基于主题模型的通用文本匹配方法

黄振业[1] 莫淦清[1] 余可曼[2]

[1](浙江金融职业学院信息技术学院 浙江 杭州 310018)

[2](杭州平治信息技术股份有限公司 浙江 杭州 310030)

摘 要 检测长文本和短文本相似性的应用场景越来越多,文本对的一致性检测大多可以统一抽象成文本相似性的比较问题。该问题的难点在于短文本是零散的,从而很难判断其属于哪个领域及其背景知识,也难以引入词嵌入来解决在通用场景的具体文本匹配问题。基于这个问题,提出一种新的基于文本聚类主题模型的轻量方法,不需要利用额外的背景知识来匹配通用文本相似性。在两个经典测试样本数据集上的实验结果表明,该方法的文本相似性检测效率非常高。

关键词 自然语言处理 文本匹配 主题模型 吉布斯采样

中图分类号 TP391 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2024.05.045

GENERAL TEXT MATCHING BASED ON TOPIC MODEL

Huang Zhenye[1] Mo Ganqing[1] Yu Keman[2]

[1](*School of Information Technology, Zhejiang Financial College, Hangzhou 310018, Zhejiang, China*)

[2](*Hangzhou Pingzhi Information Technology Co., Ltd., Hangzhou 310030, Zhejiang, China*)

Abstract The similarity measurement between a long text and a short text relatively has more and more application scenarios, and the consistency judgment on these text pairs can be abstracted as a comparison problem of text similarity. The challenge is that the short text is sparse, it is difficult to determine which domain it belongs to and it is also difficult to introduce word embedding to solve the specific text matching problem in general scenarios. Aiming at this problem, this paper proposes a lightweight approach based on topic model with text clustering which can match generalized long-short texts without using extra related background knowledge. The experimental results on two typical test sample datasets show the text similarity detection efficiency of the proposed method is very high.

Keywords Natural language processing Text matching Topic model Gibbs sampling

0 引 言

评估文本相似性的语义相似性被普遍应用到很多现实场景中。例如,通过判断是否一篇新闻的内容契合其标题所表达的含义,来帮助新闻审查人员提升审查效率;通过评判电子商务网站上的商品标题和商品详情页的内容是否匹配,来提升网站后台对的商品是否合规的审查效率等。

虽然解决文本内容语义相似性的方法已经有很多种[1],但是对文本相似的检测场景依然存在很多挑战。递归卷积神经网络应用到文本分类方法[2]和基于记忆网络的大规模智能问答理解方法[3]在解决文本相似问题时,需要先假设被检测文本具有相当的规模。当被比较的文本含有短文本,而且该短文本缺失相关背景信息时,这些方法会遇到短文本特征稀疏问题。为了解决短文本特征稀疏问题,可以利用外部知识,也可以利用预训练的词嵌入,还可以利用通用或者是专业的领域知识。某些传统的方法引入外部数据来帮助解决文本的相似性问题,例如:基于未标记背景知识改进短文本分类方法[4]利用未标记语料库作为“背景知识”,把语料库作为文本比较的中介。另外,也有一些基于

收稿日期:2021-01-28。**黄振业**,讲师,主研领域:自然语言处理,区块链技术。**莫淦清**,副教授。**余可曼**,高工。

文本表现模型的方法,例如:Doc2Vec[5]。近些年,也有一些利用神经网络体系的方法,例如:孪生网络[6](Siamese-LSTM)来评估句子间的语义相似性。此外,还有最新的隐含主题比较方法[7]。该方法基于领域的预训练词嵌入来进行文本比较。总之,在这些方法的模型中,必须要使用背景知识或者预训练嵌入单词来解决真实场景中的文本相似性检测问题。

然而在真实的场景,适合的背景数据并不容易定义和选取。对于通用的文本相似性比较问题,判断长文本和短文本分别是属于哪个领域是很困难的。但如果使用基于词嵌入方法时选取的通用知识或者通用单词语义关系,又会带来性能问题:短文本特征稀疏问题和高维的向量会引起内存和计算时间的大量消耗。此外,保持外部知识的更新和预训练词嵌入任务的更新也会更加困难。另外,这些基于常规文本表现模型的方法也不适用于文本相似性比较任务。这是因为,在某种程度上,短文本不可避免地引入了干扰。而文本表现模型引入了干扰的话,对于长文本中的非主题词,文本比较的最终结果会不稳定[7]。

为了解决这个问题,本文先引入一个假定:如果短文本的主题是匹配长文本的,那么可以假定它们的语义是相似且一致的。本文重点解决如何基于主题模型来从长文本中提取主题,然后来判断短文本是否和长文本语义相似。针对文本相似性比较场景,提出了一种新的主题模型。该模型主要针对基于 Dirichlet 混合模型[8](Dirichlet Multinomial Mixture, DMM),另外还提出一种评判文本相似性是否一致的评估方法。

本文方法通过对长文本比较场景的通用特征研究,做了以下假定:1) 短文本中的每个单词最多只出现一次,当短文本是标题的时候,这是短文本都会具备的特征。2) 在一次比较中,长文本中的一个单词会出现多次。当长文本是内容部分时,这是长文本都会具备的特征。3) 当短文本在通用的比较场景匹配长文本时,短文本中重要的单词也会出现在长文本中,同时短文本中的每个单词之间的强相关性也会出现在长文本中。这使得对于一个具体的比较任务,长文本可以被利用作为语义的辅助信息,称之为自辅助(Self-Auxiliary)增强。本文希望能从长文本它自己中挖掘出针对具体任务的语义关系,而不是从外部特定领域和通用知识来挖掘。4)假定在该文本相似性比较任务中仅仅只存在唯一的主题[8-10]。基于这四个通用文本相似性比较任务的假定,本文提出了一种新的方法。如图 1 所示,它扩展了 DMM 模型相应的比较任务。本文方法采用的模型使用了吉布斯采样[11](Gibbs Sampling)过程。

