Computer Science and Application 计算机科学与应用, 2023, 13(10), 1918-1927
Published Online October 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.1310190

Hans汉斯

基于MS-Cluster与Prompt-Learning话题检测与追踪技术

李　崭，杜晓童，黄　浩，任秋霖

中国电子科技集团公司第十研究所，四川 成都

收稿日期：2023年9月16日；录用日期：2023年10月16日；发布日期：2023年10月24日

摘　要

话题检测与追踪技术随着信息处理技术以及人工智能技术的发展，已经取得了较好的发展，但在实际应用中，由于算法标注数据需求高、训练代价大，很难较好的落地应用。本文提出了基于MS-Cluster与Prompt-Learning的话题检测追踪技术，通过聚类分析过程初步进行话题聚合，在此基础上通过提示学习推理进行话题补偿，完成话题检测与追踪过程。该方法在包含13个话题的测试数据集上进行测试验证，证明该方法在零样本与低样本标注情况下有较好效果，同时相较于其他主流话题检测追踪技术在准确率与召回率上都有提升。

关键词

话题检测追踪技术，提示学习，小样本学习，聚类分析

Topic Detection and Tracking Technology Based on MS-Cluster and Prompt-Learning

Zhan Li, Xiaotong Du, Hao Huang, Qiulin Ren

The 10th Research Institute of China Electronics Technology Group Corporation, Chengdu Sichuan

Received: Sep. 16th, 2023; accepted: Oct. 16th, 2023; published: Oct. 24th, 2023

Abstract

Topic detection and tracking technology has been developing well with the development of information processing technology and artificial intelligence technology. However, in practical applications, it is difficult to achieve good deployment due to the high demand for algorithm annotated data and the large training cost. This article proposes a topic detection and tracking technology

文章引用: 李崭, 杜晓童, 黄浩, 任秋霖. 基于 MS-Cluster 与 Prompt-Learning 话题检测与追踪技术[J]. 计算机科学与应用, 2023, 13(10): 1918-1927. DOI: 10.12677/csa.2023.1310190

based on MS-Cluster and Prompt-Learning. The method performs topic aggregation through clustering analysis and topic supplementation through prompt learning reasoning to complete the topic detection and tracking process. The method was tested on a dataset of 13 topics, and it showed good results in the case of zero-shot learning and few-shot learning, and it outperformed other mainstream topic detection and tracking technologies in terms of accuracy and recall rate.

Keywords

Topic Detection and Tracking Technology, Prompt-Learning, Few-Shot Learning, Clustering Analysis

1. 引言

话题检测与追踪技术[1] (Topic Detection and Tracking, TDT)是近年提出的一项信息处理技术，这项技术旨在帮助人们应对日益严重的互联网信息爆炸问题，对新闻媒体信息流进行新话题的自动识别和已知话题的持续追踪。由于网络信息数量庞大，形式多样、传播迅速，互联网新闻报道冗余多、议题发散、易漂移，与一个话题相关的信息往往孤立地分散在很多不同的地方并且出现在不同的时间，仅仅通过这些孤立的信息，人们对某些话题事件难以做到全面的把握。通过话题检测与追踪技术，能够达成针对一个话题实现多维度、多时间节点的聚合关联，实现新话题的自动识别和已知话题的持续追踪，提高信息获取的价值。当前，话题检测与追踪主要研究方向是通过对文本数据流的文本数据进行识别，对数据的边界进行划分，从而实现突发性话题的发现检测、话题的发展追踪以及话题发展变化的探测。

话题检测与追踪技术研究起始于上世纪 90 年代，经过 30 年左右的发展[2]，由于其任务具有主题类别未知性、数据突发性等特点，形成核心解决方案思路主要包括两大类：非监督学习任务预测[3]与监督学习任务预测[4]。

基于非监督学习任务预测[5]的方法主要是采用主题模型[6]、聚类分析等机器学习过程，在数据特征、主题特征[7]学习表征基础上，通过非监督预测过程对特征相似的数据进行聚合，实现数据的自主的划分，形成话题脉络。基于监督学习任务预测的方法可分为多分类模型以及序列分类模型。通过分类标签预测，在标签基础上对数据进行组织聚合，形成话题检测追踪结果。

