doi:10.6041/j.issn.1000-1298.2020.07.028

面向食品安全事件新闻文本的实体关系抽取研究

郑丽敏[1,2]　齐珊珊[1]　田立军[1,2]　杨　璐[1,2]

(1.中国农业大学信息与电气工程学院,北京 100083;2.食品质量与安全北京实验室,北京 100083)

摘要:为解决从大规模网络文本中快速、准确识别食品安全事件并进行实体关系抽取受中文复杂语法特性限制的问题,提出一种基于依存分析的面向食品安全事件新闻文本的实体关系抽取方法 FSE_ERE(Entity relation extraction of food safety events, FSE_ERE)。该方法结合句子的依存分析结果和实体关系抽取模型,对非结构化中文文本进行无监督的实体关系抽取,并引入一种将文本相似度结合到 PU 学习(Positive and unlabeled learning)的半监督分类方法,利用改进的特征加权处理方法提高分类精度,使得 FSE_ERE 方法能够在高质量的食品安全事件新闻文本中完成实体关系抽取工作。实验结果表明,FSE_ERE 方法在食品安全事件新闻文本数据集和多类型混合新闻文本数据集上的实体关系抽取均达到了先进的性能,F 值分别达到了 71.21% 和 67.42%,证明了 FSE_ERE 方法的有效性和可移植性。

关键词:食品安全事件;实体关系抽取;依存分析;PU 学习;文本相似度

中图分类号:TP391;TP311　　**文献标识码**:A　　**文章编号**:1000-1298(2020)07-0244-10　　**OSID**:

Entity Relation Extraction of News Texts for Food Safety Events

ZHENG Limin[1,2]　QI Shanshan[1]　TIAN Lijun[1,2]　YANG Lu[1,2]

(1. *College of Information and Electrical Engineering*, *China Agricultural University*, *Beijing* 100083, *China*
2. *Beijing Laboratory of Food Quality and Safety*, *Beijing* 100083, *China*)

Abstract: In order to solve the problem of fast and accurate identification of food safety events from large-scale Web texts and the extraction of entity relations, which is limited by the complex grammatical characteristics of Chinese, a method of entity relation extraction based on dependency parsing for news texts of food safety events FSE_ERE (entity relation extraction of food safety events) was proposed. This method combined the dependency parsing results of sentences with the entity relation extraction model to conduct unsupervised entity relation extraction for unstructured Chinese texts, and also introduced a semi-supervised classification method combining text similarity with positive and unlabeled learning (PU learning) classification method, which used an improved feature weighting processing method to improve the classification accuracy. That can make the FSE_ERE method to complete the entity relation extraction work in the high-quality news text of food safety events. The experimental results showed that the FSE_ERE method achieved advanced performance in entity relation extraction on food safety event news text dataset and multi-type hybrid news text dataset, and the F-measure achieved 71.21% and 67.42% respectively, which proved the effectiveness and portability of the FSE_ERE method.

Key words: food safety events; entity relation extraction; dependency parsing; positive and unlabeled learning; text similarity

0　引言

食品安全事件频发,注水肉、过期奶粉等事件严重影响了民众的生活,造成了严重的后果[1]。网络上各种新闻文本的数量迅猛增长,如何快速、准确地获取食品安全事件新闻文本,并理清其中的关系脉络是一项耗时、耗力的工作。食品安全事件新闻文本的分析梳理对于消费者和管理者均具有重要意

收稿日期:2019-11-01　修回日期:2019-12-24

基金项目:国家重点研发计划项目(2017YFC1601803)、现代农业产业技术体系北京市生猪产业创新团队项目(BAIC02-2019)、国家蛋鸡产业技术体系项目(CARS-40-K27)和"十二五"国家科技支撑计划项目(2013AD19B09)

作者简介:郑丽敏(1962—),女,教授,主要从事计算机视觉与人工智能研究,E-mail:zhenglimin@cau.edu.cn

义：消费者能够从杂乱冗余的大量数据中快速获取事件的主要信息，对事件的发展走向有系统的认知，提前做出正确的预防或应对，减轻事件带来的伤害；管理者利用梳理出来的信息快速决策，及时发布并通知、提醒各部门或消费者采取相应措施等。实体关系抽取能够从半结构化和非结构化的信息源中抽取出实体及实体之间的语义关系，在数据挖掘、问答系统、知识图谱构建等研究中均扮演着重要角色，是实现分析梳理的基础，受到越来越多研究者的关注[2-4]。

