Computer Science and Application 计算机科学与应用, 2022, 12(7), 1795-1800
Published Online July 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.127180

Hans汉斯

融合主题模型和图神经网络的无监督文档聚类模型

张出阳[1]，赵晓鹏[2]，柴变芳[1]

[1]河北地质大学信息工程学院，河北 石家庄
[2]河北省财政厅信息中心，河北 石家庄

收稿日期：2022年6月20日；录用日期：2022年7月19日；发布日期：2022年7月26日

摘 要

TextING (Inductive Text classification via GNN)模型是一种流行的图神经网络文本分类方法，其为每个文档构建词共现文档图，基于GCN (Graph Convolutional Networks)在所有文档词图上学习文档表示，进而通过监督的方式训练文档分类模型。但该方法需要大量文档类别标签，且基于词图的文档表示不能充分学到整个文档集合的全局特征。针对此问题，提出一种无监督的文本分类模型。该模型首先利用ETM (Embedd Topic Model)主题发现模型学习包含全局词特征的文档表示，并对ETM学到的文档主题表示进行Kmeans聚类作为文档的伪类标，再利用TextING训练文档分类模型。在真实文档数据集上的结果表明该方法比主流无监督文档聚类准确性高。

关键词

文档聚类，主题发现，图神经网络，词表示

Unsupervised Document Clustering Model Based on Topic Model and Graph Neural Network

Chuyang Zhang[1], Xiaopeng Zhao[2], Bianfang Chai[1]

[1]School of Information Engineering, Hebei GEO University, Shijiazhuang Hebei
[2]Information Center of Hebei Provincial Finance Department, Shijiazhuang Hebei

Received: Jun. 20th, 2022; accepted: Jul. 19th, 2022; published: Jul. 26th, 2022

Abstract

TextING (Inductive Text classification via GNN) model is a popular text classification method

文章引用: 张出阳, 赵晓鹏, 柴变芳. 融合主题模型和图神经网络的无监督文档聚类模型[J]. 计算机科学与应用, 2022, 12(7): 1795-1800. DOI: 10.12677/csa.2022.127180

based on graph neural network. According to the Graph Convolutional Networks (GCN), the document representation is learned on all the document word graphs, and then the document classification model is trained by supervision. However, this method requires a large number of document category labels, and the word graph-based document representation cannot fully learn the global characteristics of the entire document set. To solve this problem, an unsupervised text classification model is proposed. The Embedd Topic Model (ETM) was used to learn the document representation containing global word features. Kmeans clustering was applied to the document topic representation learned by ETM as the pseudo class standard of the document and then use TextING to train the document classification model. The results on real document datasets show that the proposed method is more accurate than the mainstream unsupervised document clustering.

Keywords

Document Clustering, Topic Discovery, Graph Neural Networks, Word Embeddings

1. 引言

当今社会正处于数据日益呈爆炸式增长的时代，从如此庞大的数据中发掘出反映出的现实问题的知识具有重要的意义。文本分类是一种典型的知识发现任务，传统的文本分类方法，比如贝叶斯、K 近邻和支持向量机等。这种方法主要依赖于人工提取的文本特征，再利用浅层模型实现文档的分类。

近年来，深度学习技术利用神经网络自动提取面向任务的文档特征，用于文档分类。基于图神经网络的方法可充分利用文本的关系学习文档表示，更好地用于文档分类。TextGCN [1]利用图神经网络对文本进行分类，解决了长序列和非连续词的交互问题，它为数据集构建成一个图，从全局学习节点特征，然而它在模型训练是需要庞大的参数和极大的空间和内存消耗，容易产生显存爆炸和梯度爆炸的问题。Text-level-gnn [2]解决了 TextGCN 内存消耗过大和泛化性能差的问题，它为每个文档单独的构建图来学习节点特征，但由于在构图时对于每队单词之间的边都是固定的，并不对所有单词之间具有适应性且因为从全局角度构图导致其在训练时的数据集中必须包含测试文件。TextING [3]考虑了每个文档细粒度的交互信息，为每个文档建立图从局部学到词节点特征，为每个文档计算文档级的向量表示用于学习文档分类模型，尽管该模型在构图方面有了很大改进，但其在数据集选取方面有很大限制，只有有标签数据集才能为其所用。