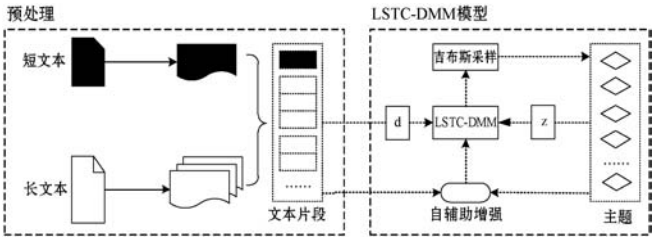

图 1　本文模型的长短文本匹配

由于词嵌入会带来不相关的全局干扰和性能问题,同时在很多场景中选择背景知识是比较困难的,本文方法既不需要引入全局外部数据或通用词嵌入,也不需要选择与具体领域相对应的背景知识。整个过程可以被看作是一种与人类习惯相一致的冷启动的语义理解过程,类似于没有相关背景知识的小孩第一次阅读一段文本内容并提取该文本的中心思想的过程,这个过程应该是轻量和通用的。在新闻标题-新闻内容匹配实验中,使用了两个来自于现实的相关数据集,实验结果显示本文方法的效率明显更高。

本文主要的贡献概括起来有三点:

(1) 利用了被比较的文本之间长度的差异,开发了一种高效、轻量与通用的模型来解决文本相似性匹配问题;而该模型不需要限制在某个特定领域,也不要求被比较的文本必须具有特定的长度。

(2) 相对于目前业界其他先进通用文本比较方法,本文方法对于通用的文本相似性的比较更加高效,开拓性地把主题模型应用到解决文本相似性匹配的问题上。实验结果也证明了相对于基准水平,本文模型在准确率和速度上具有优势。

(3) 细化统一的相似度衡量标准,这种衡量标准可以被用在各种文本区分方法上,从而实现在相同标准上对这些方法进行评估。同样的,该评估方法也可以用于比较不同文本区分方法在挖掘文本中潜在的主题上的能力。就对主题模型提取潜在变量的效率上来说,在通用数据集上对文本相似性进行评估的方法可以被认为是一种对主题模型进行评估的通用方法。

1　背　景

传统的文本匹配方法通常会使用一些数值估算值,例如余弦相似度和距离来评估文本间的语义相似度,或者在使用分类算法或聚类算法完成文本分类之后判断是否文本是语义相关的等。有几种匹配方法依赖于分布式语义模型,该模型对应从分布式语料库分析[12]来进行一种语义空间构建。所以文本匹配问题和四个方面紧密相关[13]:文本语义空间的表现模型、语言单元(词、句、段落和文本)的比较、匹配算法以及

相似度测量方法。基本上,有以下三种主要的文本匹配方法。

1.1 几何方法

各种几何方法在用上下文向量定义的语义空间中对两个语言单元的相对位置进行评估。这些方法把单词匹配到一个多维空间,因此那些相关的单词被当作是在更高维度空间中的某个特定的点,并基于这些特定点来进行文本比较。这类方法中,文本也可以通过在向量空间中的对应的向量来表现,因此也能够通过测量向量间的距离来比较单词,例如:Doc2Vec。词移距离[14](Word Mover's Distance,WMD)依赖词嵌入(Word Embeddings)的预训练。词移距离把两个文档之间距离的定义为把所有的词从一个文档转换到另一个在词嵌入空间的最优转换代价。但是在长文本场景中,必须考虑词移距离的计算复杂性。隐含主题比较方法[7]基于特定领域的词嵌入模型来提取文档的隐藏主题,同时使用几何测量方法来进行文档比较。但是,对于匹配问题,隐含主题比较方法高度依赖领域背景知识以及特定领域的词嵌入。

1.2 统计方法

统计方法被当作为经典的主题模型方法:概率潜在语义索引[15](Probabilistic Latent Semantics Analysis,PLSA)和 LDA 文档主题生成模型[16](Latent Dirichlet Allocation,LDA)。概率模型依赖两种概率分布:(1) 单个单词和给定的主题相关的概率;(2) 单个文本引用了一个主题的概率。为了解决短文本特征稀疏问题、高维问题和海量数据的问题,最近很多对主题模型的改进[9-10,17-18]被提出。本文方法在解决匹配任务时,将长文本作为对文本主题的富语义关系的辅助知识,应用基于主题模型来识别被比较文本的主题。由于依赖本地特定任务相关的知识,因此本文模型是轻量级的。

1.3 神经网络方法

最近几年,神经网络方法也被应用于文本匹配。Siamese-LSTM 被用于评估句子间的语义相似性。上下文对齐递归神经网络模型[19](Context-Aligned RNN,CA-RNN)被用来进行句子相似性建模。这些方法更侧重于解决长度差不多的句子的相似性问题,这些方法也依赖于预训练语料或者词嵌入。

2 本文模型

正如之前的讨论,短文本和长文本的主题如果相匹配,那么文本对被认为是语义一致的。本文模型通过挖掘文本的主题来判断长短文本是否属于同一个聚类。

传统的主题模型,例如:PLSA 和 LDA,遵循一个假定:即每个文本是在一个主题集合上进行建模,而且也能被很好地作用在长文本上。然而,传统主题模型被认为不适用于短文本,所以也就不能用于在文本相似性比较任务中发现主题。而且直接用通用文本聚类算法把被比较的文本分别划分到两个聚类来判断它们是否语义相关也非常困难。为了解决短文本的特征稀疏问题、高维问题和海量数据问题,基于伪文档的短文本主题模型[18]和短语主题模型(PTM)[20]提出了通过伪文档和辅助数据,基于潜在语义结构发现和文本聚类的解决方法。

受此启发,在通用文本相似性比较任务中,本文将长文本作为同时作用于短文本和长文本及其自身的辅助数据。当被比较双方的文本长度差异极大时,可以忽略在短文本中的语义信息,而只把长文本作为辅助信息来提供匹配任务中的语义相关信息。在本文模型中,将长文本和短文本的比较从各自的单主题来生成也是很重要的。本文方法由三个过程组成:数据预处理、主题推导和相似度评估。

2.1 预处理

对于数据预处理,先要将比较任务中原先的长文本切分成文本片段;每一个文本片段是原先长文本的一部分,称为长文本片段(Long-Text Snippets)。而比较任务中的短文本可以直接作为短文本片段(Short-Text Snippet)。对长文本片段和短文本片段进行数据预处理,遵循以下数据预处理规则:

(1) 把字母转换成小写。

(2) 删除非拉丁字符和停顿词(Stop Words)。

(3) 使用 Word-Net Lemmatizer of NLTK3 对单词实施词干提取(Stemming)。

(4) 删除长度小于 2 或大于 15 的单词。

(5) 在短文本片段中,删除那些只出现在短文本中而没有出现在长文本中的单词。即删除在短文本中文档频率小于 2 的单词。

(6) 删除长文本中长度小于 5 的单词。要求长文本和短文本的长度差距不能过大。

2.2 Dirichlet 混合模型

LDA 文档主题生成模型把每个文档 d 作为一个在主题之上的概率分布 θ_d。每个主题 z 被建模为在固定词表 V 中的单词之上的分布式概率。DMM 被用于解决数据特性稀疏问题和短文本上下文有限的问题。该模型是一种通用概率模型,其基于假设:一个文本是从