针对上述两种思路，都存在一定局限性，非监督任务预测过程中，由于一般不存在参数最优化学习过程导致预测效果存在较大瓶颈；监督任务预测过程中则需要大量高质量标注语料进行训练，且预测数据类别与数据标签需要与训练数据有较高的拟合关联程度，否则预测效果将无法达到预期。

综上所述，话题检测与追踪技术当前技术瓶颈是需要实现低资源学习情况下达到较好的预测效果。这样使得话题检测与追踪技术在不同的样本数据与不同领域的应用分析中，可以通过较少的数据标注干预，达到预期效果。

2. 技术现状

话题检测与追踪技术当前主要研究集中在监督任务预测与非监督任务预测上。监督任务预测是通过将话题检测任务构建成为事件探测、提取、分类等任务进行识别[8]，再对数据进行组织聚合形成话题检测结果。其中，Bekoulis [9]等提出了一种子事件话题检测模型，通过文本信息的时序性，将检测任务构

建成为序列标记任务进行预测分析。Araki [10]等提出了一种基于逻辑回归的多分类器，通过特征工程构建进行事件间的关系识别，由此进行话题事件的检测分析。但监督学习任务存在标注需求量十分巨大，且由于采用分类标签监督学习形式，不能对开发域话题进行扩展等问题，导致其领域应用难度较高。

非监督任务预测是通过将话题检测任务构建成为聚类任务、主题发现任务等，通过对数据特征化，基于特征相似的数据为同一话题的假设下进行推理，实现话题检测追踪任务。其中，张帆[11]等人提出了一种改进的 Single-Pass 聚类算法，通过时间特征以及特征词汇的表征改进，在话题分析任务中取得了较好效果。Wartena [12]等人提出了通过关键词聚类分析进行主题聚合的算法，核心是通过相似计算与主题词汇提取改善话题检测效果。Xie [13]等人在话题检测研究中，提出通过在高维向量空间中，映入特征选择与激励机制，优化信息特征表达。张小明[14]等，提出了一种增量聚类算法进行自动话题检测，通过增量聚合的模式在验证中不仅一定程度上提升了准确率，还通过增量计算减少了计算代价。Ge [15]等提出了一种主题分析模型进行话题检测，其主题分析模型采用构建过程中通过采用关键短语代替独立词作为特征，实现主题模型对文本主题特征的优化，优化话题检测过程。Pang [16]等人提出了针对短文本的话题检测的新模型，可以通过词汇的共现网络构建实现信息间关联从而进行主题信息检测。非监督学习任务通过聚合流程优化以及特征优化，一定程度上可以提高话题检测追踪的效果，但针对话题信息内涵发散、漂移，很难通过非监督任务对信息进行聚合，使得话题检测追踪效果不佳。

针对以上问题，本文提出了一种基于 MS-Cluster 与 Prompt-Learning 的话题检测追踪技术，结合的监督学习与非监督学习技术，采用聚类分析与特征学习进行话题的聚合分析，在此基础上通过预训练模型的提示构建[17]与预测对话题聚合结果进行补偿。实验结果表明，本文提出的方法在零样本与少样本[18][19]标注情况下，大幅度提升了话题检测的效果，充分证明了方法的有效性，下面将详细介绍该技术。

3. 基于 MS-Cluster 与 Prompt-Learning 的话题检测追踪

基于 MS-Cluster (Multi-Section Cluster)与 Prompt-Learning 的话题检测追踪技术包括以下几个步骤：(1) 数据特征化：对文本数据进行特征学习表征，分为语义特征学习与时间特征学习。针对语义特征采用 Word2Vec 模型[20]作为文本语义特征学习以及表达模型，Word2Vec 模型通过对输入词汇的上下文词汇进行预测，实现语义信息的学习，采用三角核函数对时间特征进行编码，实现时间特征表达，最后将时间特征与语义特征进行拼接得到文本特征实现文本特征学习；(2) 聚类分析：基于文本特征表达，对采用 MS-Cluster 文本数据集合进行聚类分析，得到基于聚类分析的话题聚合分析结果；(3) 话题补偿推理：基于预训练模型提示工程构建与 prompt-learning，通过话题间的关系推理，对话题完备性进行补偿，得到话题检测与追踪结果。下面将详细介绍算法核心步骤。

