计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2023/17(12)-2861-19
doi: 10.3778/j.issn.1673-9418.2303083

主题模型自动标记方法研究综述

何东彬[1],陶　莎[2],朱艳红[3],任延昭[4],褚云霞[1+]
1. 石家庄学院 河北省物联网安全与传感器检测工程研究中心,石家庄 050035
2. 中国农业大学 农业农村部农业信息化标准化重点实验室,北京 100083
3. 石家庄邮电职业技术学院 河北省物联网智能感知与应用技术创新中心,石家庄 050021
4. 北京工商大学 计算机与信息工程学院,北京 100048
+ 通信作者 E-mail: 53945776@qq.com

摘　要:主题模型常用于非结构化语料库和离散数据建模,抽取隐含主题分布。由于主题发现结果采用词列表形式,理解其含义较为困难。尽管通过人工标记可生成更具解释性和易理解的主题标签,但成本巨大缺乏可行性,而自动主题标记的研究为解决该问题提供了方法和思路。首先对当前最为流行的狄利克雷分配主题模型进行阐述与分析,并根据主题标签三种不同表现形式,基于短语、摘要和图片,将主题标记方法分为三种类型;之后围绕提高主题的可解释性,以生成的不同类型主题标签为线索,对近年来的相关研究成果进行梳理、分析和总结,并对不同标签的适用情境和可用性进行探讨;同时根据不同方法的特点进一步分类,重点对基于词法、子模优化和图排序方法生成摘要主题标签进行定量和定性分析,从学习类型、使用技术和数据来源出发,对比不同方法的差异;最后对主题自动标记研究存在的问题和趋势发展进行讨论,基于深度学习、与情感分析结合并不断拓展主题标记应用的场景,将是未来发展的重点和方向。

关键词:主题模型;潜在狄利克雷分配(LDA);主题标记;主题标签
文献标志码:A　　**中图分类号:**TP391.1

Survey of Automatic Labeling Methods for Topic Models

HE Dongbin[1], TAO Sha[2], ZHU Yanhong[3], REN Yanzhao[4], CHU Yunxia[1+]
1. IoT Security and Sensor Test Engineering Research Center of Hebei Province, Shijiazhuang University, Shijiazhuang 050035, China
2. Key Laboratory of Agricultural Informatization Standardization, Ministry of Agriculture and Rural Affairs, China Agricultural University, Beijing 100083, China
3. Hebei Province IOT Intelligent Perception and Application Technology Innovation Center, Shijiazhuang Posts and Telecommunications Technical College, Shijiazhuang 050021, China
4. School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China

Abstract: Topic models are often used in modeling unstructured corpora and discrete data to extract the latent topic. As topics are generally expressed in the form of word lists, it is usually difficult for users to understand the meanings of topics, especially when users lack knowledge in the subject area. Although manually labeling topics

基金项目:河北省重点研发计划项目(22320301D);北京市科技计划课题(221100007122003);河北省农业科技成果转化项目(V1672275144902);石家庄学院博士科研启动基金项目(23BS018)。
This work was supported by the Key Research and Development Program of Hebei Province (22320301D), the Science and Technology Project of Beijing (221100007122003), the Agricultural Science and Technology Achievements Transformation Project of Hebei Province (V1672275144902), and the Doctoral Research Fund Project of Shijiazhuang University (23BS018).
收稿日期:2023-03-24　　**修回日期:**2023-07-17

can generate more explanatory and easily understandable topic labels, the cost is too high for the method to be feasible. Therefore, research on automatic labeling of topic discovered provides solutions to the problem. Firstly, the currently most popular technique, latent Dirichlet allocation (LDA), is elaborated and analyzed. According to the three different representations of topic labels, based on phrases, abstracts, and pictures, the topic labeling methods are classified into three types. Then, centered on improving the interpretability of topics, with different types of generated topic labels utilized, the relevant research in recent years is sorted out, analyzed, and summarized. The applicable scenarios and usability of different labels are also discussed. Meanwhile, methods are further categorized according to their different characteristics. The focus is placed on the quantitative and qualitative analysis of the abstract topic labels generated through lexical-based, submodular optimization, and graph-based methods. The differences between separate methods with respect to the learning types, technologies used, and data sources are then compared. Finally, the existing problems and trend of development of research on automatic topic labeling are discussed. Based on deep learning, integrating with sentiment analysis, and continuously expanding the applicable scenarios of topic labeling, will be the directions of future development.

Key words: topic model; latent Dirichlet allocation (LDA); topic labeling; topic label

主题模型(topic model)是一种从非结构化数据中自动提取隐含语义主题的生成概率模型,常用于大规模语料库和离散数据建模。该模型将语料库中的文档理解成特定隐含主题的分布,因而可以按照隐含的语义特征来发现抽象的主题,并通过词列表的形式表示。目前最为流行的主题模型是2003年由Blei等人[1]提出的潜在狄利克雷分配模型(latent Dirichlet allocation,LDA),在文本分类、异常检测、推荐系统、文本摘要、观点抽取、词义归纳、情感分析、信息检索等诸多领域[2-3]得到广泛应用,并快速发展。但由于主题采用词列表形式,如表1所示,通常会对用户正确理解造成一定的障碍。特别是在用户缺乏主题领域相关背景知识的情况下[4],其对主题的理解可能是破碎、片面和不准确的。

表1 APNews某主题的top-20主题词

Table 1 top-20 terms of a topic in APNews

Index	Topic terms
1 to 10	million company said bank billion offer will corp new business
11 to 20	inc plan sale share also firm finance manag stock invest

为提高主题模型发现结果的可解释性,通常的做法是进行主题标记[5-6]。具有特定领域知识的专家给出的主题标签通常更容易理解,对主题的说明也更加充分和准确[7]。但面对海量的语料数据,人工标记主题工作耗时费力,甚至成为不可能完成的任务。此外,局限于个人认识,标签的客观性也会受到影响。因此,利用机器进行自动标记,可提高主题标记的效率并增强准确性和客观性[8]。

本文与凌洪飞等人[9]的主题自动标记综述文献相比,不同于其以生成来源为线索,对现有主题标记方法进行分类比较,本文按照生成主题标签的不同形式分类,创新性地采用文本特征表示方法与主题标记模型所使用的技术相结合,并从这两个层面对现有研究成果进行总结,详细描述了不同方法的建模过程和适用场景,从全局和微观两个不同视角对现有方法进行阐述和分析。结合具体应用和相关领域的创新性研究,指出基于预训练语言模型以及多种深度学习技术相融合的方法应是未来突破的重点和方向。

1 主题模型介绍

对主题建模的研究,在早期通常利用空间向量模型[10-11]将相关文本聚合到同一类簇下。但该方法只对文本进行简单分类,未深入挖掘文本所蕴含的语义信息,也未对用户理解挖掘结果提供帮助。

为解决上述问题,Deerwester等人[12]提出了潜在语义索引或潜在语义分析方法,利用文本语义挖掘出更深层次的聚类信息(主题)[13]。该方法利用奇异值分解,通过将数据从高维空间映射到低维语义空间,以获得抽象的主题分布,并降低了整体开销。缺点是时间复杂度高,通过分解矩阵发现的主题解释性不强,不能区分一词多义的情况[14]。

针对该问题,Hofmann[15]提出一种概率潜在语义索引(概率潜在语义分析)模型,认为一篇文档由多个主题组成,且主题词服从于多项式分布。由于隐含了高斯分布假设,更符合文本特性。因其利用强

化期望最大化算法训练模型参数，所以解决了同义词和多义词问题。由于其并非完备概率模型，会逐渐增长并出现过拟合现象[14]。

2003年，Blei等人[1]在概率潜在语义索引基础上，提出一种由文档、主题和单词构成的三层贝叶斯概率模型，潜在狄利克雷分配（LDA）将文档看作词袋的集合，根据主题分布，以及词对主题的隶属度，生成集合中的文档。LDA不仅克服了模型随语料数量增长而逐渐增大的缺点，同时也避免了过拟合问题。如图1所示。

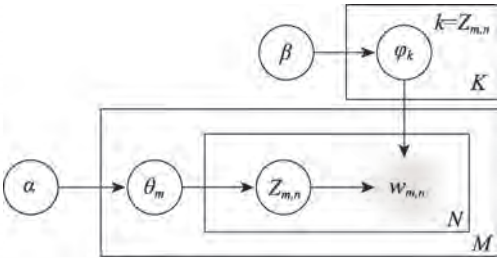

图1 LDA模型表示

Fig.1 Graphical representation of LDA model

LDA本质上是一种无监督学习算法，其生成一篇文档的过程如下：

1. 文档长度为服从泊松分布值 N

2. 从Dirichlet(α)分布中取出第 m 个文档的主题分布 θ_m

3. For $n=1$ to N：

4. 为当前第 m 个文档的第 n 个词从多项式分布 θ_m 中抽取出一个主题 $Z_{m,n}$

5. 从Dirichlet(β)分布中取出第 k 个主题的主题词多项式分布 φ_k，且 $k=Z_{m,n}$

6. 选择一个服从 φ_k 多项式分布的词 $w_{m,n}$，作为第 m 个文档的第 n 个词，将其写入该文档

α 和 β 是先验参数，采用近似估计算法（变分期望最大化或折叠吉布斯采样）来估计参数 θ_m 和 φ_k。前者推断速度快，但模型参数估计不如后者准确；后者易于实现，缺点是较前者收敛速度慢。

LDA模型出现后，因其拥有良好的先验概率假设和简单高效的抽样推理算法，逐渐成为主题建模事实上的标准化模型[9]，并广泛应用于文本分类、检索、摘要和主题演化等领域，开启了主题模型研究的热潮，相关研究成果也不断涌现[16-21]。

2 主题标记方法

对现有主题自动标记方法，单纯按照主题标签的表现形式分类，有三种类型：基于短语、摘要和图片的主题标签[22]。如图2所示，列举了生成这三类主题标签所对应的所有主题自动标记方法。

使用形式简洁的短语或概念[2-3,7-8,23-36]对主题进行标记，标签短小凝练，具有高度概括性，适合用户快速浏览主题内容。例如对基于APNews语料库[2,4,37]进行LDA主题建模，其中某一主题中的top-20主题词如表1所示，可采用上位词“economy”作为短语标签来解释主题。

虽然采用短语主题标签可帮助用户理解主题，但在主题含义丰富或涉及领域较为宽泛时，因长度较短，实际效果不能令人满意。若当前短语本身具有多重含义，在缺乏前后文的情况下，无法确定其真实含义。此外，如果当前主题中的重要主题词之间缺乏内在联系，则很难找到一个合适的短语去准确地描述主题。对用户来说，一个不恰当的主题标签只会使得正确理解主题变得更加困难[4,37]。