实体关系抽取方法有半监督式、远程监督式和无监督式 3 种[3]。其中，半监督式的实体关系抽取需要选取少量的种子，种子的品质会直接影响抽取效果，且受人的主观影响明显[5]；远程监督式的实体关系抽取需要大规模知识库的支撑，但适用于各领域的大规模知识库很难找到，所以该方法并不适用于众多领域[3,6-7]；无监督式的实体关系抽取无需任何人工标注数据、预定义关系类型等，适用于开放领域的关系抽取[8-9]。目前，英文的实体关系抽取研究已经达到较高的水平，由最初的开放式信息抽取系统 TextRunner[10] 发展到 O-CRF[11]、ReVerb 系统[12]、Ollie 系统[13]等，性能不断提高。中文实体关系抽取却发展缓慢，主要是由于中文语法具有复杂多变、无标准句式、实体参数位置不固定等特点，导致中文文本的实体关系抽取难度远远高于英文文本。文献[14]提出第一个开放领域实体关系抽取系统 ZORE，在语义层面进行研究，具有有效性，但随着召回率的提高，准确率下降趋势过于明显。文献[15]提出用于知识获取的中文开放信息抽取的 CORE 系统，证明了从中文语料库中抽取关系而不向 IE 系统输入任何预定义词汇和关系的可行性，但并未在大规模的新闻文本数据集上进行充分的实验。之后针对不同的数据类型，在 ZORE、CORE 的基础上出现了 GCORE[16]、C-COERE[17] 等系统，性能得到了优化。

但是这些方法对所有的文本采取相同的处理方式，未充分考虑食品安全事件新闻文本的以下特性：发生主题、涉事食品、食品种类、涉事企业、企业负责人、涉事人员、发生时间、发生地点、发生原因、发生规模、导致结果、产生影响及危害等，无法对网络上食品安全事件新闻及时预警，在一定程度上降低了事件时效性。针对这一问题，本文提出一种基于依存分析的食品安全事件新闻文本的实体关系抽取方法 FSE_ERE，充分考虑中文新闻文本的语言特性，利用 LTP 工具[18]对句子进行分词、词性标注、命名实体识别处理后，对各个语言单位内成分之间的依存关系进行分析，揭示句子的句法结构。再结合这些知识和构建的实体关系抽取模型抽取出其中包含的实体关系三元组，实现中文新闻文本中实体和关系的自动抽取，无需任何人工干预。质量高且类别明确的文本能有效提高抽取模型和依存分析结果的匹配度，从而提高抽取性能。因此在实体关系抽取过程中引入半监督的 PU 学习分类方法，创造性地将文本相似度结合到 PU 学习分类方法中，通过改进的特征选取与加权处理方法提高分类的精度，以节省时间和人力。

1　FSE_ERE 方法

FSE_ERE 方法主要包含两部分内容：①为了获取更多高质量的文本数据，在大规模的新闻文本中利用基于 PU 学习的分类模型提取食品安全事件新闻文本。②在提取的文本的基础上，利用基于依存分析的模型进行实体关系抽取工作。

1.1　分类模型描述

分类问题是机器学习的一个重要组成部分，目前大多数分类方法是根据已知样本的某些特征后判定新样本的类别[19-20]。文本分类一般要经过文本预处理、特征选择、分类器训练和性能评估 4 个步骤[21-22]。本文主要解决的问题是在众多互联网文本中，在只含有积极样例的情况下，快速地挑选出高质量的食品安全事件类文本，以便进行实体关系抽取工作。本文中出现的积极样例是食品安全事件新闻文本，消极样例是非食品安全事件的其他各个类别的新闻文本，未标记样例是大规模的网络新闻文本。