3.1. 数据特征化

数据特征化是通过文本语义特征模型表达的语义特征与文本时间特征拼接，得到数据特征学习表达结果。本文文本语义特征化采用 Word2Vec 模型进行特征计算，时间特征化采用三角核函数进行特征表达。

3.1.1. 语义特征学习

语义特征模型采用 Word2Vec 模型，模型通过输入词汇对其上下文词汇进行预测的学习任务，实现文本语义特征学习。首先对文本进行分词处理，对分词结果进行 one-hot 编码，得到词汇的 one-hot 特征向量 x_k，其中 x_k 的维度为 $1*V$ 。其次，初始化编码矩阵 w^I_{V*N} ，矩阵中参数初始化采用随机初始化，其中 w^I 表示编码矩阵，矩阵维度为 $V*N$ 。通过计算公式：$h_k = x_k * w^I_{V*N}$ ，得到特征向量 h_k，其中 h_k 表示 one-hot 特征向量 x_k 通过编码矩阵 w^I_{V*N} 进行特征降维的向量，其矩阵维度为 $1*N$ 。初始化解码矩阵 w^O_{N*V} ，矩阵

中参数初始化采用随机初始化，通过计算公式：$y_k^j = h_k * w_{N*V}^O$ 得到词汇的解码 one-hot 特征向量 y_k^j，其中 k 表示输入词汇的索引，j 表示需要预测的上下文词汇的索引。然后，通过采用函数 softmax 将输出的特征向量 y_k^j 进行归一化处理，得到概率分布特征向量 p_k^j，对 p_k^j 与词汇 j 的 one-hot 特征向量采用交叉熵进行误差衡量，通过最小化交叉熵对词汇的上下文学习，实现文本的语义学习。模型采用一个词汇对其上下文总共 C 个词汇进行预测学习，其损失函数为：

$$\text{Loss} = \sum_{j \in C} -x_j \log p_k^j$$

其中 x_j 为词汇 j 的 one-hot 特征向量，C 表示词汇 k 的上下文词汇。然后通过最小化损失函数与 BP 算法，对模型的编码矩阵与解码矩阵的参数进行更新，完成模型训练，得到文本语义模型。

完成语义模型训练后，对文本的语义特征进行表达。首先对文本进行分词处理以及停用词过滤处理，再对得到的文本词汇集合进行频率统计，得到 N_{word} 个词汇，则文本特征向量 v_{text} 计算公式为：

$$v_{text} = \frac{f_i * \sum_{i=1}^{N_{word}} v_i}{\left| \sum_{i=1}^{N_{word}} v_i \right|}$$

其中，v_i 为文本中的第 i 个特征词汇的特征向量，f_i 为特征词汇的出现频率。

3.1.2. 时间特征表达

时间特征学习采用三角核函数，对时间特征进行特征构建，通过三角变化公式特征化，可得时间相似度计算公式为：

$$Sim_t = \left(\frac{\alpha}{N_{td}} \right)^2 * \sum_{i=1}^{T_d} \left(\cos\left(\frac{\pi}{2} * \frac{t_a - t_b}{T_{span}} * \frac{i}{N_{td}} \right) \right)$$

其中 $T_{span} > t_a - t_b, t_a > t_b$ 即三角函数内的取值范围在 $[1, \pi/2]$ 且单调递减，可推导出 Sim_t 随着 $t_a - t_b$ 单调递减，使得时间特征化符合话题的分布特性，时间越相近则相似度越高，时间越相远则相似度越低。α 为时间特征权重因子，T_{span} 为时间跨度长度，t_a，t_b 分别表示文本 a、b 的时间信息。通过对相似计算公式进行展开，得到时间特征向量 $V_{timeemb} = \left(U_1, U_2, \cdots, U_{2N_{td}-1}, U_{2N_{td}} \right)_{1*N_{timeemb}}$，具体如下：

$$U_{2i-1} = \frac{\alpha}{N_{td}} * \cos\left(\frac{\pi}{2} * \frac{t}{T_{span}} * \frac{i}{N_{td}} \right) \quad i \in [1, N_{td}]$$

$$U_{2i} = \frac{\alpha}{N_{td}} * \sin\left(\frac{\pi}{2} * \frac{t}{T_{span}} * \frac{i}{N_{td}} \right) \quad i \in [1, N_{td}]$$