一个单一主题生成的,同时被应用到短文本主题模型[10,21]。因为长文本和短文本分别只有一个主题,所以该模型适用于文本相似性比较场景中的特征化。因此本文采用 DMM 作为模型的基础。

在集合 D 中生成一个文本 d 的过程如图 2 所示。先选取一个主题赋值给文档,然后主题/单词 Dirichlet Multinomial 模块从被选取的同一个主题生成文档中的所有单词。另外,还选取了两个 Dirichlet 分布分别为主题/单词的分布 ϕ 和文档/主题的分布 θ,其中 α 和 β 是这两个分布对应的超参数(hyper-parameters)。文档 d 中的所有单词从同一个主题 z_d 生成(从 Multinomial 分布 ϕ_{zd} 生成)。

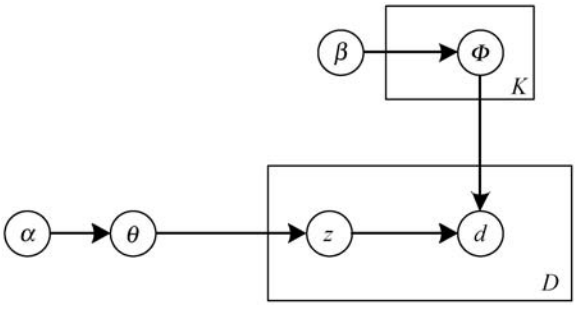

图 2　DMM 图形模型

该 DMM 生成过程如下:

步骤 1　通过采样得到主题的分布:θ ~ Dirichlet(α)。

步骤 2　对每个主题 k ∈ {1,2,⋯,K}:

计算出主题词的分布:distribution φ_k ~ Dirichlet(β)。

步骤 3　对每个文档 d ∈ {1,2,⋯,D}:

步骤 3.1　采样得到 z ~ Multinomial(θ)。

步骤 3.2　对每个单词 w ∈ $w_{d,1}, w_{d,2}, \cdots, w_{d,nd}$:

采样得到单词 w ~ Multinomial(φ_{zd})。

精确计算每个文档 d 的混合组成 z 通常是困难的。然而有很多类似的推导技术可以用来解决该问题,比如变分推断[16](Variational Inference)和吉布斯采样。在对主题 z 做每个吉布斯采样迭代时,引入了以下两个条件分布[10]:

$$p_1(z_d = z \mid \vec{z}_{\neg d}, \vec{d}) \propto \frac{\vec{m}_{z,\neg d} + \alpha}{D - 1 + K\alpha} \frac{\prod_{w \in d} (n^w_{z\neg d} + \beta)}{\prod_{i=1}^{N_d} (\vec{n}_{z,\neg d} + V\beta + i - 1)} \tag{1}$$

$$p_2(z_d = z \mid \vec{z}_{\neg d}, \vec{d}) \propto \frac{\vec{m}_{z,\neg d} + \alpha}{D - 1 + K\alpha} \frac{\prod_{w \in d} \prod_{j=1}^{N^w_d} (n^w_{z\neg d} + \beta + j - 1)}{\prod_{i=1}^{N_d} (\vec{n}_{z,\neg d} + V\beta + i - 1)} \tag{2}$$

式中:N_d 是文档 d 中的单词数;N^w_d 是单词 w 在文档 d 中出现的次数。而 $m_{z,\neg d}$ 是在不考虑文档 d 的情况下主题 z 的文档数目。此外 $n^w_{z,d}$ 是单词 w 在主题 z 中的出现次数;$n_{z,\neg d}$ 是在主题 z 中,不考虑当前文档 d 相关时出现的所有单词数。在每轮吉布斯采样迭代中,文档 d 的主题不断的根据条件分布 $p(z_d = z \mid \vec{z}_{\neg d}, \vec{d})$ 来被重新赋值。该条件分布采纳了每个文档属于每个主题的概率。这里 z_d 代表了文档 d 的主题从 z 中被删除掉了。在式(1)和式(2)中,当假设每个单词可能在每个文档中出现最多一次时,使用式(1)。而当假设一个单词被允许在文档中出现多次时,使用式(2)。对一个特定的文本相似性比较任务场景,本文使用这两个公式。

事实上,本文研究发现在短文本中极少出现重复单词。特别是在短文本中,总是会出现那些语义上重要的单词,例如在新闻标题中的短文本。在现实的场景中,长文本中往往允许重复的单词出现。所以对文档文本比较任务中的短文本片段和长文本片段,本文提出的新模型均基于这两个假设,并包含了这两个主题采样器。

2.3　自辅助增强

在文本相似性比较任务中,短文本的文本长度很短,通常缺少单词关联信息,也不存在它自带的富语义关系。如果长文本语义匹配了相应的短文本,那么当前长文本的上下文中,必然存在和短文本这边的继承语义之间的联系。反之,如果长文本不匹配短文本,不太可能从长文本那边挖掘到和短文本这边关联的语义关系。因此如果不引入外部知识,就很难确认被比较文本两边的语义之间的直接联系。

如果把相对来说具有更多相关信息的长文本片段作为比较双方的增强辅助信息,那么在同一个主题的所有文本片段中,长文本片段更加重要。由于所有这些长文本片段事实上有相同的主题,因此做一个假设:长文本片段更可能被重新赋值给其他长文本片段的主题,它被赋值给短文本片段主题的可能性较小。基于这些特性,本文提出了一种自辅助增强来平滑采样过程,从而能提供更好的准确率。事实上,自辅助增强显著地提升了本文模型对文本相似性比较任务的效果。

2.4　主题推导

本文方法使用和 DMM 模型相同的生成过程和图形表现。正如本文之前讨论的,本文方法在主题推导上和 DMM 模型是不同的。针对通用文本相似性比较任务,在本文方法中,对主题推导过程的吉布斯采样算法进行了改进。