为克服短语主题标签的缺陷，通常需要信息丰富、描述充分的方式。因此，采用长文本来描述主题成为更佳的选择。长文本主题标签可单独使用，也可作为短语标签的补充[4]。目前，主要通过文本摘要技术来生成长文本主题标签，以提供更丰富和多样的信息，帮助用户充分理解主题的内涵[4,37-41]。

此外，还有研究者采用图片或文本配图形式的主题标签[22,42-44]在特定场景下对主题进行解释。由于图片标签表达较为生动、直观，且具有跨越语言鸿沟的优势，对帮助用户理解主题具有积极作用。三种不同主题标签的优缺点及适用场景总结如表2所示。

表2 三种类型的主题标签

Table 2 Three types of topic labels

主题标签	优点	缺点	适用场景
短语	简单、友好、易理解	难以充分解释内涵丰富的主题	信息检索、观点抽取、文本分类等
摘要	具有丰富的表现力，表达充分	不够简洁和概括，不利于用户快速浏览	科研热点检测、科研热点趋势分析与发现、多文本摘要
图片	直观，易于理解，跨越语言鸿沟	实现困难，某些抽象主题难以用图片描述	常用于推荐系统和情感分析系统内容展示

Sorodoc等人[43]认为不同主题应选择不同类型的主题标签，某些主题可能适用短语解释，有些主题可能适合长文本描述，另外一些主题可能更适合采用图片展示。综合来看，三种不同形式的主题标签各有特点，使用时需要考虑具体的应用场景。不论采

图2 主题标记方法

Fig.2 Topic labeling methods

用哪种形式，对主题标签的质量要求是没有区别的，生成的标签应符合如下标准[2,4]：(1)相关性，生成主题标签与主题在语义上高度相关；(2)覆盖性，生成主题标签包含更多主题词，则多样性越强，冗余度越低；(3)区分性，不同主题标签间的区别性越大，说明所表达语义的区分度越高，标签质量更高。

3 基于短语的主题标记方法

三种类型的主题标签差异明显，其适用的范围和场景也不同。本章将按主题标签的类型，对不同的主题标记方法分类阐述。

Aletras 等人[45]认为，短语主题标签对用户更友好，更容易理解和使用。在文献检索任务中，短语标签可以很好地概括主题主旨，短时间内帮助用户检索到更多的相关文献[45]。此外，当用户需要快速了解语料库内包含文档的种类、范畴等信息时，简短且概括性强的短语标签就成为不二之选。目前，主题自动标记方法大多采用短语主题标签，详情列表如表3所示。

3.1 基于统计方法

早期的主题自动标记研究通常将主题词的频率视为基本特性之一[34]，大部分研究利用文本的浅层特征，例如基于 BOW 和 N-gram[2-3,7-8,34,46]生成候选标签，再通过主题和候选标签间的相似度排序确定最优主题标签。

Mei 等人[2]最早提出一种使用短语或 N 元语法对

表3 基于短语的主题标记方法

Table 3 Topic labeling method based on phrases

作者	方法	文本表示	数据来源	描述
Mei等人[2]	统计	BOW、N-gram	内源	基于浅层特征，利用候选短语和主题间的KL散度值排序
Magatti等人[8]	统计	BOW	内源	基于一组多种相似性度量方法，利用谷歌目录（已停用）寻找最优主题标签
Lau等人[7]	统计	BOW	外源	基于维基百科语料分析，采用不同方法计算标签与主题间的相关评估分数
Mao等人[46]	统计	BOW、N-gram	内源	使用块分析和 N 元语法检验从文档中抽取候选标签，利用KLD进行排序
Kou等人[3]	统计	W2V、LTV	内源	使用稠密向量进行主题和候选标签向量化，解决了一词多义问题
Tiwari等人[34]	统计	BERT、GloVe	内源	使用词嵌入向量化短文本中的分层主题和候选标签，利用Cosine计算相似度
Shahriar等人[36]	统计	W2V、D2V	内源	基于数据驱动的挖掘框架，利用情感词和方面术语以增强分类性能
Hulpus等人[24]	图排序	BOW	外源	使用Wikipedia Categories、YAGO和DBpedia Ontology来获取候选标签
Aletras等人[23]	图排序	BOW	外源	利用Wikipedia产生候选标签，使用图排序方法对标签进行排序
Sanjaya等人[28]	图排序	BOW	外源	构建一个以主题、词、Wikipedia和候选术语为结点的异构图，用于集成学习
Kim等人[27]	本体	BOW	外源	基于Wikipedia建立UniDM本体，由决策树和K近邻方法建立主题映射
Zosa等人[30]	本体	SBERT[47]	外源	主题基于SBERT向量化，将本体映射转换为一种语言无关的多标签分类
Kinariwala等人[32]	本体	BOW	内源	top主题词输入CEPS_Ontology本体，获得最多计数的上位词作为主题标签
Allahyaria等人[35]	本体	BOW	外源	将本体概念和LDA集成在OntoLDA框架中，依赖于本体语义相似性排序
Bhatia等人[25]	神经网络	W2V、D2V	外源	引入doc2vec和word2vec向量编码，训练了一个有监督向量回归模型NETL
Kozono等人[29]	神经网络	W2V、D2V	外源	利用NETL[25]生成候选标签，并应用Mao等人[46]的主题自动标记方法选择标签
Popa等人[31]	神经网络	BART[48]	内源	基于BART微调模型建立候选标签池，参考主题词概率分布增大评分权重
Alokaili等人[33]	神经网络	embedding	内源	主题词编码为词嵌入到双向GRU中，输出到前馈神经网络以获取标签概率

注：对于文本表示，BOW为bag of words，W2V为word2vec，D2V为doc2vec，LTV为letter trigram vectors。

主题进行标记的方法，将主题标记过程视为一个优化问题，即单词分布间的KL（Kullback-Leibler）散度最小化，以及主题模型和主题标签间的互信息最大化。其主要利用短语的浅层特征，并根据当前短语和主题模型之间的KL散度对候选标签排序，以选取最优短语标签。Mao等人[46]采用与Mei等人[2]类似方法，使用块分析（chunking parsing）和 N 元语法检验（N-gram testing）[49]方法从文档中抽取候选标签。不同的是，其利用了层次主题模型的结构化信息，分别基于全局词频权重和Jensen-Shannon散度对每个主题的候选标签进行排序，以获得最佳的主题标签。

相对Mao等人[46]提出的对层次主题模型主题自动标记方法，Magatti等人[8]更早提出一种利用Google Directory（谷歌目录服务已于2011年7月21日停用）构造主题树对层次主题模型进行自动主题标记的方法ALOT（automatic labeling of topics）。该方法包含两部分：首先通过谷歌目录（当前已停用）构造的主题树获得主题层次结构，然后基于一组相似性度量（Cosine、Overlap、Mutual、Dice、Tanimoto和Jaccard），来寻找最优的主题标签。通常，层次主题模型将主题组织为层次结构，其中每个主题都是从更通用的主题中派生而来。这种层次结构可以帮助人们更好地理解文本的主题结构，因此也能在更高的概率上生成质量更佳的主题标签。

实践中，主题标记任务通常需要大量的标注数据来训练标记模型。然而，由于数据收集和标注的成本很高，很难在每个领域中都获得足够的标注数据。因此，迁移学习成为解决这个问题的一种有效方法。Lau等人[7]提出了一种使用无监督学习技术对候选标签进行排序的主题标记方法。对给定主题，首先选择top-N个主题词在英文Wikipedia（https://www.wikipedia.org/）中进行查询，并从得分最高文档的标题中获取候选标签。Lau认为一个良好的主题标签应该和主题词之间存在着某种较强的联系，因此使用了以下几种基于词法的关系评估措施：点间互信息（pointwise mutual information）、T检验（student's t-test）、Dice系数（Dice's coefficient）、皮尔森卡方检验（Pearson's χ^2 test）和似然比检验（likelihood ratio test）等。Lau使用了一个宽度为20的滑动窗口，在维基百科语料上进行分析，以获取候选标签和主题词词频统计信息，最后计算每个主题的top-10主题词与候选标签间的多个关系评估得分，并将同一个候选标签不同的评估分数进行算数平均，最终根据每个候选标签的平均分数获得最佳主题标签。

采用迁移学习方法，利用从外源性语料库中已获得的知识，不仅可以减少目标领域的标注数据量需求，也能获得更丰富和概括能力更强的候选主题标签，模型的泛化能力也得到提升。但该类方法也存在一定的局限性：首先，外源语料库与当前文本集应存在一定的共通性，否则难以实现主题标签的迁移；此外，外源性主题标签往往未出现在当前文本集中，对目标主题的覆盖度和准确性可能会存在偏差。

在主题标记任务中，相对于使用传统的BOW和N-gram，使用稠密向量表示文本有助于提高标记模型的性能，因为其可以更好地表示文本之间的相似性和差异性，所以使模型更准确地分类文本。Kou等人[3]使用OpenNLP[50]对给定主题的全部文档进行解析，抽取出包含top-10关键词的短语，作为候选标签集。为评估主题与候选主题标签之间的相关性，将该主题与候选标签映射到同一向量空间，并基于LTV（letter trigram vectors）、CBOW（continuous bag-of-words）和Skip-gram[51]三种不同词向量表示计算余弦相似度，以选择得分最高的标签。

word2vec[51]是一个包含CBOW和Skip-gram两种模型的框架，只能对单个单词进行建模，无法直接处理文档级别的语义关系。相比之下，doc2vec[52]可以将整个文档表示为一个向量，从而能够更好地处理文档级别的语义关系。基于此，为获取COVID-19大流行期间的热点事件，Shahriar等人[36]提出一个基于word2vec和doc2vec的主题框架SATLabel，用于从COVID-19相关的推文中提取主题并自动标记。该框架利用情感术语和方面术语的单字特征通过LDA输出主题聚类，从情感词和方面术语中各取20个组成不同向量集，利用软性余弦相似度找到与主题最为接近的主题标签。