3.2. MS-Cluster 聚类分析

MS-Cluster 聚类分析包括三个核心过程，包括：聚类初始化，聚类划分以及聚类终止三个过程。

3.2.1. 聚类初始化

聚类初始化过程是对数据集合中的数据进行特征计算，形成特征向量集合，用于后续聚类分析计算。其中，设参与话题聚合的文本数量为 i，每篇文章的特征向量 V_{emb}^i，对每篇单独形成一个聚类点，对初始聚类点进行聚合，形成一个簇中聚类点数量为 i 的聚类簇，完成聚类初始化。

3.2.2. 聚类划分

聚类划分过程是对每个待划分的聚类簇进行裂变，形成多个新的聚类簇的过程。其中，设裂变的数

量为 N_{dis}，根据当前聚类簇中的聚类点分布情况，对聚类簇进行中心点推举，将类簇中推举 N_{dis} 个中心点作为新的聚类中心点。推举方式采用聚类点价值评估算法，首先推选候选中心点，以聚类点局部密度 $pi > pi_{scoremin}$ 为条件推选候选中心点，其中局部密度计算公式为：

$$pi = \sum_{dist<dist_{min}} \frac{1}{1+\frac{1}{dist}}$$

其次，根据推选出的 K 个中心点，计算每个中心点的评估价值，价值计算公式为：

$$V_{score}^{a} = pi_a * \sum_{i\neq a, i\in[1,k]} e^{\left(\frac{dist_{a,i}}{dist_{min}}\right)^2}$$

通过价值评估，得到 N_{dis} 个聚类中心点，完成中心点推举，其中，$pi_{scoremin}$ 为局部密度的最小阈值，$dist_{min}$ 为局部密度计算的最小距离，$dist_{a,i}$ 为聚类点 a 至聚类点 i 的距离。其中，聚类点间的距离计算公式为：

$$dist = 1 - \frac{vec_a \cdot vec_b}{|vec_b| * |vec_b|}$$

在此基础上通过 N_{dis} 个聚类中心点，进行类簇划分。非聚类中心点选择相似度最高的聚类中心点加入，形成类簇，类簇形成后，更新类簇中心，对非聚类中心点重新计算类簇划分，迭代此计算过程，直到所有的非聚类中心点不再更新其所属类簇，完成类簇划分，形成新的 N_{dis} 个聚类簇。其中聚类中心更新计算公式为：

$$v_{center} = \frac{\sum_{i\in c_{cluster}} Vec_i}{\left|\sum_{i\in c_{cluster}} Vec_i\right|}$$

其中 $c_{cluster}$ 为当前类簇中所有聚类点，Vec_i 为当前类簇中第 i 个聚类点的特征向量，聚类点间的相似度计算公式为：

$$Sim = \frac{vec_a \cdot vec_b}{|vec_a| * |vec_b|}$$

其中，vec_a、vec_b 分别表示文本 a 与文本 b 特征向量。

3.2.3. 聚类终止

聚类终止过程是对每个划分的聚类簇进行评分，检测其是否可以停止继续划分子类。其中，设聚类增益最小阈值为 g_{errmin}，判断聚类簇中的信息增益是否大于设定的最小阈值，当聚类增益 $g_{err} < g_{errmin}$，终止当前的当前类簇划分；当聚类增益 $g_{err} > g_{errmin}$，对当前类簇进行划分，得到子类簇，同时对划分生成每个新生成的类簇进行聚类划分，直至所有类簇聚类增益 $g_{err} < g_{errmin}$，完成所有类簇的聚类划分。

其中，计算类簇划分对类簇带来的误差增益，其计算公式如下所示：

$$g_{err} = Err_{cluster}^{c} - \sum_{i\in cN_{dis}} Err_{cluster}^{i}$$

其中 $Err_{cluster}^{c}$ 为划分前聚类簇的聚类簇误差，$\sum_{i\in cN_{dis}} Err_{cluster}^{i}$ 为新划分的 N_{dis} 类簇的聚类误差和。聚类误差的计算公式为：

$$Err_{center} = \sum_{i\in c_{cluster}} \left(1 - \frac{vec_{center} \cdot vec_i}{|v_{center}| * |vec_i|}\right)$$