对单一比较任务,将短文本作为一种短文本片段 s_0,同时使用 Gensim[22] 句子生成器,将长文本切为多

个短文本片段:$s_1,s_2,\cdots,s_{L-1}$。任务中的全文本被转换为 L 个片段,如果一些片段来源于长文本,可能会有几个片段是长文本中的原句。把 L 预设为所有长文本片段和短文本片段加起来的总数量。在这些文本片段中,第一个片段 s_0 肯定是来源于短文本,而其他的片段则来源于长文本。本文算法如算法 1 所示,相关变量的说明如表 1 所示。

算法 1 本文算法

1. 输入:文档列表 $\vec{d}(s_0,s_1,\cdots,s_{k-1})$。
2. 输出:为每个文档打上类别标签 $\vec{z}$。
3. 初始化以下三个计数变量为 0:m_z——属于第 z 类的文档数量;n_z——属于第 z 类的词的数量,n_z^w——第 z 类中的词语 w 的重复数量。
4. 初始化。循环遍历 N 个文档(其中任一个用 d 表示):

通过多项分布 z ~ Multinomial(1/K),随机为文档 d 抽取出一个类别标签 z_d

属于第 z 类的文档数量 m_z 加 1

属于第 z 类的词的数量 n_z 加 N_d(文档 d 的单词数)

内循环。遍历文档 d 当中的每个单词(w):

第 z 类中的词语 w 的重复数量 n_z^w 加上 N_d^w(单词 w 在文档 d 中出现的次数)

内循环结束

外循环结束。

5. 吉布斯采样过程。循环 I 次迭代:

循环遍历 N 个文档(其中任一个用 d 表示):

记当前文档 d 的主题为 $z=z_d$

属于第 z 类的文档数量 m_z 减 1

属于第 z 类的词的数量 n_z 减去 N_d

内循环开始。遍历文档 d 当中的每个单词(w)

第 z 类中的词语 w 的重复数量 n_z^w 减去 N_d^w(单词 w 在文档 d 中出现的次数)

内循环结束

(为文档 d 采样标签)

如果 d 是短文档,采样的分布使用式(1):

$z_d \leftarrow z \sim p_1(z_d=z \mid \vec{z}_{\neg d},\vec{d})$;(Equation 1)

否则如果 d 是来自长文档的片段,采样分布使用式(2):

$z_{d_{s0}},z_{d_{\neg s0}} \leftarrow z \sim p_2(z_d=z \mid \vec{z}_{\neg d},\vec{d})$;(Equation 2)

其中,进一步引入可解释性的增强因子,用来优化采样结果:

$\eta_d=\frac{n_{\neg z_{s0}}}{N_{\neg s0}}$;可解释为长文本片段的被分配的类别,在匹配任务中的影响力。其中分子部分表示当前采样步骤中,和分配给短文本的类别不相同的长文本片段的数量,分母部分表示长文本片段总数。因此,引入增强因子之后的采样结果如下表示:

$z_d \leftarrow \eta_d z_{d_{\neg s0}}+(1-\eta_d)z_{d_{s0}}$;

更新相应的计数记录,在下个循环中使用。

属于第 z 类的文档数量 m_z 加 1

属于第 z 类的词的数量 n_z 加上 N_d

内循环开始。遍历文档 d 当中的每个单词(w)

第 z 类中的词语 w 的重复数量 n_z^w 加上 N_d^w(单词 w 在文档 d 中出现的次数)

内循环结束

外循环结束。

表 1 本文中使用到的符号定义

符号	定义
D,N	语料库中文档的数量
V	词表中的单词数
K	预定义的潜在主题数
$\vec{d}$	语料库中的文档
$\vec{z}$	每个文档的主题标签
I	迭代次数
m_z	主题 z 中的文档数
n_z	主题 z 中的单词数
n_z^w	单词 w 在主题 z 中出现次数
N_d	文档 d 中的单词数
N_d^w	单词 w 在文档 d 中出现的次数
n_d^z	文档 d 中和主题 z 相关的单词数
$n_{z,d}^w$	文档 d 中的单词 w 在主题 z 中的出现次数
ϕ	主题/单词分布
θ	文档/主题分布

第一步,将所有的片段随机地分配给 K 主题聚类,这里的 K 是实验预设置的。按照精准主题发现方法[11]的聚类过程,在每个吉布斯采样迭代中,按照条件分布 $p(z_d=z \mid \vec{z}_{\neg d},\vec{d})$ 把每个片段 s_i 重新赋值给一个主题。对于 I 轮迭代(这里设置 I 为 10 次),采用 DMM 混合模型的短文本主题聚类方法[10]中的通用设置对所有的这些片段进行转换。为了计算出匹配分,使用赋值给和该短文本片段主题相同的长文本片段的数量,除以长文本片段总数作为计算结果。匹配分的计算式为:

$$Score_{\text{本文模型}}=\frac{N_{S\in S_1\cdots S_{L-1}}^{t_{S_0}}}{N_{S_{\neg 0}}} \tag{3}$$

式中:$N_{S\in S_1\cdots S_{L-1}}^{t_{S_0}}$ 代表在最终的聚类结果标签集合中被赋值给同一个短文本主题 t_{S_0} 的长文本片段的数量,$N_{s\neg 0}$ 是长文本片段的总数。

与普通的文本聚类方法不同,本文模型把长文本的文档片段作为比较中的分子。所以,当计算每个文档片段 d 的 $p(z_d=z \mid \vec{z}_{\neg d},\vec{d})$ 时,一个自辅助增强 η_d 引入到了本文模型中。这样可以使长文本片段更可能地被重新赋值给其他长文本片段的主题,而不大可能会被重新赋值给短文本片段的主题。因此,文本相似性比

较任务就得到了比直接使用通用算法更高的准确率。在这一采样过程中,利用算法 1 来计算增强因子 η_d。

3　实验与结果分析

本节对两个公开可用数据集进行实验。这两个公开数据集是典型的文本相似性比较场景:新闻文章的标题-内容数据和科学论文的标题-摘要数据。这两个数据集没有预设的专用领域信息、外部知识背景和通用词嵌入。对比本文方法和其他先进方法在这两个数据集上进行文本相似性比较任务的效果以及计算执行的时间。实验结果表明在效果和效率上,本文模型都明显领先。

3.1　数据集

3.1.1　新闻文章[23]