1.1.1　关键特征

在文本预处理过程中，分词和去停用词是主要步骤。由于目前的自然语言处理工具仍存在一定的缺陷，无法全面、准确地识别出文本中存在的领域专有名词，尤其是食品安全事件领域中特有的食品名称、发生原因（即引起食品安全事件发生的具体因素）等。例如，“毒鸡蛋”是食品安全领域中出现的一种问题食品的名称，分词工具通常会将其分词为“毒”和“鸡蛋”两部分。但是“鸡蛋”只是普通食品的名称，并不能作为食品安全事件的问题食品，这就造成食品安全事件的主体食品判定的错误，影响事件的分析研究。因此，领域词典在分词、词性标注、命名实体识别过程中发挥着重要作用，能够辅助自然语言处理工具更全面地、更准确地识别出文本中的重要信息，还能够帮助选取重要特征，提高分类精度。

通过对食品安全事件统计分析和对中文新闻文

本表达特点进行研究,发现与其他类型的新闻相比,不论食品安全事件新闻文本的完整程度如何,通常会包含以下特性:涉事食品、发生原因、涉事企业和发生地点 4 项,因此将这 4 项作为关键特征。为了保证它们的正确性,分别构建了关于 4 项关键特征的领域词典,并将这 4 个词典称为关键特征词典。关键特征词典中的词汇是从国家药品监督管理局、食品伙伴网等网站的相关模块中爬取的专有名词,共 273 709 个,各个特征项对应的领域词典中包含词的个数统计结果如表 1 所示。其中发生原因包括食品添加剂、真菌毒素、污染物、农兽药方面的专有词汇;发生地点包括省级行政区、地级市、县级市和县。

表 1 各个特征项的领域词典中包含的词个数

Tab. 1 Number of words in domain dictionary of each feature item

关键特征项	领域词典的词	领域词典中词汇的个数	总数
涉事食品	各种食品名称	148 884	148 884
发生原因	食品添加剂	623	7 805
	真菌毒素	3 570	
	污染物	124	
	农兽药	3 488	
涉事企业	各个企业名称	114 822	114 822
发生地点	省级行政区	34	2 198
	地级市	333	
	县级市	370	
	县	1 461	

预处理时,对文本进行清洗,包括去除链接、空格、无意义字符,并利用分词工具对文本进行分词操作后,在分词系统中引入上述关键特征词典,能够明显提高分词的准确率。此外,在得到每个文本的分词结果后,还需要进行去停用词处理,因为这些停用词虽然词频高但是对文本分类贡献小。则文档集所有剩余的分词结果构成了一个词典向量。该词典向量与关键特征词典中存在一些相同的词汇,为了避免特征重复,删除词典向量中这部分重复的词汇。

1.1.2 特征模板生成

TF - IDF 算法是一种目前最为常用且非常有效的特征提取方法,根据计算的特征权重评估每个特征对文本的重要程度。本文采用 TF - IDF 方法计算所有特征词在每篇文档中的特征权重,但传统的 TF - IDF 没有考虑特征词在类间分布状况的影响。所以本文在 TF - IDF 中引入特征选择效果较好的卡方统计量(Chi-square, CHI)方法进行修正。

CHI 用于表示特征词与类别之间的相关程度,CHI 越高则表示相关程度越高,对应的特征词不仅更能代表某个类别,还具有更高的权重。CHI 计算公式为[22-23]

$$V_{\mathrm{CHI}}(t_j, C_i) = \frac{(AD - BC)^2 |X|}{(A + C)(B + D)(A + B)(C + D)} \tag{1}$$

式中 V_{CHI}——卡方统计量(CHI)

t_j——第 j 个一般特征词

C_i——第 i 个类别

$|X|$——数据集中的文档总数目

其中 A、B、C 和 D 的含义如表 2 所示。

表 2 特征与类别关系

Tab. 2 Relationship between features and categories

参数	属于类别 C_i	不属于类别 C_i
包含特征词 t_j 的文档数目	A	B
不包含特征词 t_j 的文档数目	C	D

此外,文本关键特征也能够明显区分类别间的差异,对分类产生较好的影响。所以将涉事食品、发生原因、涉事企业、发生地点 4 项关键特征补充到选取的特征词后面,生成特征模板。虽然关键特征对应的词汇集合与类别相关性最大,但是它们在文档中出现的次数并不多,导致了其权重低。所以在改进的关键特征权重计算方法的基础上还引入了关键特征因子 λ,以实现加权处理。λ 是经过大量实验后得出的一个经验系数,本文取值为 3。