其中，$c_{cluster}$ 为当前类簇中所有聚类点，v_{center} 为当前类簇的中心聚类点，vec_i 为当前类簇中聚类点的特征向量。

3.3. 话题补偿推理

本文在话题补偿推理阶段，针对聚类产生的各个类簇进行类簇间的话题关系推理，通过推理结果对话题进行聚合补偿，增强话题聚合程度。在话题补偿任务中，本文采用文本生成模型，通过生成模型的标签概率映射，计算话题的关联性，实现话题的零样本推理学习(zero-shot)以及少样本推理学习(few-shot)。

话题补偿推理采用 ERNIE 预训练模型，在此基础上通过 prompt 模板构建文本推理任务，推理两篇文本的话题相关性，并基于 prompt-learning 进行模型微调，优化话题推理效果。在此基础上，将文本间的话题相关性，通过话题间相似推理计算公式，映射到话题与话题间的相关性推理，其公式如下所示：

$$SimTopic_{a,b} = \frac{\sum_{i\in a, j\in b} same\_topic_{i,j}}{n_a * n_b}$$

其中 a，b 分别表示不同的两个话题，$\sum_{i\in a, j\in b} same\_topic_{i,j}$ 表示 a，b 两个话题中两两属于相同话题的数据求和数量，n_a 与 n_b 表示话题 a，b 中的数据总量。通过话题间的相关性推理，将高度近似的话题进行聚合，提高话题的完备性，实现话题的补偿推理。

4. 实验

4.1. 数据集

本文采用的数据集为自筹数据集合，其中数据来源于包括新浪微博、网易新闻等社交媒体以及新闻门户网站，通过数据爬虫，数据清洗，形成测试数据集。其中，数据集时间跨度从 2012 年至 2019 年，包括 13 个话题，共 2957 条数据，每条数据包括新闻标题、新闻内容、新闻时间，数据集话题信息具体分布如图 1 所示：

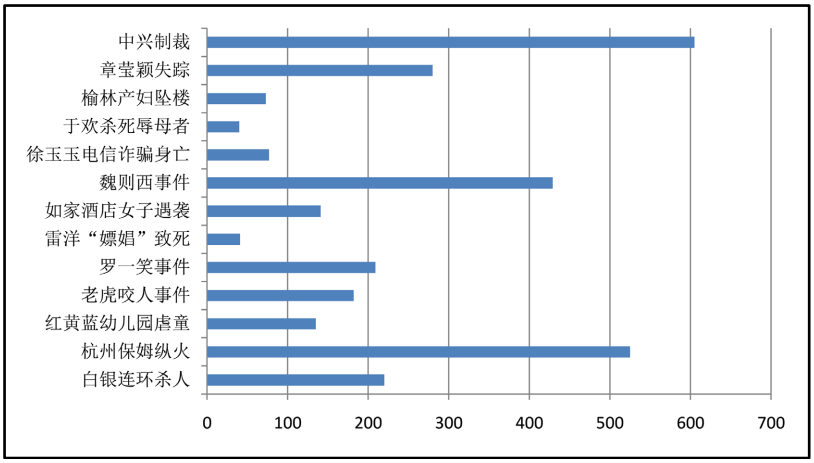

Figure 1. Dataset topic distribution map
图 1. 数据集话题分布图

4.2. 评估指标

本文在话题检测追踪任务采用的评价指标主要采用 2 个指标，Precision(准确率)以及 Recall(召回率)，其具体计算公式如下所示：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

在计算公式中，TP 表示两个样本属于同一话题并正确计算成为同一话题的数量，FP 表示两个样本不属于同一话题并错误计算成为同一话题的数量，FN 表示两个样本属于同一话题并错误计算没有成为同一话题的数量。

4.3. 实验结果与分析

4.3.1. 模型实验结果

针对本文提出的话题追踪算法，对算法的聚类分析过程以及话题补偿过程分别进行实验验证。对于聚类分析过程，核心参数包括两个部分：一、针对聚类过程中聚类中心选择上，有两种模式可以进行选择，通过价值计算进行中心点的推选以及随机选择两种模式；二、针对聚类停止条件增益最小阈值为 g_{errmin} 选择，该值主要用于控制聚类结束条件。聚类分析过程实验验证结果如图 2 所示：