这个数据集的内容包含了公开的 3 824 篇新闻的数据:标题、子标题和文本。这些新闻的内容单词数大多在 1 000 到 5 000 之间,而且标题长度基本上都不超过 15 个单词。这个数据集中的文本相似性的长度差异是比较大的,符合本文的假设,部分数据如表 2 所示。这些文章是一个项目在 2016 年 12 月到 2017 年 3 月之间收集的,其中新闻的来源是 ABC news、CNN news、The Huffington Post、BBC News、DW News、TASS News、Al Jazeera News、China Daily 和 RTE News。

表 2　新闻标题-内容数据集预处理样例

类型	原始文本数据	预处理后的文本数据
短文本	People Of Faith Are Pledging To Protect People Under Threat By Trump's New Policies	people of faith are pledging to protect people under by trumps new policies
长文本	Many people in our nation, and indeed around the world, are scared by the things happening in Washington. Those most affected …	many people in our nation and indeed around the world are scared by the things happening in Washington those most affected …

3.1.2　科学论文[24]

这个数据集包含了机器学习会议的科学信息。数据来自于 IEEE Xplore Digital Library 和 Machine Learning Applications (ICMLA)的国际会议的官方网站。这个数据集包含该会议的 448 篇论文,大多数论文的标题不超过 15 个单词,并且它们的摘要不超过 500 个单词,部分数据如表 3 所示。可以看到在这些标题-摘要数据集中被比较文本的长度差异只比新闻标题-新闻内容数据集的长度差异小一点点,同样符合本文的假设。

表 3　论文标题-内容数据集预处理样例

类型	原始文本数据	预处理后的文本数据
短文本	A Machine Learning Tool for Supporting Advanced Knowledge Discovery from Chess Game Data	machine learn support advanced knowledge discovery chess game data
长文本	In the current era of big data, high volumes of a wide variety of data of different veracity can be easily collected or generated …	current era big data high volume wide variety data different veracity easily collect generate high …

3.2　实验配置

3.2.1　本文方法的配置

本文模型中,先将长文本转换为多个片段,并在一个比较任务中,把这些长文本片段和短文本放在一起。在按照前面说的规则处理完成所有的数据之后,通过随机交换数据库中短文本某些行的值来产出一半的反面样例。本文模型对于这两个数据集的迭代次数均设置为 10。这样在聚类中的文档上做吉布斯采样过程速度很快,而且其计算执行时间也能保持稳定。本文模型把超参数设置成常规的值:$\alpha = 0.1$ 和 $\beta = 0.1$。

在文本相似性比较场景,只有"是"与"否"两个匹配结果。因此基于被比较文本对各自只有一个主题的假定,最多只存在两个主题。所以 K 理想的值是 1 或 2。本文 K 值设置为 2(或者略高于 2)。这是因为,如果 K 的值较小的话,更容易把属于同一个正向标注组(Ground Truth Group)的文档赋值到相同的聚类[10]。设置成这个值非常适合文本相似性比较任务。这也对本文模型从长文本端发现更多的辅助信息非常有帮助。在测量计算中,越多的长文本片段具有和短文本片段相同的主题,文本相似性就更有可能匹配上。本文模型打分评估方法定义在式(3)中。图 3 和图 4 展示了本文模型在两个不同的数据集的 K 值上的测量准确率、召回率和 F1 分数。可以看出本文模型方法在 $K > 2$ 以后,测量准确率、召回率和 F1 分数达到了稳定态。

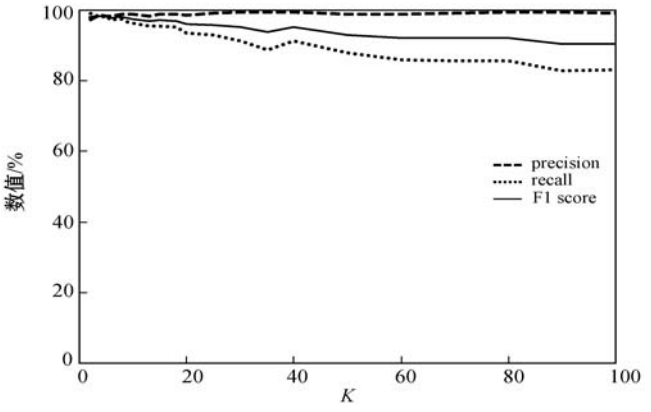

图 3　新闻标题-内容数据集的 K 值的准确率/召回率/F1 分数

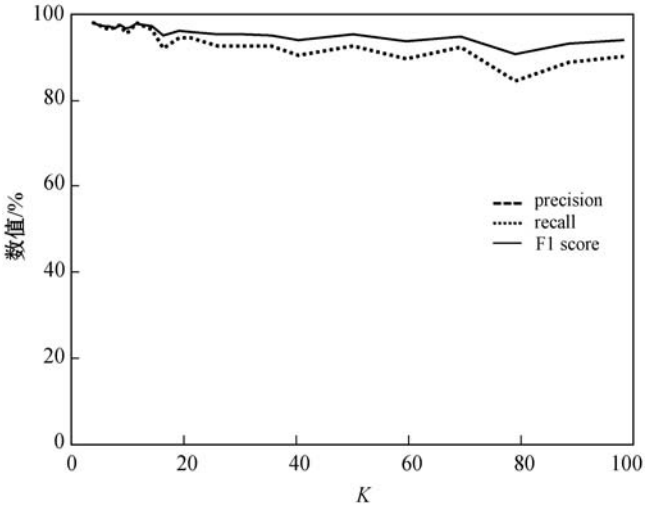

图 4 论文标题-摘要数据集的 K 值的准确率/召回率/F1 分数

此外,需要给长文本选择一个的片段数目 L。图 5 和图 6 显示了本文模型在不同的 L 值在两个数据集上的三个实验结果的对比。可以看出,在本文模型中,测量结果值虽然很稳定,但当 L 值增大的时候,测量结果会下降。这是因为在这个实验场景中,如果 L 变大,那么每个文本片段中的单词会变得更稀疏,也就是说,它需要等待更长的时间来完成采样过程中的收敛。然而如果 L 相对较小的话(例如 L 设置为 2,这就意味着长文本不能先被去除,而是被直接放进模型中),采样过程达到稳定会比较困难。在本文模型中,对文本相似性比较任务,选择把 L 设置为较小的值。这是因为按前面提到的,当 L 的值较小时,在相关长文本片段中,有更多的自辅助信息。这样本文模型中,就能更多地用这些辅助信息来确保相识度测量的稳定性,如图 5 和图 6 所示。同时,考虑到需要针对比基准实验来做时间复杂度的优化,L 设置为 3。