TextING 文本分类模型训练需要提供文档标签来指导模型的训练，而现实的数据集大多都是没有标签的。主题模型 ETM [4]能够从全局学到每个文档的主题表示，为得到更好的文档类别标签提供了更有效的特征表示。因此可以利用基于主题表示的文档聚类解决数据无标签问题，提出了一种融合 TextING 和 ETM 的文档聚类模型——TextING_ETM。其利用 ETM 学习每个文档的主题向量，通过聚类算法 Kmeans [5]对主题向量进行聚类[6]生成类标以此作为文档的伪标注，TextING 中利用伪标注对模型参数进行指导学习到文档的表示，以此保证同时利用了单词的局部和全局的信息来进行无监督的文本分类。

2. 文档聚类模型 TextING_ETM

文档聚类模型——TextING_ETM 框架如图 1 所示。首先利用 ETM 主题模型从词的全局角度学习数

据集中各个文档的主题表示，利用深度聚类[7]的思想，用 Kmeans 对主题表示的文档进行聚类得到文档伪类标。TextING 模型基于文档图表示学习词和文档表示，利用伪标注指导模型参数学习。

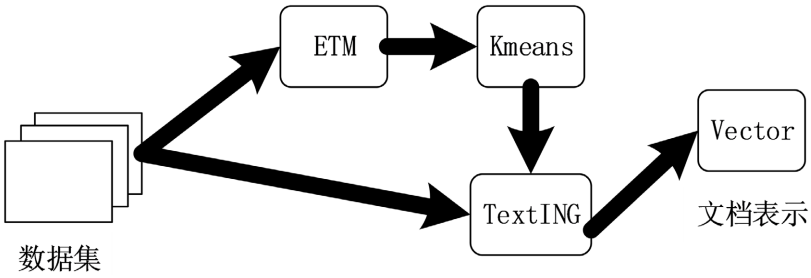

Figure 1. Model framework diagram
图 1. 模型架构图

2.1. 文档伪标签提取

基于词嵌入的主题模型 ETM 生成每个文档的过程如下：1) 根据 $LN(0,I)$ 为每个文档分配一个主题 θ_d；2) 对文档中的每个词：a) 指派一个主题 $Z_{dn} \sim Cat(\theta_d)$；b) 指派词的主题分布 $w_{dn} \sim \text{softmax}\left(\rho^{\mathrm{T}}\alpha_{Z_{dn}}\right)$。其中，$\rho$ 代表输入的词嵌入向量矩阵，$Cat(\cdot)$代表分类分布，$LN(\cdot)$代表逻辑斯蒂正态分布，NN 代表一个线性变换网络。基于 ETM 得到的文档主题表示，采用 Kmeans++，获取文档的类别标签。

2.2. 基于图神经网络的文档分类模型

首先基于文档词序列对每个文档构建一个文档图，顶点表示词，边表示词间存在共现关系(表示为 $G=(V,E)$，V 表示顶点集，E 表示边的集合)。首先利用预训练模型学习的词向量初始化图节点的词表示；然后基于词图聚合词共现词的特征来更新词的向量；最后根据词向量和软注意力机制得到文档表示，用于文档分类预测。在更新词向量时采用门控图神经网络的核心——GRU (Gated Recurrent Unit)对词表示进行充分更新，给定词嵌入矩阵 $h \in R^{|V|\times d}$，GRU 进行词向量更新过程如下：

$$a^t = Ah^{t-1}w_a, \tag{1}$$

$$z^t = \sigma\left(w_z a^t + u_z h^{t-1} + b_z\right), \tag{2}$$

$$r^t = \sigma\left(w_r a^t + u_r h^{t-1} + b_r\right), \tag{3}$$

$$h_t^{\sim} = \tanh\left(w_h a^t + u_h\left(r^t \odot h^{t-1}\right) + b_h\right), \tag{4}$$

$$h_t = h_t^{\sim} \odot z^t + h^{t-1} \odot \left(1 - z^t\right), \tag{5}$$

其中，a^t 代表初始的此向量矩阵，$\odot$ 代表哈达玛积，A 代表邻接矩阵，σ 是 sigmoid 激活函数，w,u,b 是可训练的权重参数和偏置系数，z 和 r 是更新门和遗忘门以确定邻居信息对当前节点嵌入的贡献程度，$h_t^{\sim}$ 是更新过程中的候选变量，h_t 是最终的词向量表示。