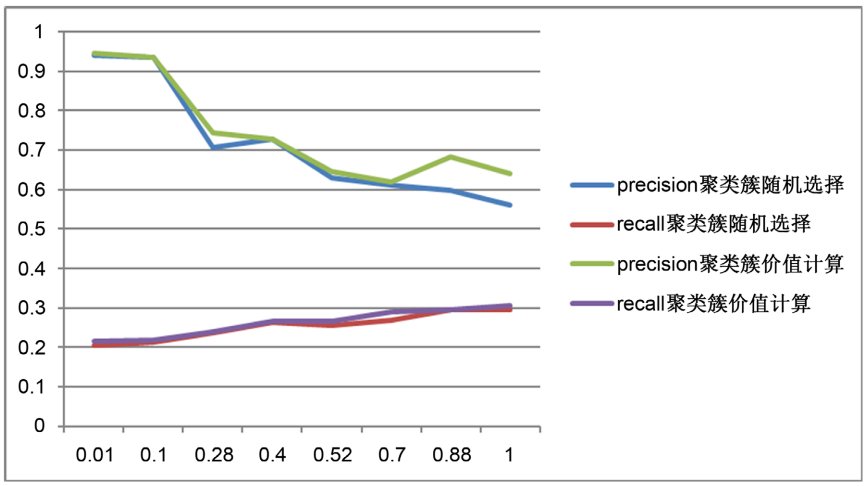

Figure 2. Parameter adjustment effect
图 2. 调参效果图

从结果可以看出，针对聚类中心选择模式上，价值计算中心点推选的模式相较于随机选择的模式在 Precision 与 Recall 指标上有较少的提升，但提升并不明显。针对聚类停止条件增益最小阈值为 g_{errmin}，随着增益最小阈值为 g_{errmin} 的增加，聚类结果的 Precision 指标明显下滑，Recall 指标缓慢上升。由此实验可见，聚类过程中能够达到较高的 Precision 指标，通过结果具体分析得知，主要体现在聚类分析在短时话题热点的探测聚合上有较好的效果；但聚类过程中的 Recall 指标提升很难，通过结果具体分析得知，主要体现在聚类分析在长时连续话题的追踪聚合上很难达到较好的效果。

对于话题补偿过程，核心是验证采用预训练模型 + Prompt-Tuning 特征学习模式，能否增强聚类分析结果的聚合程度，以提升算法在长时连续话题的追踪聚合能力。同时，验证零样本学习与小样本学习在话题分析补偿中的效果提升。在实验中，聚类分析过程采用的算法参数如下，增益最小阈值设为 g_{errmin} 值 0.01 以及聚类中心选择采用价值计算模式。话题补偿过程实验验证结果如图 3 所示。

从结果可以看出，在话题补偿阶段，随着小样本学习的样本数量提升，话题检测与追踪结果的 Precision 指标出现微量的下滑，其主要是由于话题补偿阶段的话题聚合，引入了少量误差数据引起