相比word2vec和doc2vec只能处理局部上下文，GloVe[53]是一种使用全局统计信息生成词向量的方法，该方法不仅考虑了局部的上下文，还考虑了整个语料库的全局统计信息，因此可以更好地处理稀有词汇，但仍无法处理上下文信息。BERT（bidirectional encoder representations from transformers）[54]是一种预训练语言模型，可用于处理上下文动态相关性信息，在许多自然语言处理任务中都优于其他方法。因此，针对层次主题中不同级主题词和候选标签之间可能不存在共同术语，以及无法通过词匹配了解二者相关关系的问题，Tiwari等人[34]使用200维的GloVe和384维的BERT all-Mini-LM-L6-v2（https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2）词嵌入，向量化语料库、主题以及候选标签，通过动态上下文语义的引入，确保层级标记的主题一致性，可有效利用主题间的并列和从属关系来提高主题标记的有效性和准确性。

3.2 基于图排序方法

自动主题标记算法在生成候选标签后，通常会计算其与主题之间的距离（相关关系）并以此排序选择最优的主题标签。例如，Mei等人[2]利用KL散度，Mao等人[46]采用一组相似性度量方法，包括Cosine、Overlap、Mutual、Dice、Tanimoto和Jaccard距离，Lau等人[7]使用PMI（pairwise mutual information）、t-test、χ^2 test和LLR（log likelihood ratio）等方法。但上述方法仅限于直接计算候选标签和主题间的关系，并未利用到候选标签间的相互关系和信息。而图排序算法可以利用结点间的关系，通过随机游走过程发现网络中的重要结点或路径，相较于前述方法仅单纯依赖向量空间中的距离（关系）计算方法，可以有效利用不同候选标签间的相关关系选出更具代表性和概括性的标签。

Hulpus等人[55]提出一种基于数据结构化信息的主题自动标记方法，分为四个阶段：（1）通过LDA模型发现主题；（2）利用DBpedia（http://wiki.dbpedia.org/）的结构化数据，将top-N主题词与其中的具体概念联系起来，进行词义消歧；（3）所获概念作为结点，不同结点的分类概念在DBpedia中的从属关系表示为边，构建候选标签图；（4）利用图排序的随机游走算法，迭代得到全部结点的聚焦信息中心度（focused information centrality）值，最后选择得分最高的结点（对应DBpedia中的概念）成为主题标签。

相比Hulpus等人的方法，Aletras等人[23]提出一种基于图排序的无监督主题标记方法。除了利用外源性的知识库，还引入了谷歌搜索引擎，其覆盖了大量的互联网信息，可以提供与特定主题相关的多样化和广泛的搜索结果。这些搜索结果可以作为构建主题标记的基础，从中提取关键词、短语和主题相关的内容。首先，利用Wikipedia产生候选标签[7]。然后，对候选标签排序，包括三个步骤：（1）基于Bing搜索引擎，使用top-N主题词进行检索。（2）利用OpenNLP（http://opennlp.apache.org/）对返回结果中的标题句子进行形式化标记，并将标记词和搜索结果中的元数据作为结点构建无向图，基于维基百科作为参考语料库来计算词的共现率，并使用归一化逐点互信息

(normalized pointwise mutual information，NPMI)[56]给图中的两个相邻结点的边赋值。为避免偶然出现的词共现所导致的噪声，设定只有当NPMI>0.2时才认为两个结点间存在链接。(3)通过PageRank算法[57]进行排序，每个结点(候选标签)的得分按照其中所包含的所有关键词的权重求和，选择分数最高的作为主题标签。

对于生成候选标签子集后再利用图排序算法获取最优解的办法，在排序过程中未再考虑候选标签和主题间的关系，可能会导致主题标签的重心发生偏移。针对该问题，Sanjaya等人[28]利用Lau等人[7]的方法生成候选标签子集后，构建了一个包含主题、词、维基百科文章和候选标签的异构图，引入了更多维度的相关性特征，对排序结果的改善具有积极的意义。但该方法并未考虑不同类型的结点之间是否天然具有平等的关系，以及对投票结果会产生怎样的影响。特别的，Sanjaya认为如果能够获取语料库中主题领域与候选标签间关系的先验知识，可能对其排序方法的最终排序结果有积极的影响。

3.3 基于本体方法

短语主题标签通常将一或多个单词组合成一个短语来描述主题。其存在以下问题：(1)多义性问题。同一个短语在不同的上下文中可能具有不同的含义。例如“apple pie”可以是一种美食，或是一个品牌。(2)歧义性问题。同一个短语可能被用于描述不同的主题。例如，“social media”可能被用于描述互联网媒体领域或用于描述社交网络。(3)连贯性问题。一些短语可能不具备连贯性，难以形成一个完整的主题。例如，“in the news”可用于描述不同的主题，但之间并不存在明显关联。(4)预定义问题。使用短语主题标签需事先定义，因此无法处理一些新出现的词汇或短语。

一些研究者[27,30,32,35]尝试使用本体(ontology)方法来解决上述问题。先验知识是本体方法非常重要的一部分，为本体的构建和推理提供了基础。本体方法的核心目标之一就是将先验知识形式化地表示为概念、关系和约束的集合，并利用这些先验知识进行推理和语义处理。

本体可以通过手动构建或自动构建(如从现有文本中抽取概念和关系)得到。然后，对于给定的文本，本体方法可以将其表示为一个向量，该向量反映了文本与本体中各个概念之间的关系。例如，可以使用基于本体的词嵌入技术(如word2vec)来生成文本向量。最后，可使用机器学习方法(如逻辑回归、朴素贝叶斯、支持向量机等)来构建分类器，输入文本向量后给出合适的主题标签。

使用本体方法进行主题标记，通常利用本体中的语义信息对文本数据进行理解和分析，或通过推理机制发现文本中隐含的语义关系和概念；其目标是将本体中的先验知识与文本特征结合，获取更准确的主题标记结果。Allahyaria等人[35]将本体概念和主题模型集成在一个框架OntoLDA(如图3所示)中，每个主题表示为概念上的多项分布，每个概念是单词上的多项分布。通过本体概念和主题，以及本体概念之间的关系就可以确定文档的主题。与已有研究类似，整个过程分为两个阶段：(1)抽取并筛选出与主题密切相关的候选主题标签。先确定本体概念集 $C=\{c_1,c_2,\cdots,c_i,\cdots,c_C\}$，然后对当前第 j 个主题 φ_j 和第 i 个本体概念 c_i，根据OntoLDA主题模型的边缘概率公式 $p(c_i|\varphi_j)$ 选取边缘概率最高的 K 个本体概念，构建主题语义图。(2)针对每个主题，提取其主题图作为子图，根据语义相似度进行图排序，以获取最适合的主题标签。

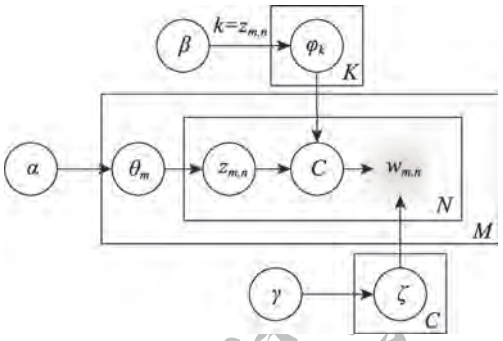

图3 OntoLDA模型表示

Fig.3 Graphical representation of OntoLDA model

Allahyaria等人提出的方法将本体概念集成到LDA中，提高了主题模型的内聚性，但该方法较为复杂，适用性不高。与其他主题模型一样，OntoLDA的性能受到主题数的影响。如果主题数设置得不当，可能会导致一些主题被合并或分裂成不合理的子主题，降低了本体概念和主题的匹配度，从而削弱了主题标签的准确性。

为提高本体概念和主题的匹配度，Kim等人[27]提出一种基于社会网络分析(social network analysis，SNA)和本体的方法来标记科研文献中有影响力的主题。该方法利用SNA方法选择热点主题，为增强主题的可解释性，构建了一个建立在Wikipedia上的挖掘结

果集UniDM本体,考虑到主题间的相互关系,可利用多种方法在UniDM和主题间进行映射,例如通过决策树和K近邻等方法建立起对主题的映射关系,有效提高了主题和候选标签的匹配程度,对最优的主题标签的选择具有积极作用。

为进一步提升对主题候选标签和主题关系的理解,提取文本的层次特征,将深度学习技术与本体方法结合,Zosa等人[30]针对多语言主题自动标记,提出一种本体映射方法,将主题映射到与语言无关的新闻本体中的概念。Zosa将本体映射问题视为一个多标签分类任务,利用一个基于Transformer的微调预训练语言模型SBERT(sentence-BERT)[47],将主题表示为上下文相关的嵌入形式。其中,一个主题可以被分类为属于本体中的一个或多个概念。需要注意的是,新闻本体分类概念与具体语言无关,其目的是为了无需额外训练就可以在多种语言上生成主题标签。

本体方法的优点是能够利用丰富的领域知识,提高主题标记的准确性和一致性,但建立和维护一个本体结构需要耗费大量的人力和时间。为降低构建本体的成本,Kinariwala等人[32]使用开源软件tool-Protégé生成了一个涉及“体育”“犯罪”“政治”和“环境”四个领域的本体CEPS-Ontology,并将主题中的top词汇作为输入,获得最多归属计数的上位词被选为该主题的标签。该方法局限于上述四个领域的新闻语料,并且需要事先构建本体作为主题标签池,主题标记质量与特定本体相关,其适用性受到限制,只能应用于特定领域。从上可知,本体方法对新领域或新概念的适应性较差,需要手动或自动扩展本体结构以涵盖新的概念和关系。

3.4 基于神经网络

主题标记是一项极具挑战性的NLP任务,目前仍面临诸多困难,包括:(1)多义性和歧义性。由于主题是由主题词集构成,同一个词或短语可能在不同的上下文中具有不同的含义,可能导致主题标记失效。(2)数据稀疏性。对于某些主题,训练数据中可能只包含很少的示例,使得准确地标记这些主题变得困难。(3)领域特定性。主题标记的性能可能会受到领域特定的词汇和表达方式的影响,需要具备跨领域的泛化能力。(4)多语言问题。在多语言环境下,同一个主题可能会以不同的词汇和表达方式出现,需要解决跨语言的主题标记问题。

针对上述问题,不同研究者将神经网络技术应用于自动主题标记任务。Bhatia等人[25]提出了一个基于word2vec[51]和doc2vec[52]的主题自动标记模型NETL(neural embedding topic labelling)。标记过程分为两个阶段:首先,生成候选主题标签集合;然后利用一个有监督学习的排序模型对候选标签排序。关键步骤详述如下:

第一阶段,参照Lau等人[7]的方法产生候选标签。Bhatia利用Wikipedia[7]语料训练doc2vec模型,并使用文档嵌入表示文档的标题 a。若给定主题为 T,则 a 与 T 的相关性定义为 $rel_{d2v}(a,T)$,若基于word2vec,则相关性定义为 $rel_{w2v}(a,T)$,且最终相关性定义为 $rel_{d2v+w2v}(a,T)$。上述公式定义如下所示:

$$rel_{d2v}(a,T)=\frac{1}{|T|}\sum_{v\in T}\cos(E^{d}_{d2v}(a),E^{w}_{d2v}(v)) \tag{1}$$

$$rel_{w2v}(a,T)=\frac{1}{|T|}\sum_{v\in T}\cos(E^{w}_{w2v}(a),E^{w}_{w2v}(v)) \tag{2}$$

$$rel_{d2v+w2v}(a,T)=rel_{d2v}(a,T)+rel_{w2v}(a,T) \tag{3}$$

其中,$E^{d}_{d2v}(a)$ 为 a 的文档嵌入表示,$E^{w}_{d2v}(v)$ 为主题词 v 的文档嵌入表示,$E^{w}_{w2v}(a)$ 为 a 的词嵌入表示,$E^{w}_{w2v}(v)$ 为主题词 v 的词嵌入表示,$|T|$ 为主题词的个数,cos为余弦相似度。

第二阶段,利用CrowdFlower(https://www.crowdflower.com/)获得人工标注[7]以及候选标签的四个特征数据,训练基于多特征的回归模型NETL,对候选主题标签进行重排序。四种特征数据包括:(1)候选标签和主题词间的字母三元组(letter trigram)重叠统计[3];(2)令 a 为结点,Wikipedia中的超链接为边,构建有向图,利用PageRank方法[57]获得每个结点的权重;(3)词的个数;(4)候选标签与top-10主题词的重叠个数[7]。

Bhatia等人的研究结果表明,利用神经网络获得词嵌入和句嵌入,可以学习到单词和文本的语义表示,从而更好地捕捉主题之间的语义关联和差异。此外,神经网络通过上下文窗口或序列模型来捕捉词语之间的关联,更好地理解主题在上下文中的含义和语义,最终提高主题标记任务的准确性。

对层次主题,如果使用NETL[25]直接进行主题标记,且只有主题词作为输入,则生成的主题标签与子主题缺乏联系,而且可能出现重复。针对该问题,Kozono等人[29]提出一种NETL的改进模型,将子主题获取的主题标签和相关的文档作为输入,根据Mao等人[46]的主题自动标记方法,利用主题间的兄弟及父子关系,基于doc2vec和word2vec获得不同向量编码,并计算候选标签成绩,选择排序后的top-10主题

标签。同理，对上一级主题依次迭代求取每个父主题的主题标签。该方法属于两阶段标记方法，首先生成候选标签集，然后进行排序。问题在于，从现有的内源性语料库或外源性的知识库中可能找不到合适的概括性的短语标签，此时主题标记的结果与实际相差可能会非常大。

为了获得与主题相关性更高、覆盖性更强的主题标签，Alokaili等人[33]提出一种基于seq2seq模型的主题标记方法，可生成当前语料库或知识库中不存在的短语标签。该模型的编码器和解码器均采用循环神经网络(recurrent neural network，RNN)，将主题词编码为300维的词嵌入输入到双向GRU(gated recurrent unit)中，解码器生成一系列词嵌入，作为前馈神经网络的输入，最终选择概率值最高的词作为主题标签。由于产生候选主题标签使用生成式神经网络，实时性可能会差一些，但生成的主题标签相关性和概括性可能会更好。

相比GRU、LSTM(long short-term memory)等传统RNN，Transformer在并行计算、长期依赖建模、全局信息获取、编码器-解码器结构、模块化和可扩展性等方面具有明显的优势，因此在自然语言处理任务中取得了显著的性能提升。为进一步提升自动主题标记模型的效果，Popa等人[31]提出了一个基于BART(bidirectional and auto-regressive transformers)[48]的NETL的改进模型，该模型也采用了经典的两阶段主题标签生成方法。第一阶段，构建了NETL标记器，生成用于微调BART的数据集(标签候选子集)。为避免过拟合，NETL labeler采用了top-5主题词，通过嵌入相似度，并参考主题词概率分布增大在评分中的权重。除此之外，还利用N-gram从语料库中抽取名词性短语作为候选标签。第二阶段，利用seq2seq模型构建了一个从主题到候选标签的一个一对多序列映射，其中主题表示为由空格分隔的前20个主题词的串联字符串。预训练模型BART在生成的数据集上进行微调训练后，最终的预测模型BART-TL可以为任意单个主题输出主题标签。

Popa等人构建的主题标记模型基于BART，一种大规模预训练语言模型，由FAIR团队于2019年推出。BART基于Transformer架构，并使用海量的无标注数据训练。大规模的预训练模型可以学习主题和候选标签的文本表示，提取深层特征，优化标记过程，增强泛化性，在不同的领域中都能获得较好的标记结果。还能通过共享嵌入空间或联合学习多语言表示来解决跨语言的主题标记问题。

3.5 讨论

在表3中，根据数据来源，主题标记所依赖的语料库可分为内源性和外源性两种。前者仅限于语料库自身，后者需要依靠外部扩展知识，或借助外部数据以更广泛(或具象)的表现形式(例如图片)来描述主题以及主题间的关系，以提高生成主题标签的准确性和多样性[7,22-23,25,42,58]。

内源性主题标记方法[4, 38]基于自身语料库来抽取或生成主题标签，在语义相关性上更接近原始语料库。但该方法对语料库要求较高，只有规模足够大，语料足够丰富时，生成的主题标签才能够反映主题的本质。而那些规模较小、文本较短、表达欠规范语料库，很难抽取出高质量的候选标签。但信息足够丰富的大规模语料库，进行文本解析和抽取候选标签所消耗的资源也十分可观。

外源性主题标记方法[7,23-25]通常会利用外部知识库中已有的先验知识来抽取和选择主题标签。外源性语料库的内容更全面，范围更广阔，提供的方法或服务更新颖，得到高质量主题标签的可能性也更高。但也存在一定缺陷，例如：(1)外源性内容或服务通常来自互联网，产生和消亡的速度都很快，例如Google Directory分类目录服务已在2011年停止；(2)主题可能不存在于外部源中；(3)其他一些不可控因素，对依赖外源语料或服务的主题自动标记模型的稳定性，可能会造成一定的影响。

当前主题标记的研究重点在于候选标签的生成，以及标签排序算法的选择上。通常基于外源性方法生成的主题标签概括性更好，但实现复杂度也更高。如果主题在外部语料库中不存在，则标记工作会比较困难，此时结合内源性语料生成候选标签可能会是更好的选择。此外，基于稠密向量建模方法的局限性在于主题标签生成的质量依赖于词向量的质量，其质量又受到语料库的影响[3]。因此，利用外源性语料库中所蕴涵的更为丰富的语义表示和先验知识，引入预训练语言模型，并基于其建构主题自动标记模型可能是一个更好的选择。

4 基于摘要的主题标记方法

对于内涵较为丰富的主题，短语标签的表达能力受限于其长度，通常无法对主题给予全面和充分的描述。面对短语主题标签解释能力不足的问题，基于摘要的方法对主题进行标记逐渐受到研究者的重视，该类研究多采用抽取式摘要方法对主题进行标记，表4概述如下。

表4 基于摘要的主题标记方法

Table 4 Topic labeling method based on summaries

作者	方法	类型	描述	特点
Basave等人[38]	词法特征	无监督	采用抽取式摘要方法,将所有句子作为候选句,通过平均主题分布概率、平均Tf-idf、最大边界相关和图排序四种方法度量句子与主题间的相关性,并对句子评分	利用成熟摘要技术,不依赖外源语料
Barawi等人[39]	词法特征	无监督	针对情感主题,选择与主题一致且情感耦合的句子构成主题标签,选择句子时需要考虑两个标准间的平衡问题	可应用于其他情感主题模型的改进
Wan等人[4]	子模优化	无监督	基于子模优化(Submodular)的两阶段主题标记建模方法,使用三个不同的单调子模函数表达句子的中心度,衡量句子与主题的相关性、覆盖性和可区分性	采用贪心算法,时间复杂度高
He等人[37]	图排序	无监督	主题标记过程分为三个阶段:候选句子识别,基于候选句整体中心性排序(TLRank-C)和基于提高与抑制结点投票率策略的图排序算法(TLRank-G),以实现冗余控制	全局最优解,时间复杂度低
Kozbagarov等人[41]	EM方法	无监督	采用BERT对语料库的向量化,基于MSSC和k-means获得指定主题数量的聚类,应用EM算法获得模型参数,选择一个嵌入最接近给定聚类质心的代表性句子作为主题标签	采用词嵌入聚类方法,时间复杂度高
He等人[40]	注意力	有监督	基于注意力机制的三层神经网络模型,利用成对注意力组件,从相反方向对句间关系建模,有效抑制冗余,为分类器提供高质量关系编码,生成高质量主题标签	利用注意力机制对句间关系编码,方法复杂

4.1 基于词法特征

Basave等人[38]认为外源性的主题自动标记方法并非总是适用的,这是因为主题词有时并不存在于外部源中。因此提出一种内源性多文本摘要算法框架[38],利用四种不同方法评估所有候选句与主题的相关性:(1)SB(sum basic),对给定主题,利用句子所包含主题词的边缘分布概率均值评分;(2)混合词频逆文档频率(Hybrid Tf-idf),选取对主题具有较高隶属度文档中的句子,采用Tf-idf均值进行评分;(3)最大边界相关(maximal marginal relevance,MMR)[59],在计算句子与主题的相关性时,避免与已有句子产生叠加冗余,以均衡评分;(4)根据句间相似度,利用TextRank[60]对句子评分。

根据实验结果发现,基于词频的方法优于SB、TextRank和MMR。通常新闻事件很难在外源语料中找到相关内容,因此只能依赖内源性语料生成主题标签。Basave等人[38]提出的方法基于词法特征,关注单个词汇的统计信息,无法利用词间的语义关系和上下文信息,难以充分理解主题和候选标签的真实含义,导致生成的摘要缺乏准确性和表达力。

使用多维特征可以从不同角度对文本进行建模,包括语义、句法、结构、情感等。通过综合考虑不同特征之间的关系和权衡,可以更好地理解文本的含义、结构和上下文关系,从而生成更优质的摘要主题标签。