计算关键特征的权重时,应统计关键特征 p_g 对应的关键特征词典中的词汇在文档 x_i 中的频率,并计算关键特征的逆文档频率(Inverse document frequency, IDF),最后计算出关键特征在文档 x_i 中的权重。计算公式为

$$D(w_{p_g}) = \lambda(p_g) T_{p_g} I_{p_g} \tag{2}$$

式中 w_{p_g}——关键特征的权重

$D(w_{p_g})$——关键特征在文档 x_i 中的权重

T_{p_g}——关键特征的 TF 值

I_{p_g}——关键特征的 IDF 值

T_{p_g} 的主要思想是:关键特征 p_g 是一类特征的集合,如果 p_g 在文本中出现的不同词汇数多且频次高,说明这篇文档描述了很多关于 p_g 的内容,与 p_g 相关程度高,则可以认为文档属于 p_g 相关的类别。I_{p_g} 的主要思想是:据统计分析,关键特征 p_g 涉及的某些词汇在大多数文档中出现频率都比较低,但这些特征词对文本分类的作用却十分明显,它们对分类贡献率高却容易被忽略掉,所以 I_{p_g} 被用于表示关键特征 p_g 对于整个文档集的重要程度,即当包含 p_g 的文档数目越少时,p_g 对文本分类贡献率会越高。T_{p_g} 和 I_{p_g} 的计算方法分别为

$$T_{p_g}=\frac{\sum_{p_g} n(D_{p_g},x_i)}{\sum_{k} n_{k,x_i}} \tag{3}$$

$$I_{p_g}=\lg\frac{|X|}{1+N(p_g)} \tag{4}$$

式中　D_{p_g}——p_g对应的关键特征词典中的词汇

$n(D_{p_g},x_i)$——D_{p_g}在文档 x_i中出现的频次

n_{k,x_i}——文档 x_i中词汇 k 出现的次数

$N(p_g)$——包含关键特征 p_g的文档数目

在式(4)中,分母项加 1 是对其进行了平滑处理,防止该词语不在语料库中时导致的除数为零现象发生。

最后,由于大多数文档长度不一样,TF－IDF 算法会出现偏向于长文本的情况,所以需要对 TF－IDF 算法的计算结果作统一的归一化处理。同时将特征词的 CHI 进行对数化处理,以解决权重不均衡问题。综上所述,本文改进后生成的特征模板中,一般特征权重计算公式为

$$w_{t_j,p_g}=\lg(V_{\mathrm{CHI}})\frac{F_{t_j}E_{t_j}}{\sqrt{\sum_{j=1}^{m_1}(F_{t_j}E_{t_j})^2+\sum_{g=m_1+1}^{m_2}(T_{p_g}I_{p_g})^2}} \tag{5}$$

关键特征的权重计算公式为

$$w_{t_j,p_g}=\lambda(p_g)\frac{T_{p_g}I_{p_g}}{\sqrt{\sum_{j=1}^{m_1}(F_{t_j}E_{t_j})^2+\sum_{g=m_1+1}^{m_2}(T_{p_g}I_{p_g})^2}} \tag{6}$$

式中　m_1——一般特征词的数目

m_2——关键特征的数目

F_{t_j}——第 j 个一般特征词的词频

E_{t_j}——第 j 个一般特征词的逆文档频率指数

利用向量空间模型(Vector space model, VSM)方法对文本进行文本向量化表示,用于文本分类器的训练。对于一篇食品安全事件新闻文档 x_i,其向量表示为

$$\boldsymbol{x}_i=(w_1,w_2,\cdots,w_i,\cdots,w_{m_1},\cdots,w_j,\cdots,w_{m_1+m_2})\quad(1\leqslant i\leqslant m_1\leqslant j\leqslant m_1+m_2) \tag{7}$$