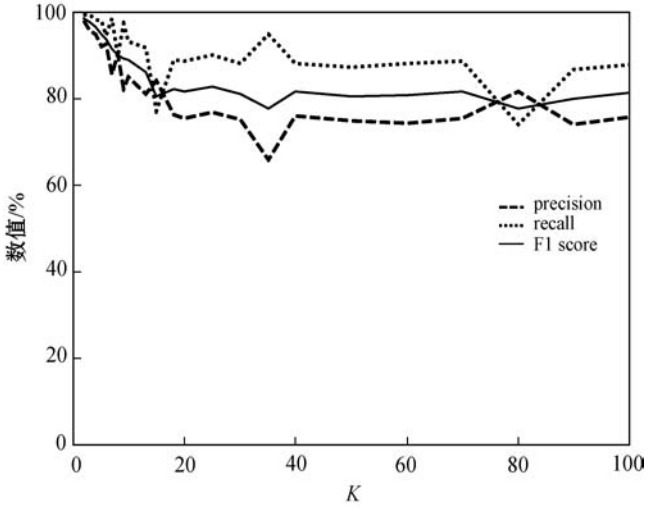

图 5 标题-内容数据集的 L 值的准确率/召回率/F1 分数

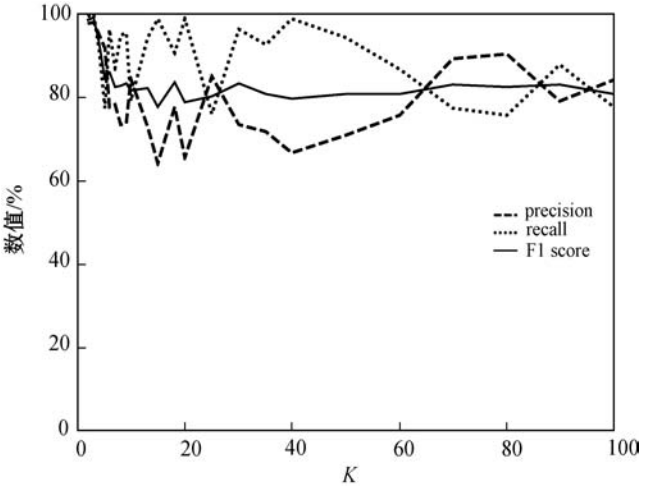

图 6 标题-摘要数据集的 L 值的准确率/召回率/F1 分数

在算法 1 中,DMM 模型的采样迭代的时间复杂度是 $O(KDS)$,这里 K 是主题数量,D 是文档数,S 是文本平均长度。根据特定的文本相似性比较任务的特征,本文方法改进了模型,采用相对小的固定的 K 值和 L 值来尽量缩减任务的时间复杂度。这样得到了在主要指标上好的平衡。可以看出使用本文模型,所有的实验结果并不稳定。这是因为不同文本规模的数据集的长度有点不同。

3.2.2 基线配置

除了不把长文本切分成多个子片段,在和其他基线模型的比较实验中,尽可能和本文方法的数据预处理方法保持一致。对每个基线,在这两个数据集上分别做 20 次实验来观察平均的结果和那些实验的波动范围。

(1) Doc2Vec。使用 Doc2Vec 模型的预训练,在每个比较任务中,将短文本和长文本分别向量化,然后计算它们各自向量的余弦相似度。

(2) 词移距离。采用预训练单词向量来分别把短文本和长文本向量化。然后在每个计算任务中,用 Gensim 工具来计算距离。当测量文本相似度时,值越小,被比较的文本就更加相似。

(3) SiameseLSTM。SiameseLSTM 方法中,相关模型需要先从数据集合(新闻或论文)上训练出来。然后被生成的模型按文本向量的方式来读入输入文本,同时也把它的最后隐藏状态作为每个文本的向量表现。两种表现间的相似度被用来作为两种文本比较的语义相似度的估计值。

(4) 隐藏主题模型。使用原论文中的代码来计算隐藏主题相关性的余弦相似度。但是因为是通用的比较场景,没有一个具体的领域,原论文中一个特定的词嵌入类目是不能用的,也不适用于通用的文本相似性

比较任务。所以这里使用全英文 Wiki 作为词嵌入。

3.3 评　估

不同的文本匹配方法使用了不同的计分标准,如表 4 所示。例如,如果用距离来衡量是否两个文本语义能匹配,距离的值越小,两个文本的匹配度越高。如果用余弦相似度来衡量两个文本的语义匹配相似度,余弦值越接近 1,两个文本越匹配。如果是用聚类算法来判断两个文本是否属于同一个聚类,那计分范围就是在 0 和 1 之间,这里“1”意味着两个文本语义上属于同一个聚类,“0”意味着两个文本语义上不匹配。表 4 中,加下划线的分数代表不匹配。

表 4　五种文本匹配方法原始分上的计分标准

方法	原始分数	阈值	指标类型
Doc2Vec	0.543 1　0.529 1　0.180 6 0.262 6　0.537 9　0.396 4 0.353 0　0.231 3　0.543 0 0.453 4　0.450 3　0.384 8 0.541 6　0.418 3　0.305 0 0.131 4　0.481 8 …	0.39	Cosine Similarity
WMD	3.656 9　4.068 0　4.659 6 4.862 4　3.629 8　2.768 1 4.973 5　5.003 5　3.411 9 3.678 3　4.510 9　4.497 2 3.154 6　3.791 9 …	4.31	Distance
Siamese LSTM	0.511 3　0.308 1　0.551 7 0.534 3　0.437 4　0.566 6 0.558 9　0.684 6　0.501 2 0.537 8　0.189 6　0.368 9 0.399 3　0.571 9　0.445 1 0.381 3　0.541 8 …	0.31	Cosine Similarity
HT	0.122 1　0.091 4　0.093 8 0.084 1　0.106 1　0.118 0 0.075 7　0.121 2　0.126 4 0.073 1　0.068 4　0.129 2 0.094 0 …	0.09	Cosine Similarity
本文方法	0.666 7　0.090 9　0.25 0.0　0.0　0.428 5　0.166 7 0.0　0.0　0.285 7　0.5 0.0　0.0 …	0.1	DMM Score

而本文使用统一的评估方法[7]来评估不同相似度算法的效果。在训练集上得到匹配的阈值分界,在验证的时候使用这个阈值来判断是否匹配。

3.4 讨　论

本文描述的算法步骤如下:

(1) 在训练过程中,通过本文算法,计算出每一对文档的匹配分数,得到训练阶段的匹配分数列表。

(2) 根据匹配分数和文档匹配的真实标签分数,计算得到基于统计学意义上的临界阈值分数。针对当前数据集的临界阈值分数可以判断文档是否匹配。例如,针对新闻数据集的分数阈值是 0.1,是本文算法在训练集上计算得到的,如表 4 所示。

(3) 在划分的验证集中测试验证时,使用了训练得到的判断阈值。例如,当通过本文算法计算出一对测试文档的匹配分数后,如果该分数大于该阈值(比如 0.1),则表示该测试文本对是匹配的;反之,则为不匹配的。由此来比较本文算法的有效性。

采用上述算法步骤,对各数据集在测试阶段的数据上做验证,得到的文本匹配测试结果(比如准确率等指标)。相对其他算法在相同数据集的划分下的测试结果,本文算法的实验结果有显著的改善。

如表 5 所示,由第一个数据集(新闻标题-内容文本对)的实验结果看出,本文方法在准确率、召回率、F1 分数以及综合平衡性等方面都比其他方法好。这可能是本文方法适用于文本相似性长度差异较大的数据集,而在新闻标题和内容数据集上,正好被比较文本对的长度差别比较大。

表 5　新闻标题-内容匹配任务的结果

指标	本文方法	Doc2Vec	WMD	Siamese LSTM	HT
准确率	0.964 ± 0.002	0.929 ± 0.014	0.910 ± 0.015	0.705 ± 0.015	0.956 ± 0.015
召回率	0.997 ± 0.002	0.995 ± 0.009	0.898 ± 0.025	0.980 ± 0.014	0.961 ± 0.005
F1 分数	0.991 ± 0.005	0.942 ± 0.006	0.905 ± 0.005	0.820 ± 0.009	0.959 ± 0.005
任务耗时/s	27.12 ± 3.75	71.93 ± 0.4	40.74 ± 4.93	32.81 ± 1.17	33.64 ± 3.62
是否需要预训练数据	No	Yes	Yes	Yes	Yes

如表 6 所示,由第二个数据集(论文-摘要文本对)的实验结果能够看出只有 Doc2Vec 方法在准确率上比本文方法略好一些。这可能是因为论文-摘要文本对数据集上的被比较文本对的长度差距相对来说小一点,而 Doc2Vec 方法正好在这种场景有较好的准确率[14]。但如果使用通用外部嵌入词,Doc2Vec 方法的性能问题和语义噪声问题就会出现,因此 Doc2Vec 方法的召回率就会比本文方法差。

表 6 论文标题-摘要匹配任务的结果

指标	本文方法	Doc2Vec	WMD	Siamese LSTM	HT
准确率	0.942 ± 0.024	0.964 ± 0.02	0.939 ± 0.022	0.614 ± 0.034	0.912 ± 0.036
召回率	0.994 ± 0.005	0.955 ± 0.021	0.924 ± 0.024	0.991 ± 0.008	0.939 ± 0.024
F1 分数	0.957 ± 0.022	0.931 ± 0.017	0.931 ± 0.011	0.685 ± 0.015	0.925 ± 0.015
任务耗时/s	5.97 ± 0.63	14.11 ± 0.24	5.77 ± 0.95	17.25 ± 1.06	7.23 ± 0.77
是否需要预训练数据	No	Yes	Yes	Yes	Yes

4 结 语

本文提出了一种不需要依赖外部背景知识或特定领域的嵌入词,而且快速有效的通用文本相似性匹配方法。尽管在某些场景中,例如使用同义词的场景,不得不引入相关的外部数据,但对绝大多数更通用的应用场景是不需要引入外部数据的,例如官方新闻、论文、正式商业描述和正式咨询系统等。为了解决文本相似性比较任务,本文基于通用特征,并在 DMM 模型的基础上使用吉布斯采样过程进行了扩展。自辅助方法的目的是增强文本相似性比较的效果。该方法也可以用于其他非监督任务。在两个公开的非特定领域的数据集上,对四种其他先进的匹配技术的基线方法上进行了实验,并使用一致的评估方法来比较实验的结果。结果表明本文方法在通用的应用场景中避免了具体领域的限制,相对其他的基线具有更好的效果。

本文还存在一些不足之处,如目前只是在英文数据集上进行了实验并取得了不错的效果。但本文提出的相似性匹配方法是否适用于中文等多语言场景有待进一步验证,这同时也是下一步的研究方向。

参 考 文 献

[1] 王仲远,程健鹏,王海勋,等. 短文本理解研究[J]. 计算机研究与发展,2016,53(2):262 - 269.

[2] Lai S W, Xu L H, Liu K, et al. Recurrent convolutional neural networks for text classification[C]//9th AAAI Conference on Artificial Intelligence,2015:2267 - 2273.

[3] Bordes A, Usunier N, Chopra S, et al. Large-scale simple question answering with memory networks [EB]. arXiv. 1506.02075,2015.

[4] Zelikovitz S, Hirsh H. Improving short-text classification using unlabeled background knowledge to assess document similarity[C]//17th International Conference on Machine learning,2000:1183 - 1190.

[5] Le Q, Mikolov T. Distributed representations of sentences and documents[C]//31th International Conference on Machine Learning,2014:1188 - 1196.

[6] 胡朝举,梁宁. 基于深层注意力的 LSTM 的特定主题情感分析[J]. 计算机应用研究,2019,36(4):1075 - 1079.

[7] Gong H Y, Sakakini T, Bhat S, et al. Document similarity for texts of varying lengths via hidden topics[C]//56th Annual Meeting of the Association for Computational Linguistics,2018:2341 - 2351.

[8] Nigam K, McCallum A K, Thrun S, et al. Text classification from labeled and unlabeled documents using EM[J]. Machine Learning,2000,39(2 - 3):103 - 134.

[9] Nguyen D, Billingsley R, Du L, et al. Improving topic models with latent feature word representations[J]. Transactions of the Association for Computational Linguistics,2015,3:299 - 313.

[10] Yin J H, Wang J Y. A Dirichlet multinomial mixture model-based approach for short text clustering[C]//20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2014:233 - 242.

[11] Griffiths T L, Steyvers M. Finding scientific topics[J]. Proceedings of the National Academy of Sciences, 2004, 101(S1):5228 - 5235.

