 不同聚类算法的准确率

Fig. 8 Accuracy of different clustering algorithms

由图 8 可知，在使用本文深度学习模型完成文本特征的提取后，通过结合不同的聚类算法得到的准确率也不同。可以观察到，K-means 和混合高斯聚类的准确率是相同的，而层次聚类和谱聚类的准确率均稍有提高，密度聚类算法的准确率相比层次聚类和谱聚类的准确率稍有改善，而本文算法的组合方式优于其他聚类组合方式。经过分析，传统的聚类算

法大部分是基于度量计算的，距离近的被划分为一类，鲜有考虑样本的权值，只是通过移动聚类中心来对样本进行划分，而且部分聚类算法对初始聚类中心敏感；本文使用的 SOM 聚类算法可以避免预先设定聚类中心所带来的干扰，同时通过竞争学习修改获胜神经元及邻域样本权值，使其靠近聚类中心，可以很好地适应数据分布，解决了传统聚类算法易陷入局部极小值和对初始值敏感的问题，最终达到比较好的聚类准确率。

结束语　需求分析阶段是整个软件生命周期的重要阶段，前期的需求分析直接决定后期工作的成败。将软件需求文本按照功能性聚类，也可以指导开发人员对软件进行模块化设计。本文针对软件需求文本的特点和传统聚类算法的局限性，融合注意力机制和金字塔卷积网络对文本特征进行深度挖掘，同时保留原始信息，融合最终卷积后的特征交由 SOM 自组织完成聚类输出。由于需求文档的边界不清晰、描述容易模糊且有二义性，未来需要结合领域知识来进一步提高聚类性能。

参 考 文 献

[1] TONG Z X，MA P J，DING X，et al. Requirement Research on Demand Clustering and Demand Optimization Method Based on Natural Language Understanding [J]. High Technology Letters，2015，25(3)：257-269.

[2] MÖLLER K H. Ausgangsdaten für Qualit? tsmetriken－Eine Fundgrube für Analysen[M]// Software-Metriken in der Praxis. Springer，1996.

[3] BOEHM B W，ROSS R. Theory-W Software Project Management：Principles and Examples[J]. IEEE Transactions on Software Engineering，1989，15(7)：902-916.

[4] DAVIS A M. Software requirements：objects，functions，and states [M]. PTR Prentice Hall，1993.

[5] XIAO W N，ZHANG W Q，WANG L L. Research on a Software Needs Analysis Risk Assessment Model Based on BP Neural Network [J]. Computer Science，2011，38(4)：199-202.

[6] WANG Y M，HAN F，WANG H P，et al. Software Demand Risk Model Based on Grey Clustering Evaluation and Its Application[J]. Computer Engineering and Design，2006(18)：3497-3500.

[7] DEVLIN J，CHANG M W，LEE K，et al. BERT：Pre-training of Deep Bidirectional Transformers for Language Understanding [M]// Encyclopedia of Systems Biology. New York：Springer，2013.

[8] VASWANI A，SHAZEER N，PARMAR N，et al. Attention Is All You Need[J]. arXiv：1706.03762，2017.

[9] MARGHNY H. MOHAMED，MOHAMMED M. Abdelsamea：Self Organization Map based Texture Feature Extraction for Efficient Medical Image Categorization [J]. arXiv：1408.4143，2014.

[10] MARTIN J，KLEINROCK L. Excerpts from：An Information Systems Manifesto [J]. Communications of the ACM，1985，28(3)：252-255.

[11] ZHAO W，ZHANG L，MEI H，et al. A Program Clustering Method Based on Functional Demand Hierarchy Condensation [J]. Journal of Software，2006(8)：1661-1668.

[12] JIANG B，YE L Y，PAN W F，et al. Service clustering method based on demand function semantics [J]. Journal of Computers，2018，41(6)：1035-1046.

[13] SUN Z Y，LIU G S. Research on neural network clustering algorithm for short texts [J]. Computer Science，2018，45 (S1)：392-395.

[14] HU W S，YANG J F，ZHAO M. Demand analysis based on grey clustering algorithm [J]. Computer Science，2016，43 (S1)：471-475.

[15] MIKOLOV T，CHEN K，CORRADO G，et al. Efficient Estimation of Word Representations in Vector Space[J]. arXiv：1301.3781，2013.

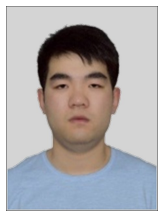
[16] COATES A，NG A Y. Learning feature representations with K-means[J]. Lecture Notes in Computer Science，2012，7700：561-580.

[17] ZEPEDA-MENDOZA M L，RESENDIS-ANTONIO O. Hierarchical Agglomerative Clustering[M]// Encyclopedia of Systems Biology. New York：Springer，2013.

[18] COMON P. Independent component analysis，A new concept [J]. Signal Processing，1994，36(3)：287-314.

[19] BLEI D M，NG A Y，JORDAN M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research，2012，3：993-1022.

KANG Yan，born in 1972，master supervisor，is member of China Computer Federation (CCF). Her main research interests include machine learning and so on.



CUI Guo-rong，born in 1995，master. His main research interests include natural language processing and so on.