Barawi等人[39]认为,对情感主题建模,如果只基于词法特征的相关性,将导致模型趋于选择信息量有限的短句,难以捕捉有效的情感信息,无法生成适合的情感类主题标签。Barawi提出一种对情感主题进行自动标记的模型,建模过程中引入情感维度特征的处理,选择与主题一致且情感耦合的句子构成候选主题标签集;排序算法主要考虑了句子与主题的相关性 $Rel(s|t_{l,z})$ 和句子对情感的覆盖度 $Cov(s|t_{l,z})$ 两方面内容,候选标签成绩的计算公式定义如下:

$$L(s|t_{l,z})=\alpha\cdot Rel(s|t_{l,z})+(1-\alpha)\cdot Cov(s|t_{l,z}) \tag{4}$$

其中,s 为候选句,l 为情感标签,z 为给定主题。实验证明,该方法对情感主题进行自动标记优于其他基线方法,属于一种通用方法,无任何特定的依赖关系,可直接应用于任何情感多项式分布主题模型的改进。

4.2 基于子模优化方法

摘要生成方式通常有生成式(Abstractive)和抽取式(Extractive)两种。前者对生成文本的文法和语法要求严格,实现较为困难;后者从原文档中抽取句子组成摘要文本,实现简单且无生成文本的文法和语法问题。因此现有研究大多采用抽取式方法生成摘要主题标签[4,37,61],通常分为两个阶段:首先对语料库中的句子进行评分,然后选择合适的句子生成摘要(主题标签)[4,62]。通常来说,抽取式方法存在一个算法的下界,由于在句子评分过程中未考虑生成摘

要时所产生的冗余，会导致句子排序的准确性被削弱[63]。因此，如何抑制由于句子重叠所导致的冗余成为抽取式摘要方法研究的难点和重点。

子模性在组合优化中具有重要作用，当目标函数具有子模性时，组合优化问题通常能够在多项式时间内得到最优或近似解[64]。Lin等人[65]首次将子模函数应用于多文档自动摘要，并将其定义为预算约束下的子模函数最大化问题。由于利用MMR构建的目标函数仍然是子模且非单调，采用一种新的贪心算法来优化目标函数[66]以保持单调不减，最终解决摘要中出现冗余的问题。根据已有研究，Wan等人[4]利用预算约束下最大化具有子模性的评分函数，提出一种子模优化（submodular）的两阶段主题自动标记方法：

第一阶段，滤除大部分与主题相关性低的句子，句子 s 与主题 θ 间的KL散度[4]计算如下所示：

$$KL(\theta,s)=\sum_{w\in SW\cup TW} p_\theta(w)\times\log\frac{p_\theta(w)}{tf(w,s)/len(s)} \tag{5}$$

根据 $KL(\theta,s)$ 对句子集合排序，分别选取与每个主题最相关的top-500句子作为候选句集合 V 。

第二阶段，对每个主题 θ ，利用子模最大化方法从 V 中寻找构成主题标签的真子集 E ，该方法通常是一个NP-hard的问题，采用贪心算法[65]获得近似最优解[4]，公式如下：

$$\tilde{E}=\operatorname{argmax}_{E\subseteq V}\{f(E)\}\quad \text{s.t.}\quad len(E)\leqslant L \tag{6}$$

$\tilde{E}$ 为主题摘要句子集合，$len(E)$ 为摘要集合 E 的长度，L 为摘要的最大长度。 $f(E)$ 为 E 的综合质量评分函数，由 $REL(E)$、$COV(E)$ 和 $DIS(E)$ 组成，分别对应 E 对给定主题 θ 的相关性、覆盖性和不同 $\tilde{E}$ 之间的可区分性[4]，公式如下：

$$f(E)=REL(E)+COV(E)+DIS(E) \tag{7}$$

实验结果表明，该方法有效抑制标记过程中的冗余产生，生成的主题标签在相关性、覆盖性和可区分性三个维度上获得较大提升。该子模函数优化模型基于贪心算法实现，虽然对设计NP-hard问题的有效逼近算法有效，但对贪心算法进行优化较为困难，通常计算代价较大，且得到的是近似最优解。

4.3 基于图排序方法

图排序是生成抽取式摘要的另一种较为重要的方法，其特点是算法收敛速度快、易于得到全局最优解[67]。其中PageRank[68]是最具代表性的图排序算法，基于图定义随机游走过程（一阶马尔可夫链），根据转移矩阵随机访问各个结点进行投票，收敛到平稳状态后，结点的最终得票率为其评分。

LexRank[69]和TextRank[60]为PageRank的改进方法，可直接用于生成主题标签，虽然句子评分可获得全局最优解，但由于图排序过程中未考虑句间冗余控制的问题，导致生成的主题标签效果不佳。

针对上述问题，He等人[37]对次模函数优化模型[4]和图排序模型[60,69]进行研究，提出一种冗余感知的、基于图排序的三阶段主题自动标记模型TLRank[37]，标记过程如图4所示。

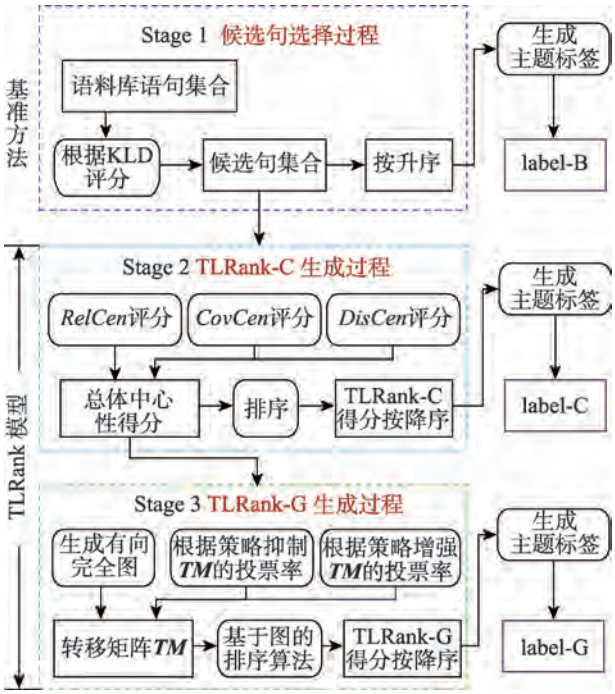

图4 TLRank主题标记过程

Fig.4 TLRank topic labeling process

第一阶段，抽取候选句[4]，为每个主题 θ 生成相应的候选句集合 $CSSet$ 。

第二阶段，借鉴子模函数优化模型中的奖励函数，从相关性、覆盖性和区分性三方面获得句子的综合中心性评分，公式分别定义如下：

$$RelCen(s,\theta)=\frac{\exp(KLD(TPS(\theta),s)^{-1})}{len(s)^a} \tag{8}$$

$$CovCen(s,\theta)=\frac{\sum_{w\in SW} p_\theta(w)\times tf(w,s)}{len(s)^a} \tag{9}$$

$$DisCen(s,\theta)=\frac{\sum_{w\in SW} p_\theta(w)\times tf(w,s)}{\sum_{\theta^*\in U}\sum_{w\in SW} p_{\theta^*}(w)\times tf(w,s)} \tag{10}$$

为使用一个统一尺度来衡量句子的整体质量，定义整体中心性（$OverAllCen$），公式如下：

$$OverAllCen_y = \alpha RelCen(s_y,\theta)+\beta DisCen(s_y,\theta)+(1-\alpha-\beta)CovCen(s_y,\theta) \quad (11)$$

式中，$OverAllCen_y$ 为候选句 y 的整体中心性，s_y 代表候选句 y，并有 $\alpha>0$，$\beta>0$，$\alpha+\beta<1$。

第三阶段，根据句子的综合中心性评分和句间相似度，提出一种抑制和扩张策略：通过构建一个正定的转移矩阵实现马尔可夫过程，使模型在图排序过程中能够感知冗余并改变投票比率，从而生成冗余度更低、更具多样性的主题标签。过程概述如下：

以 *CSSet* 中句子为结点构建有向完全图，图中结点 y 计算公式[57,60,68-69]如下：

$$p(y)=\sum_{x\in In(y)}\frac{edge_{xy}}{\sum_{z\in Out(x)}edge_{xz}}p(x) \quad (12)$$

式中，$edge_{xy}$ 为 x 指向 y 的边。当 x 的整体中心性值大于 y 时，$edge_{xy}$ 受到抑制，公式如下：

$$edge_{xy}=\frac{edge_{xy}}{e^{similarity\_Jaccard(s_x,s_y)}} \quad (13)$$

反之，扩张 $edge_{xy}$ 的值公式如下：

$$edge_{\cdot y}=edge_{\cdot y}\,r^{\left(Degree_y/\sum_{x\in CSSets}Degree_x\right)^a} \quad (14)$$

式中，$Degree_x$ 和 $Degree_y$ 为结点 x 和结点 y 的度值[69]，$edge_{\cdot y}$ 为任何指向结点 y 的边。扩张和抑制策略的目标是改变结点的投票比率，加权重要结点并抑制非重要结点。

实验结果表明，TLRank 与对照算法相比更优。但图排序模型属于无监督学习，很难获取候选句的深层特征，也无法捕捉词语、句子和文本之间的复杂关系，不能准确地理解文本的语义信息，进行更复杂和层次更深的句间关系建模，难以进一步提高生成主题标签的多样性和有效抑制标签中的冗余。

4.4 基于神经网络方法

神经网络能够学习丰富的语义表示，通过隐层的非线性变换和特征提取，可以捕捉词语、句子和文本之间的复杂关系。这使得神经网络能够更好地理解文本的语义信息，从而提高主题标记的准确性和语义一致性。

Kozbagarov 等人[41]认为，句子与单词不同，具有完整的语法和语义结构且信息量更大，更适合用来解释主题，因此未采用通用标准方法中的词法特征，而是使用了预训练语言模型 BERT 对主题和语料库文本进行向量化，并在此基础上进行自动主题标记，整个过程分为六个步骤：(1)首先将语料库中的句子依次输入 BERT 模型，获得的句嵌入为输出顶端 4 个隐层的均值；(2)如果数据规模较大，为了提高计算效率，可从语料库中随机抽取一个子集；(3)根据主题数 K 值，基于句子的嵌入表示，利用最小平方和聚类(min-sum-of-square clustering，MSSC)和 k-means 等方法进行聚类；(4)根据前述步骤获得的所有句嵌入和聚类质心，重新计算句子的概率分布值；(5)在已获聚类的基础上，计算句子的聚类分布矩阵 $F_{n\times m}$；(6)最后应用 EM(expectation maximization)算法完成指定次数的迭代计算，获得模型参数的估计量，得到句子的主题概率分布和主题在文本中的概率分布，并选择一个嵌入最接近给定聚类质心的代表性句子作为主题标签。