[12] Baroni M, Lenci A. Distributional memory: A general 1026 framework for corpus-based semantics[J]. Computational Linguistics,2010,36(4):673 - 721.

[13] Harispe S, Ranwez S, Janaqi S, et al. Semantic similarity from natural language and ontology analysis[J]. Synthesis Lectures on Human Language Technologies,2015,8(1):254 - 265.

[14] Kusner M J, Sun Y, Kolkin N I, et al. From word embeddings to document distances[C]//32nd International Conference on Machine Learning,2015:957 - 966.

[15] Hofmann T. Probabilistic latent semantic indexing[C]//ACM SIGIR Forum,2017,51(2):211 - 218.

[16] 赵乐,张兴旺. 面向 LDA 主题模型的文本分类研究进展与趋势[J]. 计算机系统应用,2018,27(8):10 - 18.

[17] Li C L, Wang H, Zhang Z Q, et al. Topic modeling for short texts with auxiliary word embeddings[C]//39th International ACM SIGIR Conference on Research and Development in Information Retrieval,2016:165 - 174.

[18] Zuo Y, Wu J, Zhang H, et al. Topic modeling of short texts: A pseudo-document view[C]//22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2016:2105 - 2114.

(下转第 349 页)

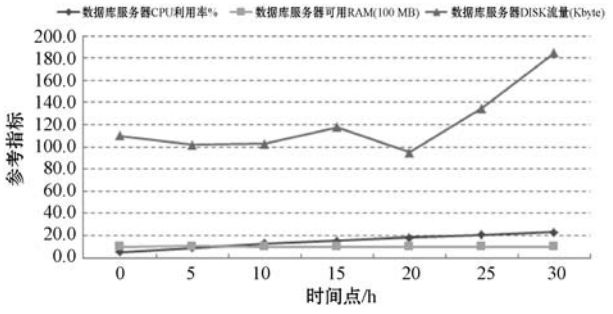

图 6　数据库服务器稳定性测试参数

从整体来看,C2TP 平台整体性能完全能够满足各项参数的要求,可以平稳运营在真实环境中。

5　结　语

C2TP 吸取了现有云培训技术的优点,尤其是在 ECS 弹性服务器方面,突破传统培训模式,为企业量身打造了与之相契合的云培训平台。针对企业私属性的特点及特色,科学地规划了与其需求相对应的培训功能模块,以积累知识、提升技能、岗前考核等为核心主导,实现企业闭环培训需求,与企业的持续发展相适应,提高从业人员的技能水平和对职业的满足感,为企业的生产与经营提供良好服务,从而不断提升企业的竞争力。今后的技术工作重点将集中在代码持续智能化集成方面,尤其是在性能监控方面。

参　考　文　献

[1] 闵丹. 支持云培训的教学资源管理平台设计与关键技术实现[D]. 北京:北京邮电大学,2019.

[2] Yu W, Kuang R, Xing R. Design and development of SCORM-based mobile learning system[C]//8th International Conference on Information Technology in Medicine and Education (ITME),2017:482 - 485.

[3] 李超,周泓. 学习管理系统综述和发展趋势展望[J]. 现代教育技术,2018,28(2):113 - 119.

[4] Liu Y, Li B, Niu J, et al. A Cloud-Based experiment platform for computer-based education[C]//2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing,2014:626 - 629.

[5] 黄乐辉,盛艳微,罗英. 基于云教育平台的移动学习模式研究[J]. 现代信息科技,2019,3(21):115 - 116,119.

[6] 刁兆勇,周建华. 大型企业标准化培训体系的构建与实施[J]. 中国标准化,2021(4):42 - 45.

[7] 塔娜. 基于云计算技术的大规模数据聚类分析[J]. 现代电子技术,2020,43(15):123 - 126.

[8] 贾琦. 云环境服务质量模型研究及应用[D]. 四川:电子科技大学,2016.

[9] Trabay D, Asem A, El-Henawy I, et al. A hybrid technique for evaluating the trust of cloud services[J]. International Journal of Information Technology,2021,13:687 - 695.

[10] 吴佩莉,张弢,张泉. 基于 SCORM 技术的多媒体课件统一播放框架与实现[J]. 计算机应用与软件,2019,36(5):108 - 111.

[11] 吴佩莉. 服务型云培训平台的闭环培训设计与实现[J]. 兰州文理学院学报(自然科学版),2020,34(3):88 - 92.

(上接第 318 页)

[19] Chen Q, Hu Q M, Huang J, et al. CA-RNN: Using context-aligned recurrent neural networks for modeling sentence similarity[C]//32nd AAAI Conference on Artificial Intelligence,2018:1232 - 1243.

[20] Quan X J, Kit C Y, Ge Y, et al. Short and sparse text topic modeling via self-aggregation[C]//24th International Joint Conference on Artificial Intelligence,2015:2270 - 2276.

[21] Zhao W Y, Jiang J, Weng J S, et al. Comparing twitter and traditional media using topic models[C]//33rd European Conference on Information Retrieval Research,2011:338 - 349.

[22] Řehůřek R, Petr Sojka. Software framework for topic modelling with large corpora[C]//Workshop on New Challenges for NLP Frameworks,2010: 45 - 50.

[23] New articles[EB/OL]. [2021 - 01 - 21]. https://dataverse.harvard.edu/dataset.xhtml? persistentId = doi:10.7910/DVN/GMFCTR.

[24] Vallejo-Huanga D, Morillo P, Ferri C. A dataset of attributes from papers of a machine learning conference[J]. Data in Brief,2019,24:103836.

(上接第 339 页)

[16] 尹毅峰,刘扬,徐明明. 一种具有可扩展性的 RFID 标签轻量级组证明协议[J]. 现代电子技术,2017,40(17):86 - 90.

[17] Xie R, Jian B Y, Liu D W. An improved ownership transfer for RFID protocol[J]. International Journal of Network Security,2018,20(1):149 - 156.

[18] Zhu F, Li P, Xu H, et al. A lightweight RFID mutual authenticationprotocolwithPUF[J]. Sensors, 2019, 19(13): 2957 - 2978.

[19] 史志才,王益涵,张晓梅,等. 一种具有隐私保护与前向安全的 RFID 组证明协议[J]. 计算机工程,2020,46(1):108 - 113.

[20] Liang W, Xie S Y, Long J, et al. A double PUF-based RFID identity authentication protocol in service-centricInternet of Things environments[J]. Information Sciences,2019,503:129 - 147.