式中　w_i——第 i 个特征对应的特征权重

w_j——第 j 个特征对应的特征权重

1.1.3　寻找消极样例和建立分类器

提出的 PU 学习分类模型采用两步法实现。

(1)寻找消极样例

第 1 步是在未标记样例中寻找一部分与积极样例极其不同的样例(反差大的样例)作为消极样例,详细流程如图 1 所示。首先将一部分积极样例放入未标记样例中,然后对未标记样例集合进行聚类。未标记样例集合经聚类后形成大小不同的簇。去除包含间谍样例的簇(认为簇中不含有消极样例),并对剩余簇内的文本进行相似度计算,删除相似度高的文本。因为对于大规模的网络食品安全事件新闻文本,同一篇新闻有很大概率在多个网站上被发布,或者即使不同新闻对同一事件的表述不完全一致但相似度也很高,这样的新闻则对于信息挖掘、关系抽取意义不大,因此这种多余的相似文本应该被去除。最后计算各个簇与积极样例集合之间的距离,选出差异最大的簇,将该簇中的文本标记为消极样例。

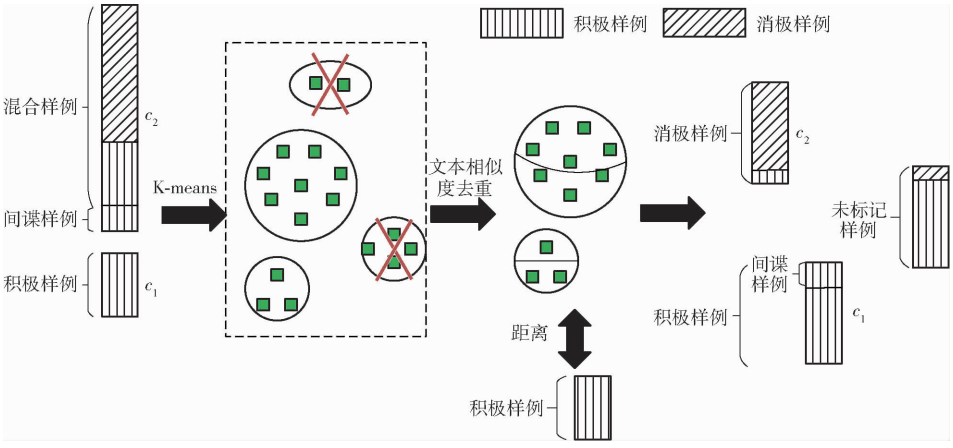

图 1　第 1 步的过程演示

Fig. 1　Process demonstration of the first step

图 1 中,采用 K-means 算法进行聚类,由于传统的 K-means 算法假设每个样本对最终聚类结果的贡献程度一样,未考虑关键特征对于聚类过程的影响,导致聚类准确率低。所以应用上述特征加权处理改进方法获得的特征能够有效解决这一问题。

此外,还需要去除重复文本以提高分类效果和文本质量,例如,对于同一事件不同描述的新闻文本,其文本相似度超过阈值时认为不同文本描述了

同一事件,只保留最近时间报道的且信息最丰富的新闻文本;对于同一涉事食品在不同地区发生的食品安全事件,根据文本的“发生地点”特征对应的地点词汇是否相同来判断是否属于同一个事件。所以删除包含间谍样例的簇后在剩下的各个簇中分别利用文本提取特征来计算文本相似度,得到的向量形式表示的文本之间以空间距离体现语义相似度[24]。对于向量化后的特征,采用最常用的余弦相似度计算方法,表示为

$$\mathrm{Sim}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\frac{\boldsymbol{x}_i\boldsymbol{x}_j^{\mathrm{T}}}{\|\boldsymbol{x}_i\|\|\boldsymbol{x}_j\|}\quad(i\neq j)\tag{8}$$