学到词向量表示后，基于注意力机制为各个词计算权重和最大池化层进行下采样操作最后计算出文档表示。计算公式如下：

$$h_v = \sigma\left(f_1\left(h_v^t\right)\right) \odot \left(f_2\left(h_v^t\right)\right), \tag{6}$$

$$h_G = \frac{1}{|V|}\sum_{v\in V} h_v + \text{Maxpooling}\left(h_1 \cdots h_V\right), \tag{7}$$

f_1 和 f_2 是两个多层感知机(MLP)，前一个计算的权重当作软注意力系数，后一个是特征非线性变换层。

最后，利用 softmax 层预测文档标签。模型优化目标为最小化交叉熵损失：

$$y_G^{\sim} = \text{softmax}\left(wh_G + b\right), \tag{8}$$

$$L = -\sum_i y_i \log \tilde{y}_i, \tag{9}$$

其中， y_i 是聚类出来的伪类标， $\tilde{y}_i$ 是图神经网络预测的类标。

3. 实验

首先介绍数据集和参数，然后用标准互信息化分数(NMI)和纯度(purity)评比模型。

3.1. 数据集及实验设置

本实验在训练过程中使用了 R8 和 20NewsGroup 数据集对模型结果进行了验证。R8 数据集包括 8 类路透社新闻数据，共有 7674 篇文档。20NewsGroup 数据集中约有 20,000 篇新闻组文档，均匀分为 20 个不同主题的新闻组集合。文本以标准方式进行预处理，包括标记化和停止字删除。首先使用 NLTK 中的 Stopwords 对数据集进行停用词过滤，然后将数据集中小于等于 2 字符长度的词并且文档中的各词词频小于 5 的词移除，接着把数据集中所有空文档和长度为一的文档删掉。经过处理 R8 数据集剩余文档 7674 篇，不重复词数 3562 个，20NewsGroup 数据集剩余文档 18,846 篇，不重复次数 18,126 个。在图神经网络构图时设置文档最大长度为 800，将长文本进行截断，设置滑动窗口的大小为 1。表 1 中详细介绍了实验过程中使用的数据集名称，训练时训练集和测试集的文档个数以及各数据集的类别。

硬件方面采用 Intel(R) Core(TM) i7-6500U CPU，8 G 内存，NVIDIAGeForce 940MX 显卡。本实验所有词向量维度都设置为 300 维，权重设置为 0.00001，学习率设置为 0.001，丢弃率设置为 0.5，TextING 的隐藏单元设置为 96，批量大小设置为 1000。

Table 1. Samples and category of dataset
表 1. 数据集的样本数及类别

数据集	训练集	测试集	类别
R8	6522	1152	8
20NewsGroup	16,012	2834	20

3.2. 无监督文本分类效果评估

本文通过标准互信息化分数(NMI)和纯度(purity) [8]对模型效果进行评比，NMI 常用于评测模型的聚类性能，purity 经常用来评测无监督模型的分类性能。NMI 和 purity 始终在 0.0 到 1.0 之间，且分数越高代表模型在下游任务的表现越好。

各数据集的主题数 Topics 对模型训练的结果有很大的影响，为了找到合适的主题数(以下简写为 T)，进行了以下实验。

表 2 和表 3 列出了 Topics 的不同取值在同一数据集上的 purity 值也不同，由以上两表可以看出在 20NG 数据集上取 Topics = 40 模型训练结果最好，在 R8 数据集上取 Topics = 30 模型训练结果最好。因此在之后的模型训练时对于 20NG 数据集设定主题数为 40，为 R8 数据集设定主题数为 30。

表 4 和表 5 说明了本文中的模型与现有的模型在相同数据集上的对比，本文提出的方法在无监督分类任务上明显优于之前已有的方法。与之前的模型相比，本文的方法加入了丰富的词网络信息，对无标