Precision 指标下滑。同时，话题检测与追踪结果的 Recall 指标大幅度提升，在零样本、小样本的情况下都有较好的效果。

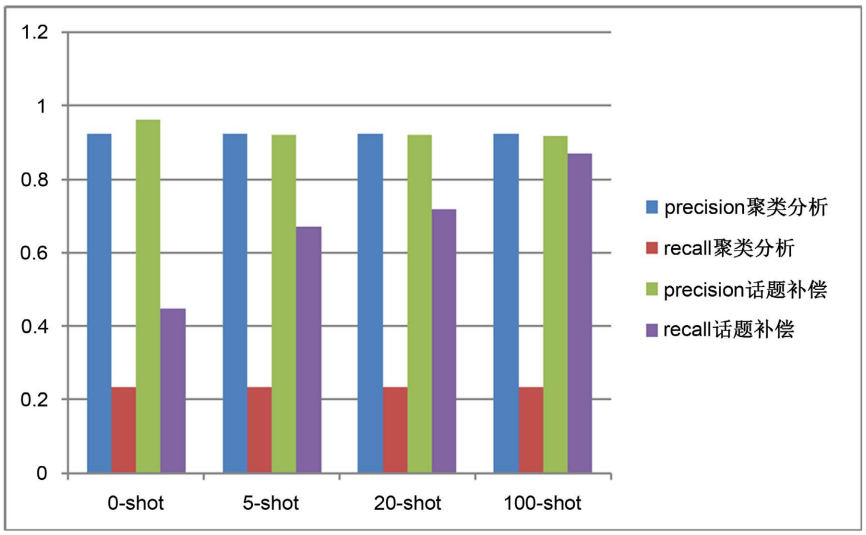

Figure 3. Topic compensation effect
图 3. 话题补偿效果图

4.3.2. 对比实验结果

话题检测与追踪任务在数据集上的比对实验结果如图 4 所示，包含了基于改进的 Single-Pass 话题检测追踪算法[11]、基于增量型聚类的话题检测算法[14]、基于 Word2Vec 与 K-means 话题检测追踪算法[21]。可以明显看出，本文提出的模型在评估指标中均取得了最优的实验测试结果，在 Precision 指标上提升 10%以上，在 Recall 指标上提升 20%以上。由此可以得出，通过聚类分析的话题聚合以及提示学习的话题补偿，相较于基于传统的聚类算法与分类算法的话题检测追踪算法，在 Precision 指标有部分幅度提升同时在 Recall 指标上有大幅度的提升，这充分证明了本文提出方法的优越性。

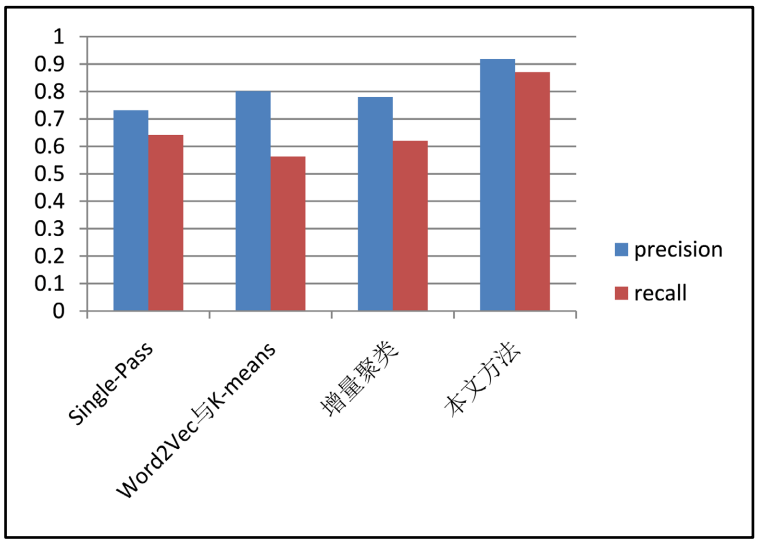

Figure 4. Comparison of experimental results
图 4. 比对实验结果图

5. 结束语

本文提出的基于 MS-Cluster 与 Prompt-Learning 的话题检测追踪技术在测试数据集上取得了最优效果，通过实验可以看出通过结合聚类分析的话题热点发现能力以及提示学习的话题补偿能力可以大幅度提升话题检测追踪的能力。同时，本文中对零样本以及少样本的话题补偿能力进行了测试，这使得算法在工程实际应用中对数据样本的标注量需求更低，便于算法的应用落地。

本文提出的话题检测追踪技术结合聚类分析以及小样本学习，使得低标注资源下算法效果能达到较好的效果，为后续相关话题检测追踪研究提供相关的参考。

参考文献

[1] Liu, P., Yuan, W., Fu, J., *et al*. (2021) Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing.

[2] Nallapati, R., Feng, A., Peng, F.C. and Allan, J. (2004) Event Threading within News Topics. *Proceedings of the* 13*th ACM Conference on Information and Knowledge Management*, Washington DC, 8-13 November 2004, 446-453. https://doi.org/10.1145/1031171.1031258

[3] Lim, K.W. and Buntine, W. (2014) Twitter Opinion Topic Model: Extracting Product Opinions from Tweets by Leveraging Hashtags and Sentiment Lexicon. *Proceedings of the* 23*rd ACM International Conference on Information and Knowledge Management*, Shanghai, 3-7 November 2014, 1319-1328. https://doi.org/10.1145/2661829.2662005

[4] 苏金树, 张博锋, 徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报, 2006, 17(9): 1848-1859.