式中 $\boldsymbol{x}_j$——第 j 个待计算文本的向量

相似度越大,说明距离越小,文本越相似。

(2)建立分类器

第 2 步,根据积极样例的集合 P、消极样例的集合 N 和未标记样例的集合 U 建立最终的分类器。具体过程如下:①将所有的间谍样例 S 都放回到积极样例集合 P 中。②给积极样例集合 P 中的每个文档 x_i 都分配固定的类标签 c_1,即 $y(c_1,x_i)=1$,且在每次迭代 EM 最大期望算法时,标签不再改变。③为消极样例集合 N 中的每个文档 x_j 都分配初始类标签 c_2,即 $y(c_2,x_j)=0$,且在每次迭代 EM 算法时,标签都会改变。④在未标记样例集合 U 中的每一个文档 x_k 都没有被分配标签,但是在 EM 算法的第一次迭代后,将会分配给每个文档一个概率标签。在随后的迭代过程中,集合 U 将通过其新分配的概率类型参与 EM 算法,例如 $y(c_1,x_k)$。⑤在集合 P、N 和 U 中重复运行 EM 算法直至收敛。

当 EM 算法结束时,将生成最终的分类器。本文将用该分类器分类食品安全事件并进行性能评估,用于后续的实体关系抽取工作。

1.2 基于依存分析的实体关系抽取

基于依存分析的食品安全事件实体关系抽取的目标是从大规模的食品安全事件新闻文本中抽取出食品安全事件中的实体及实体之间(或实体与属性值之间)的语义关系,其中实体涉及到涉事食品、涉事公司、涉事人员等;属性包括产品规格、商标形式等。面对复杂多变的中文新闻表达形式,关系抽取模型需要具有广泛性和强的鲁棒性才能够达到好的抽取效果。

(1)关系识别

动词及动词短语、名词及名词短语和位于它们前面或后面相邻的说明性修饰符均可作为关系词或关系短语。关系可以位于句子中的任意位置[16,25],能够根据模型和候选关系与句子其他成分之间的依存关系来确定元组关系。一般情况下,主语和谓语之间会通过依存关系“SBV”等来连接,谓语和宾语之间会通过依存关系“VOB”、“POB”等来连接。此外,还存在一种特殊的偏正结构,如“食药监局长×××”一句中,“局长”、“食药监”和“×××”均为名词,“局长”作为“食药监”和“×××”之间的关系,与它们之间的依存关系均为“ATT”,可抽取出实体关系三元组(食药监,局长,×××)。

(2)实体和属性识别

实体和属性识别是为了识别出每个待处理句子中的实体对(arg1,arg2),arg1 和 arg2 参数分别表示主语和宾语,arg1 为实体,arg2 为与 arg1 之间存在关系的另一个实体或者 arg1 具有的某种属性的属性值[3]。本文应用 LTP 工具分析待处理的文本,将所有句子依次进行分词、词性标注、命名实体识别和依存句法分析。还引入了涉事食品、发生原因、涉事企业和发生地点 4 个关键词典辅助分词,提高分词准确率和召回率,进而提高整体抽取性能。其中命名实体识别能够识别出句子中的所有可能实体,作为实体关系三元组的候选实体,依存句法分析对句子成分及各成分之间的语义关系进行分析,确定三元组成分。

接下来计算任意 2 个候选实体之间存在的实体数量和其他词语的数量。文献[14,26]经过统计和实验研究发现,在候选实体组成实体对后,限定每个实体对之间存在的其他候选实体数目不超过 4 个,词汇总数目不超过 5 个时,得到的三元组的准确率达到最高。这是因为句子中 2 个实体距离越远,两者之间存在关系的可能性就越小。根据依存分析的结果,检测关系词或关系短语所依赖的实体。

基于模型的实体关系抽取,是将句子的依存分析结果和基于中文语法规则的模型进行匹配完成抽取工作的。本文依据大规模新闻文本的依存分析结果中所包含的语义特征提出了中文关系抽取模型 ORE_Models,包含 ORE_Model1、ORE_Model2、ORE_Model3,具体结构如图 2 所示,图中各参数的含义如表 3 所示。

在图 2 中,关系抽取模型 ORE_Model1 多用于抽取以动词作为关系和存在介宾关系时的句子形式;关系抽取模型 ORE_Model2 多用于抽取主语,或谓语,或宾语中存在一个或多个并列情况的句子形式,其中 pred1 和 arg3 2 个节点之间由有方向的实线和虚线表示的关系所连接,但实线和虚线有且仅有一种出现,即在一个句子中不可同时存在;关系抽取模型 ORE_Model3 多用于抽取存在动补结构、偏正结构时的句子形式。每个待处理的中文句子的依存分析结果只要与模型的某一部分正确匹配且匹配成功的部分中存在可抽取的内容,就会以实体关系三元组的形式输出。其中节点及关系存在情况与可