相比传统的词法特征方法，神经网络能够更准确地获取主题和句子的含义和上下文，以提高主题标记的精度。但主题标签使用单个句子也有不足，因为单个句子受限于长度和结构，不能充分揭示主题的意义和背景。一个主题往往涉及多个相关的概念、事实或论点，需要更广泛的语境才能被准确地理解和描述。因此，利用神经网络生成的长文本主题标签，采用多个句子的摘要形式更为常见。

在生成抽取式摘要时，无监督学习很难获取候选句的深层特征，进行更复杂和层次更深的句间关系建模。因此，He 等人[40]提出一种基于注意力机制的三层神经网络主题标记模型 TLPA(topic labeling model with a paired-attention)，其结构如图 5 所示。

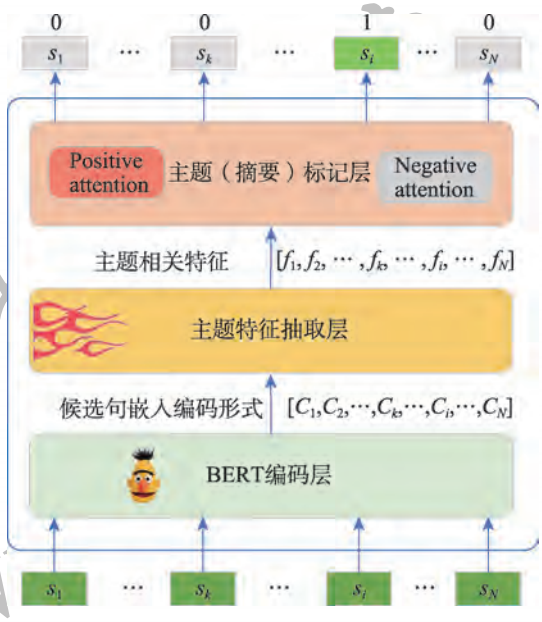

图 5 TLPA 三层神经网络结构示意图

Fig.5 TLPA three-layer neural network structure

TLPA模型底层“BERT编码层”运行于句子级别，将 N 个候选句 $[s_1, s_2, \cdots, s_k, \cdots, s_i, \cdots, s_N]$ 动态编码为前后文连续特征向量 $[C_1, C_2, \cdots, C_k, \cdots, C_i, \cdots, C_N]$；相对于BertSum[70]采用两层神经网络结构，为增进对主题的理解，增加了一个“主题特征提取层”，运行于句子级别，从 $[C_1, C_2, \cdots, C_k, \cdots, C_i, \cdots, C_N]$ 中抽取出深层特征 $[f_1, f_2, \cdots, f_k, \cdots, f_i, \cdots, f_N]$；主题标记层位于模型顶层，运行于文档级别，采用成对注意力对句间关系编码，模仿人类决策过程。最终通过线性分类器选取适合的句子，生成主题标签。

以Transformer的正向（Positive）注意力编码为例，公式定义如下：

$$PosAtt_{\text{Tran}} = [TranEncoder(X), (1/KLD)(X)] \tag{15}$$

其中，$PosAtt_{\text{Tran}}$ 为基于Transformer的正向注意力实现。将候选句集视为输入文档 D，则标记任务转化为二分类优化问题，预测句子 s_i 是否属于主题标签的概率值，公式如下所示：

$$\hat{Y}_i = \sigma(Weight(PosAtt(X_i)) - Weight(NegAtt(X_i))) \tag{16}$$

其中，$\hat{Y}_i$ 为文档 D 中的句子 s_i 属于主题标签的概率预测值，X_i 为BERT编码层输出的句嵌入或主题相关特征，例如 $PosAtt(X_i)$ 为式（16）得到的正向注意力特征，$NegAtt(X_i)$ 为负向注意力特征。

为准确而全面地评估TLPA主题标记模型的有效性，选用公开数据集SIGMOD和AP[4]。实验表明，TLPA生成的主题标签在与主题的相关性、覆盖性和区分性上显著优于图排序及其他对比方法。虽然采用基于成对注意力的句间关系编码器来模仿人类决策过程，为分类器提供高质量的句间关系编码，有效提升了模型的冗余控制水平和求解精度，但当前尚不能通过单词级和句子级的语义表示来发现潜在语义主题模式，并基于神经网络实现对文本深层特征的提取，以克服传统主题模型的局限性。

4.5 讨论

文本摘要方法通常分为单文档和多文档两种不同类型。由于发现主题结果来源于多个文档，借鉴多文档摘要方式来对主题建模结果进行标记。此外，文本摘要根据生成文本的过程又分为生成式和抽取式两种。前者生成全新的摘要文本，多样性高、冗余度小，但由于生成文本对文法和语法要求较为严格，实现困难；后者使用原文档中抽取出的句子组成摘要主题标签，实现简单，也无需担心生成文法和语法问题。但是该方法缺点也很明显，抽取句子组成的摘要通常存在句子重叠所导致的冗余。因此，结合二者的优点，使用抽取式方法生成候选句集合，并在此基础上采用生成式摘要方法生成主题标签可能是一种更加可行的策略。

当前使用文本摘要对主题进行标记的研究主要基于抽取式方法，标记任务通常分为三个过程：(1)从语料库中抽取出句子；(2)提取句子特征并排序；(3)根据排序结果选择合适的句子，以最小冗余代价生成主题标签。也有研究为优化整体效果，将第二和第三个过程合为一个整体[4,37]。

对于生成短语和摘要主题标签，一些研究采用了图排序方法。这是因为其优势在于结合图的全局信息来计算结点权重，充分利用句间关系，以抑制生成主题摘要过程中出现的冗余。例如TLRank[37]就是通过控制转移矩阵对图中结点的投票率进行增强或抑制，从而达到整体排序结果最优。但该方法中一些超参数的设置源于经验，针对不同语料库还需人工调整。因此，利用图排序进行主题标记的下一阶段的研究重点应该放在增强模型的泛化性上，对图排序过程中冗余控制的作用机理进行研究，实现模型参数的自动学习。

为进一步抑制主题标签中的冗余，解决一词多义和手工调参的问题，一种基于注意力机制的神经网络模型TLPA[40]被用于主题标记。该模型采用动态词嵌入解决了一词多义问题；利用Transformer编码器提取深层特征，有效减少了噪声信息的干扰，并增强了模型对主题的理解；采用基于成对注意力的句间关系编码器来模仿人类决策过程，为分类器提供高质量的句间关系编码，有效提升了模型的冗余控制水平和求解精度。

5 基于图片的主题标记方法

图片主题标签在特定场景下，具有更加直观的阐释能力，但不足以解释含义复杂的主题。Sorodoc等人[43]认为不同类型的主题选用不同的主题标记方式可能是更好的选择。目前采用的方法大多是基于对图片相关文本信息的解读，通过对图片进行评分的方式来选择最相关的图片主题标签，方法总结如表5所示。

与基于文本的主题标记研究不同，Aletras等人首次提出一种利用图像进行主题标记的三阶段方法[42]：首先，使用top-5主题词通过谷歌搜索英文维基百科，并将top-20搜索结果作为候选图片集；其次，候选图片具有搜索得到的元数据文本和利用尺度不变特征

表5 基于图片的主题标记方法

Table 5 Topic labeling method based on image

方法	学习类型	技术	数据来源	描述	特点
Aletras等人[42]	无监督	图排序	外源	通过检索维基百科,将得到的元数据作为候选图片集;视觉信息利用低阶图像关键特征算子所抽取,最终基于图片的文本和视觉信息,利用图排序算法进行排序	结合文本和视觉信息进行图排序
Nguyen等人[44]	无监督	LDA扩展模型	内源	通过一个多模态、多示例和多标签的潜在狄利克雷分配模型M3LDA,实现主题图片标记。模型由可视标记、文本标记和标记主题三部分组成	自动生成图片、文本标记和主题间的对应关系
Aletras和Mittal[22]	有监督	神经网络	内源	将可视化信息转换为稠密向量表示,利用16层的VGG-net进行训练,训练好的模型可以预测任意一对主题和图像之间的关联程度	具有通用性,可给出适合的图片标签

变换算法[71-72]低阶图像关键特征算子抽取的视觉信息两种模态形式;最后,将候选图片作为结点构建无向图,并排序[57]。实验证明,该方法通常可以找到适合的图片标签,其中视觉信息起到重要的作用。

使用图片对主题进行标记,最难以逾越的语义鸿沟,就是从图像的低层特征到高层语义间建立有效的关联。由于图像本身是一种多语义对象,Nguyen等人[44]在多示例多标记学习框架(multi-instance multi-label learning,MIML)[73]的基础上,提出了一种多模态、多示例和多标签的潜在狄利克雷分配模型(M3LDA),实现对LDA主题的图片标记。模型由三部分组成:可视标记、文本标记和标记主题。其中,可视标记和文本标记的主要任务都是从视觉空间或文本空间到主题标签空间的映射。而标记主题的目标在于发现和维护主题标签之间的联系,即根据不同主题将高度相关的主题标签分组,以形成图片、文本标记和主题之间一对一的客观对应关系。

Aletras和Mittal[22]在其后续研究中提出一种利用深度神经网络预测任意主题和图像间的关联程度方法。主题 T 中包含10个具有最高概率分布的主题词 $T=\{t_1, t_2, \cdots, t_{10}\}$,图片的可视化信息表示为 V,$C=\{c_1, c_2, \cdots, c_n\}$ 为对应的文本信息表示。使用词嵌入[74]方法计算 T 和 C 中所有向量均值,分别表示为 $\boldsymbol{x}_t$ 和 $\boldsymbol{x}_c$;同时将可视化信息 V 转换为稠密向量表示 $\boldsymbol{x}_v$;使用ImageNet数据集[75],利用牛津大学和谷歌共同研发的深度卷积神经网络(16层VGG-net)[76]进行训练。输入 $\boldsymbol{X}=[\boldsymbol{x}_t||\boldsymbol{x}_c||\boldsymbol{x}_v]$,输出为VGG-net所能提供最大的1 000维分类输出向量。基于公开数据集(NYT & WIKI)[42]上的实验结果证明该模型具有广泛的通用性,能准确预测主题和图像间的相关系数,给出合理的图片主题标签。

6 结论与展望

生成式概率主题模型近年来在文本分类、异常检测、推荐系统、文本摘要、观点抽取、词义归纳、情感分析和信息检索等领域[2-3]得到广泛的应用。但主题模型的发现结果通常由一组词汇的概率分布表示,会对用户理解主题造成一定的困扰,也成为主题模型进一步应用与发展的障碍,成为亟需解决的问题。