[5] 黄卫东, 陈凌云, 吴美蓉. 网络舆情话题情感演化研究[J]. 情报杂志, 2014(1): 102-107.

[6] Huang, S., Yang, Y., Li, H. and Sun, G.Z. (2015) Topic Detection from Microblog Based on Text Clustering and Topic Model Analysis. 2014 *Asia-Pacific Services Computing Conference*, Fuzhou, 4-6 December 2014, 88-92. https://doi.org/10.1109/APSCC.2014.18

[7] Pavlinek, M. and Podgorelec, V. (2017) Text Classification Method Based on Self-Training and LDA Topic Models. *Expert Systems with Applications*, **80**, 83-93. https://doi.org/10.1016/j.eswa.2017.03.020

[8] Aldawsari, M. and Finlayson, M.A. (2019) Detecting Subevents Using Discourse and Narrative Features. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*, Florence, July 2019, 4780-4790. https://doi.org/10.18653/v1/P19-1471

[9] Bekoulis, G., Deleu, J., Demeester, T., et al. (2019) Sub-Event Detection from Twitter Streams as a Sequence Labeling Problem. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Minneapolis, June 2019, 745-750. https://doi.org/10.18653/v1/N19-1081

[10] Araki, J., Liu, Z., Hovy, E., et al. (2014) Detecting Subevent Structure for Event Coreference Resolution. *Proceedings of the* 9*th International Conference on Language Resources and Evaluation*, Reykjavik, May 2014, 4553-4558.

[11] 张帆, 潘亚雄, 胡勇, 等. 基于改进 single-pass 算法的新闻话题检测与追踪技术研究[J]. 信息安全研究, 2020, 6(5): 396.1-396.8.

[12] Wang, M., Jayaraman, P.P., Solaiman, E., *et al*. (2018) A Multi-Layered Performance Analysis for Cloud-Based Topic Detection and Tracking in Big Data Application. *Future Generation Computer Systems*, **87**, 580-590. https://doi.org/10.1016/j.future.2018.01.047

[13] Xie, J., Liu, G.S. and Ning, W. (2012) A Topic Detection Method for Chinese Microblog. 2012 *Fourth International Symposium on Information Science and Engineering*, Shanghai, 14-16 December 2012, 100-103. https://doi.org/10.1109/ISISE.2012.30

[14] 张小明, 李舟军, 巢文涵. 基于增量型聚类的自动话题检测研究[J]. 软件学报, 2012, 23(6): 1578-1587.

[15] Ge, B., He, C.H., Hu, S.Z. and Guo, C. (2018) Chinese News Hot Subtopic Discovery and Recommendation Method Based on Key Phrase and the LDA Model. *Proceedings of the* 2018 *International Conference on Electrical, Control, Automation and Robotics*, Xiamen, 16-17 September 2018, 349-358. https://doi.org/10.12783/dtetr/ecar2018/26371

[16] Pang, J.H., Li, X.S., Xie, H.R. and Rao, Y.H. (2016) SBTM: Topic Modeling over Short Texts. In: Gao, H., Kim, J. and Sakurai, Y., Eds., *DASFAA* 2016: *Database Systems for Advanced Applications*, Springer, Cham, 43-56. https://doi.org/10.1007/978-3-319-32055-7_4

[17] Liu, P.F., Yuan, W.Z., Fu, J.L., Jiang, Z.B., Hayashi, H. and Neubig, G. (2021) Pre-Tain Prompt and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, **55**, 1-35.

[18] 黄彦乾, 迟冬祥, 徐玲玲. 面向小样本学习的嵌入学习方法研究综述[J]. 计算机工程与应用, 2022, 58(3): 34-49.

[19] 赵凯琳, 靳小龙, 王元卓. 小样本学习研究综述[J]. 软件学报, 2021, 32(2): 349-369.

[20] Mikolov, T., Sutskever, I., Kai, C., Corrado, G.S. and Dean, J. (2013) Distributed Representations of Words and Phrases and Their Compositionality. *Advances in Neural Information Processing Systems*, **26**, 3111-3119.

[21] 王立平, 赵晖. 融合词向量与关键词提取的微博话题发现[J]. 现代计算机, 2020(23): 3-9.