签信息的数据集进行分类可以得到不错的结果。

Table 2. Impact of changes in Topics on results of 20NG dataset
表 2. 20NG 数据集上 Topics 的变化对结果的影响

数据集	Purity			
	模型	T = 10	T = 20	T = 40	T = 80
20NG	LDA	0.1601	0.3094	**0.3137**	0.3102
	ETM	0.1612	0.3301	**0.3347**	0.3314
	TextING_TM	0.1628	0.3448	**0.3491**	0.3443

Table 3. Impact of changes in Topics on results of R8 dataset
表 3. R8 数据集上 Topics 的变化对结果的影响

数据集	Purity			
	模型	T = 15	T = 30	T = 45	T = 60
R8	LDA	0.6457	**0.7185**	0.7145	0.7165
	ETM	0.6712	**0.7395**	0.7365	0.7334
	TextING_TM	0.6814	**0.7588**	0.7416	0.7324

Table 4. NMI comparison results of algorithms on real datasets
表 4. 真实数据集上算法的 NMI 对比结果

数据集	LDA	ETM	TextING_TM
R8	0.4112	0.4779	**0.4926**
20NG	0.3610	0.3778	**0.3888**

Table 5. Purity comparison results of algorithms on real datasets
表 5. 真实数据集上算法的 Purity 对比结果

数据集	LDA	ETM	TextING_TM
R8	0.7185	0.7395	**0.7588**
20NG	0.3137	0.3347	**0.3491**

4. 结论

本文提出了一个无监督文本聚类模型。其基于文档主题表示聚类得到文档伪类标，利用图神经网络学习更好的文档表示。与主流模型在真实数据集上的对比表明：其在无监督分类任务上具有更高的准确性。

基金项目

河北省高等学校科学技术研究项目(ZD2020175)；河北省高校基本科研业务费资助；河北地质大学科技创新团队项目资助(KJCXTD-2021-11)；河北省重点研发计划项目(项目名称：基于时空大数据及深度学习的地质灾害风险识别关键技术研究编号：22375415D)。

参考文献

[1] Yao, L., Mao, C. and Luo, Y. (2018) Graph Convolutional Networks for Text Classification. 33*rd AAAI Conference on Artificial Intelligence* (*AAAI* 2019). https://doi.org/10.1609/aaai.v33i01.33017370

[2] Huang, L., Ma, D., Li, S., *et al.* (2019) Text Level Graph Neural Network for Text Classification. In: *Proceedings of the* 2019 *Conference on Empirical Methods in Natural Language Processing and the* 9*th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*), Hong Kong, 3444-3450. https://doi.org/10.18653/v1/D19-1345

[3] Zhang, Y., Yu, X., Cui, Z., *et al.* (2020) Every Document Owns Its Structure: Inductive Text Classification via Graph Neural Networks. In: *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 334-339. https://doi.org/10.18653/v1/2020.acl-main.31

[4] Dieng, A.B., Ruiz, F. and Blei, D.M. (2020) Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, **8**, 439-453. https://doi.org/10.1162/tacl_a_00325

[5] Caron, M., Bojanowski, P., Joulin, A. and Douze, M. (2018) Deep Clustering for Unsupervised Learning of Visual Features. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Computer Vision—ECCV* 2018. *Lecture Notes in Computer Science*, Vol. 11218, Springer, Cham. https://doi.org/10.1007/978-3-030-01264-9_9

[6] Li, X., Zhang, H. and Zhang. R. (2021) Adaptive Graph Auto-Encoder for General Data Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**, 1. https://doi.org/10.1109/TPAMI.2021.3125687

[7] Sun, K., Lin, Z. and Zhu, Z. (2020) Multi-Stage Self-Supervised Learning for Graph Convolutional Networks on Graphs with Few Labeled Nodes. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 5892-5899. https://doi.org/10.1609/aaai.v34i04.6048

[8] Nguyen, D.Q., Billingsley, R., Du, L., *et al.* (2018) Improving Topic Models with Latent Feature Word Representations. *Transactions of the Association for Computational Linguistics*, **3**, 299-313.