针对主题模型的自动标注问题,本文综述回顾了现有研究所采用的不同主题标签形式(短语、摘要、图片)和具体方法,从多个维度对标记方法的具体实现和使用场景进行了分析、讨论。并认为主题标记仍存在诸如准确性、扩展性、适用性和理解问题:(1)在处理复杂文本时,容易受到语言表述、噪声等因素的影响,需要进一步提高标记算法的准确性和鲁棒性;(2)目前在处理大规模文本数据时很难保证高效和实时,适用性有限,因此如何解决主题标记算法的扩展性是研究重点之一;(3)不同语言、领域的文本有不同的表达习惯、文化特征和主题偏好,需要研究如何处理不同场景下的主题标记,提高标记方法的普适性和场景适应性;(4)主题标记算法需要更好地理解文本内容和上下文关系,抽取更深层次的语义特征,以提高标记结果的语义表现和精度。

通过对现有研究的梳理、总结和分析,未来主题标记的研究将聚焦于以下方面:

(1)基于深度学习的主题标记。随着深度学习技术在NLP领域研究的不断深入,特别是三个里程碑研究成果:词嵌入[77]的文本特征向量化、word2vec[78]引入大规模预训练语言模型以及Transformer[79]的并行化处理的出现,使得很多NLP中较为困难的任务获得了创新性的解决和突破,如阅读理解、情感分析、推荐系统、信息检索、文本生成等。一些构建在

预训练模型上的系统在很多下游任务中已经超越了人类的表现。因此，利用深度学习技术进行主题标记有着广阔的市场潜力和应用场景。特别是在相关性排序过程中对冗余的控制[61]，例如利用注意力机制对候选句与主题间的相关性，以及候选句与摘要主题标签冗余度联合建模[40]。此外，基于庞大的外源性语料库的预训练语言模型应用于主题标记，以及多种深度学习方法的融合，应该是未来突破的重点和方向。基于上下文的语义理解[30,36]，能更加准确地对相关性和冗余性建模，从而得到与主题更加相关、覆盖性更好、区分性更强和更具解释性的主题标签。

（2）主题标记与情感分析技术的结合。主题标记和情感分析技术相互结合，可获得更准确的文本处理结果，例如可以分析用户对某个主题的情感倾向或从文本中提取情感总结等。Barawi等人[39]首先提出一个与主题一致且情感耦合的摘要标记模型，对情感主题的解释和理解更加有效，可扩展用于从文字语料中提取基于情感分析的不同观点总结。Shahriar等人[36]提出一个基于情感术语和方面术语的单字特征增强分类效果的数据驱动挖掘框架，通过LDA模型输出主题聚类标签，用以揭示与COVID-19大流行相关的各种问题。作者认为结合深度学习技术用于从海量的社交媒体语料中提取情感主题，并生成重要的主题标签，可有效应对社交媒体语料快速增长所导致的数据过载问题。

（3）主题标记应用场景的拓展。在未来，主题标记技术将应用于更多的场景，如智能客服、智能设备等。互联网+时代，对主题发现结果进行自动标记，通常可以用于分类和提取用户描述问题的主题，帮助虚拟助手或人工客服快速了解文本的主要内容和关键特征，将用户问题和数据库中已有的问题进行自动匹配，为用户提供更为精准的帮助，并及时响应。例如，随着消费者对金融服务投诉数量的快速攀升，CFPB（consumer financial protection bureau）[80]意识到由人类专家对这些意见进行人工审查是不可行的，因此构建了一个基于潜在狄利克雷分配的智能分析模型来对投诉意见进行智能分析。通过用户调研和在线调查发现，对客户负面情绪的安抚是提升客户服务的一个非常重要的关键因素[81]。综上可知，主题标记在智能客服中的应用，能够为用户提供更高效便捷的服务，也能为企业提供更精准的数据支持，对于提升客户满意度和企业的服务质量都具有非常重要的意义。

参考文献：

[1] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[2] MEI Q, SHEN X, ZHAI C. Automatic labeling of multinomial topic models[C]//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, Aug 12-15, 2007. New York: ACM, 2007: 490-499.

[3] KOU W, LI F, BALDWIN T. Automatic labelling of topic models using word vectors and letter trigram vectors[C]//LNCS 9460: Proceedings of the 11th Asia Information Retrieval Societies Conference on Information Retrieval Technology, Brisbane, Dec 2-4, 2015. Cham: Springer, 2015: 253-264.

[4] WAN X, WANG T. Automatic labeling of topic models using text summaries[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Aug 7-12, 2016. Stroudsburg: ACL, 2017: 2297-2305.

[5] MEI Q, ZHAI C. Discovering evolutionary theme patterns from text: an exploration of temporal text mining[C]//Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Chicago, Aug 21-24, 2005. New York: ACM, 2005: 198-207.

[6] MEI Q, LIU C, SU H, et al. A probabilistic approach to spatiotemporal theme pattern mining on weblogs[C]//Proceedings of the 15th International Conference on World Wide Web, Edinburgh, May 23-26, 2006. New York: ACM, 2006: 533-542.

[7] LAU J H, GRIESER K, NEWMAN D, et al. Automatic labelling of topic models[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Jun 19-24, 2011. Stroudsburg: ACL, 2011: 1536-1545.

[8] MAGATTI D, CALEGARI S, CIUCCI D, et al. Automatic labeling of topics[C]//Proceedings of the 9th International Conference on Intelligent Systems Design and Applications, Pisa, Nov 30-Dec 2, 2009. Washington: IEEE Computer Society, 2009: 1227-1232.

[9] 凌洪飞, 欧石燕. 面向主题模型的主题自动语义标注研究综述[J]. 数据分析与知识发现, 2019, 3(9): 16-26.
LIN H F, OU S Y. Review of automatic semantic labeling for topic models[J]. Data Analysis and Knowledge Discovery, 2019, 3(9): 16-26.

[10] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.

[11] TURNEY P D, PANTEL P. From frequency to meaning: vector space models of semantics[J]. Journal of Artificial Intelligence Research, 2010, 37: 141-188.

[12] DEERWESTER S, DUMAIS S T, FURNAS G W, et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391-407.

[13] ZHAO W Z, MA H F, HE Q. Parallel K-means clustering based on MapReduce[C]//LNCS 5931: Proceedings of the 1st International Conference on Cloud Computing. Berlin, Heidelberg: Springer, 2009: 674-679.

[14] 周厚奎. 概率主题模型的研究及其在多媒体主题发现和演化中的应用 [D]. 杭州: 浙江大学, 2017.
ZHOU H K. Research on probabilistic topic model and its application in multimedia topic discovery and evolution[D]. Hangzhou: Zhejiang University, 2017.

[15] HOFMANN T. Probabilistic latent semantic indexing[C]// Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, Aug 15-19, 1999. New York: ACM, 1999: 50-57.

[16] TEH Y W, NEWMAN D, WELLING M. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation[C]//Advances in Neural Information Processing Systems 19, Vancouver, Dec 4-7, 2006. Cambridge: MIT Press, 2007: 1353-1360.

[17] PORTEOUS I, NEWMAN D, IHLER A, et al. Fast collapsed Gibbs sampling for latent Dirichlet allocation[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Aug 24-27, 2008. New York: ACM, 2008: 569-577.

[18] CHRISTOU D. Feature extraction using latent Dirichlet allocation and neural networks: a case study on movie synopses [J]. arXiv:1604.01272, 2016.

[19] MEHROTRA R, SANNER S, BUNTINE W, et al. Improving LDA topic models for microblogs via Tweet pooling and automatic labeling[C]//Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Jul 28-Aug 1, 2013. New York: ACM, 2013: 889-892.

[20] JEON H B, LEE S Y. Language model adaptation based on topic probability of latent Dirichlet allocation[J]. ETRI Journal, 2016, 38(3): 487-493.

[21] SANTANIELLO D, COLACE F, LOMBARDI M, et al. Sentiment analysis in social networks: a methodology based on the latent Dirichlet allocation approach[C]//Proceedings of the 11th Conference of the European Society for Fuzzy Logic and Technology, Prague, Sep 9-13, 2019. Amsterdam: Atlantis Press, 2019: 1-8.

[22] ALETRAS N, MITTAL A. Labeling topics with images using a neural network[C]//LNCS 10193: Proceedings of the 39th European Conference on IR Research, Aberdeen, Apr 8-13, 2017. Cham: Springer, 2017: 500-505.

[23] ALETRAS N, STEVENSON M. Labelling topics using unsupervised graph-based methods[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Jun 22-27, 2014. Stroudsburg: ACL, 2014: 631-636.

[24] HULPUS I, HAYES C, KARNSTEDT M, et al. Unsupervised graph-based topic labelling using DBpedia[C]//Proceedings of the 6th ACM International Conference on Web Search and Data Mining, Rome, Feb 4-8, 2013. New York: ACM, 2013: 465-474.

[25] BHATIA S, LAU J H, BALDWIN T. Automatic labelling of topics with neural embeddings[C]//Proceedings of the 26th International Conference on Computational Linguistics, Osaka, Dec 11-16, 2016. Stroudsburg: ACL, 2016: 953-963.

[26] ALOKAILI A, ALETRAS N, STEVENSON M. Re-ranking words to improve interpretability of automatically generated topics[C]//Proceedings of the 13th International Conference on Computational Semantics, Gothenburg, May 23-27, 2019. Stroudsburg: ACL, 2019: 43-54.

[27] KIM H H, RHEE H Y. An ontology-based labeling of influential topics using topic network analysis[J]. Journal of Information Processing Systems, 2019, 15(5): 1096-1107.

[28] SANJAYA N A, BA M L, ABDESSALEM T, et al. Harnessing truth discovery algorithms on the topic labelling problem[C]//Proceedings of the 20th International Conference on Information Integration and Web-based Applications & Services, Yogyakarta, Nov 19-21, 2018. New York: ACM, 2018: 8-14.

[29] KOZONO R, SAGA R. Automatic labeling for hierarchical topics with NETL[C]//Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics, Toronto, Oct 11-14, 2020. Piscataway: IEEE, 2020: 3740-3745.

[30] ZOSA E, PIVOVAROVA L, BOGGIA M, et al. Multilingual topic labelling of news topics using ontological mapping [C]//LNCS 13186: Proceedings of the 44th European Conference on IR Research, Stavanger, Apr 10-14, 2022. Cham: Springer, 2022: 248-256.

[31] POPA C, REBEDEA T. BART-TL: weakly-supervised topic label generation[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Apr 19-23, 2021. Stroudsburg: ACL, 2021: 1418-1425.

[32] KINARIWALA S A, DESHMUKH S. Onto_TML: auto-labeling of topic models[J]. Journal of Integrated Science and Technology, 2021, 9(2): 85-91.

[33] ALOKAILI A, ALETRAS N, STEVENSON M. Automatic generation of topic labels[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Jul 25-30, 2020. New York: ACM, 2020: 1965-1968.

[34] TIWARI P, TRIPATHI A, SINGH A, et al. Advanced hierarchical topic labeling for short text[J]. IEEE Access, 2023, 11: 35158-35174.

[35] ALLAHYARIA M, POURIYEHA S, KOCHUTA K, et al. OntoLDA: an ontology-based topic model for automatic topic labeling[Z]. Amsterdam: IOS Press, 2009: 1-20.

[36] SHAHRIAR K T, MONI M A, HOQUE M M, et al. SATLabel: a framework for sentiment and aspect terms based automatic topic labelling[C]//Proceedings of Machine Intelligence and Data Science Applications 2021, Cumilla, Dec 2021. Berlin, Heidelberg: Springer, 2022: 63-75.

[37] HE D, WANG M, KHATTAK A M, et al. Automatic labeling of topic models using graph-based ranking[J]. IEEE Access, 2019, 7: 131593-131608.

[38] BASAVE A E C, HE Y, XU R. Automatic labelling of topic models learned from twitter by summarisation[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Jun 22-27, 2014. Stroudsburg: ACL, 2014: 618-624.

[39] BARAWI M H, LIN C, SIDDHARTHAN A. Automatically labelling sentiment-bearing topics with descriptive sentence labels[C]//LNCS 10260: Proceedings of the 22nd International Conference on Applications of Natural Language to Information Systems, Liège, Jun 21-23, 2017. Cham: Springer, 2017: 299-312.

[40] HE D, REN Y, KHATTAK A M, et al. Automatic topic labeling model with paired-attention based on pre-trained deep neural network[C]//Proceedings of the 2021 International Joint Conference on Neural Networks, Shenzhen, Jul 18-22, 2021. Piscataway: IEEE, 2021: 1-9.

[41] KOZBAGAROV O, MUSSABAYEV R, MLADENOVIC N. A new sentence-based interpretative topic modeling and automatic topic labeling[J]. Symmetry, 2021, 13(5): 837.

[42] ALETRAS N, STEVENSON M. Representing topics using images[C]//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Jun 9-14, 2013. Stroudsburg: ACL, 2013: 158-167.

[43] SORODOC I, LAU J H, ALETRAS N, et al. Multimodal topic labelling[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2017: 701-706.

[44] NGUYEN C T, ZHAN D C, ZHOU Z H. Multi-modal image annotation with multi-instance multi-label LDA[C]//Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, Aug 3-9, 2013. Menlo Park: AAAI, 2013: 1558-1564.

[45] ALETRAS N, BALDWIN T, LAU J H, et al. Evaluating topic representations for exploring document collections[J]. Journal of the Association for Information Science and Technology, 2017, 68(1): 154-167.

[46] MAO X L, MING Z Y, ZHA Z J, et al. Automatic labeling hierarchical topics[C]//Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York: ACM, 2012: 2383-2386.

[47] REIMERS N, GUREVYCH I. Sentence-BERT: sentence embeddings using siamese BERT-networks[J]. arXiv:1908.10084, 2019.

[48] LEWIS M, LIU Y, GOYAL N, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[J]. arXiv:1910.13461, 2019.

[49] CHEN J, YAN J, ZHANG B, et al. Diverse topic phrase extraction through latent semantic analysis[C]//Proceedings of the 6th International Conference on Data Mining, Hong Kong, China, Dec 18-22, 2006. Washington: IEEE Computer Society, 2007: 834-838.

[50] CHINCHOR N, ROBINSON P. MUC-7 named entity task definition[C]//Proceedings of the 7th Conference on Message Understanding. Stroudsburg: ACL, 1998: 1-21.

[51] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 27th Annual Conference on Neural Information Processing Systems 2013, Lake Tahoe, Dec 5-8, 2013. Red Hook: Curran Associates, 2013: 3111-3119.

[52] LE Q V, MIKOLOV T. Distributed representations of sentences and documents[C]//Proceedings of the 31st International Conference on Machine Learning, Beijing, Jun 21-26, 2014: 1188-1196.

[53] PENNINGTON J, SOCHER R, MANNING C D. GloVe: global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Oct 25-29, 2014. Stroudsburg: ACL, 2014: 1532-1543.

[54] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 Conference of the North Ame-

rican Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Jun 2-7, 2019. Stroudsburg: ACL, 2019: 4171-4186.

[55] HULPUS I, HAYES C, KARNSTEDT M, et al. An eigenvalue-based measure for word-sense disambiguation[C]//Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference, Marco Island, May 23-25, 2012. Menlo Park: AAAI, 2012: 1-6.

[56] BOUMA G. Normalized (pointwise) mutual information in collocation extraction[C]//Proceedings of the 2009 International Conference of the German Society for Computational Linguistics and Language Technology, Potsdam, 2009: 31-40.

[57] PAGE L, BRIN S, MOTWANI R, et al. The pagerank citation ranking: bringing order to the web[R]. Stanford InfoLab, 1999: 1-17.

[58] SMITH A, LEE T Y, POURSABZI-SANGDEH F, et al. Evaluating visual representations for topic understanding and their effects on manually generated topic labels[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 1-16.

[59] CARBONELL J, GOLDSTEIN J. The use of MMR, diversity-based reranking for reordering documents and producing summaries[C]//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Aug 24-28, 1998. New York: ACM, 1998: 335-336.

[60] MIHALCEA R, TARAU P. TextRank: bringing order into text [C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, A Meeting of SIGDAT, a Special Interest Group of the ACL, Held in Conjunction with ACL 2004, Barcelona, Jul 25-26, 2004. Stroudsburg: ACL, 2004: 404-411.

[61] HE D, REN Y, KHATTAK A M, et al. Automatic topic labeling using graph-based pre-trained neural embedding[J]. Neurocomputing, 2021, 463: 596-608.

[62] REN P, CHEN Z, REN Z, et al. Sentence relations for extractive summarization with deep neural networks[J]. ACM Transactions on Information Systems, 2018, 36(4): 1-32.

[63] REN P, WEI F, ZHUMIN C, et al. A redundancy-aware sentence regression framework for extractive summarization[C]// Proceedings of the 26th International Conference on Computational Linguistics, Osaka, Dec 11-16, 2016. Stroudsburg: ACL, 2016: 33-43.

[64] FUJISHIGE S. Submodular functions and optimization[M]. New York: Elsevier Science Inc., 2005.

[65] LIN H, BILMES J. Multi-document summarization via budgeted maximization of submodular functions[C]//Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, Jun 2-4, 2010. Stroudsburg: ACL, 2010: 912-920.

[66] LIN H, BILMES J. A class of submodular functions for document summarization[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Jun 19-24, 2011. Stroudsburg: ACL, 2011: 510-520.

[67] MALLICK C, DAS A K, DUTTA M, et al. Graph-based text summarization using modified TextRank[M]//Soft Computing in Data Analytics. Cham: Springer, 2019: 137-146.

[68] BRIN S, PAGE L. The anatomy of a large-scale hypertextual web search engine[J]. Computer Networks and ISDN Systems, 1998, 30: 107-117.

[69] ERKAN G, RADEV D R. LexRank: graph-based lexical centrality as salience in text summarization[J]. Journal of Artificial Intelligence Research, 2004, 22: 457-479.

[70] LIU Y. Fine-tune BERT for extractive summarization[J]. arXiv: 1903.10318, 2019.

[71] LOWE D G. Object recognition from local scale-invariant features[C]//Proceedings of the 1999 International Conference on Computer Vision, Kerkyra, Sep 20-25, 1999. Washington: IEEE Computer Society, 1999: 1150-1157.

[72] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[73] ZHOU Z H, ZHANG M L. Multi-instance multi-label learning with application to scene classification[C]//Proceedings of the 2006 International Conference on Neural Information Processing Systems, Vancouver, Dec 4-7, 2006. Cambridge: MIT Press, 2006: 1609-1616.

[74] LEVY O, GOLDBERG Y. Dependency-based word embeddings[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Jun 22-27, 2014. Stroudsburg: ACL, 2014: 302-308.

[75] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, Jun 20-25, 2009. Washington: IEEE Computer Society, 2009: 248-255.

[76] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv:1409.1556, 2014.

[77] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural

probabilistic language model[J]. Journal of Machine Learning Research, 2003, 3: 1137-1155.

[78] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv: 1301.3781, 2013.

[79] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems 30, Long Beach, Dec 4-9, 2017. Red Hook: Curran Associates, 2017: 5998-6008.

[80] BASTANI K, NAMAVARI H, SHAFFER J. Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints[J]. Expert Systems with Applications, 2019, 127: 256-271.

[81] SONG S, WANG C, CHEN H, et al. An emotional comfort framework for improving user satisfaction in E-commerce customer service chatbots[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers, Jun 6-11, 2021. Stroudsburg: ACL, 2021: 130-137.

何东彬(1973—),男,河北张家口人,博士,副教授,高级工程师,主要研究方向为人工智能、自然语言处理等。

HE Dongbin, born in 1973, Ph.D., associate professor, senior engineer. His research interests include artificial intelligence, natural language processing, etc.

陶莎(1985—),女,北京人,博士,副教授,主要研究方向为农产品安全信息与智能处理技术、多源数据融合技术等。

TAO Sha, born in 1985, Ph.D., associate professor. Her research interests include agricultural product safety information and intelligent processing, multi-source data fusion, etc.

朱艳红(1974—),女,河北定州人,硕士,教授,主要研究方向为人工智能、移动开发等。

ZHU Yanhong, born in 1974, M.S., professor. Her research interests include artificial intelligence, mobile development, etc.

任延昭(1986—),男,山东聊城人,博士,讲师,主要研究方向为农业大数据分析、农产品安全风险智能分析等。

REN Yanzhao, born in 1986, Ph.D., lecturer. His research interests include agricultural big data analysis, intelligent analysis of agricultural product safety risk, etc.

褚云霞(1980—),女,河北石家庄人,硕士,副教授,主要研究方向为物联网、现代教育技术等。

CHU Yunxia, born in 1980, M.S., associate professor. Her research interests include Internet of things, modern educational technology, etc.