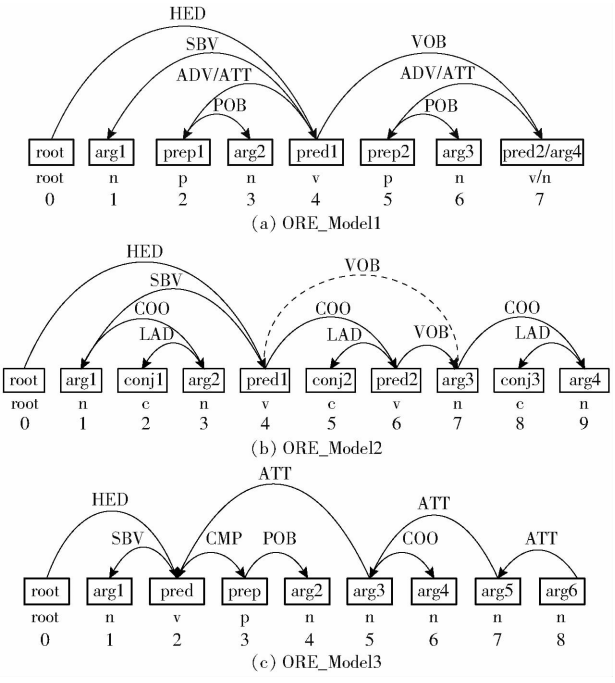

图 2　中文关系抽取模型 ORE_Models

Fig. 2　Chinese relation extraction model ORE_Models

表 3　ORE_Models 模型中参数含义

Tab. 3　Meaning of parameters in model ORE_Models

参数	含义
arg*i*	第 *i* 个实体
prep*i*	第 *i* 个介词
prep	介词
pred*i*	第 *i* 个动词
pred	动词
conj*i*	第 *i* 个连词
root	根节点
n	词性为名词
p	词性为介词
v	词性为动词
c	词性为连词
数字 0,1,…,9	词在模型中的位置编号

抽取出的实体关系三元组的情况如表 4 所示。为了便于展示,仅在表 4 的可抽取出的实体关系三元组中展示了实体和关系,但在实际抽取过程中还保存了与实体存在“ATT”等依存关系的实体修饰词汇。

表 4 中的“－”表示 2 个节点的连接组合,共同

表 4　ORE_Models 中节点及关系存在情况不同时的可抽取的实体关系三元组

Tab. 4　Extractable entity relation triples with different nodes and relations in ORE_Models

模型	模型中存在的节点	模型中存在的依存关系	可抽取出的实体关系三元组
ORE_Model1	arg1、pred1、arg4	SBV、VOB	(arg1, pred1, arg4)
	arg1、prep1、arg2、pred1、arg4	SBV、ADV/ATT、POB、VOB	(arg1, pred1, arg4) (arg1, prep1, arg2)
	arg1、pred1、prep2、arg3、pred2/arg4	SBV、ADV/ATT、POB、VOB	(arg1, pred1, arg4) (arg1, prep2, arg3)
	arg1、prep1、arg2、pred1、prep2、arg3、pred2/arg4	SBV、ADV/ATT、POB、VOB	(arg1, pred1, arg4) (arg1, prep1, arg2) (arg1, prep2, arg3)
ORE_Model2	arg1、conj1、arg2、pred1、arg3	SBV、COO、LAD、VOB	(arg1, pred1, arg3) (arg2, pred1, arg3)
	arg1、pred1、conj2、pred2、arg3	SBV、COO、LAD、VOB	(arg1, pred1, arg3) (arg1, pred2, arg3)
	arg1、pred1、arg3、conj3、arg4	SBV、COO、LAD、VOB	(arg1, pred1, arg3) (arg1, pred1, arg4)
	arg1、conj1、arg2、pred1、conj2、pred2、arg3	SBV、COO、LAD、VOB	(arg1, pred1, arg3) (arg2, pred1, arg3) (arg1, pred2, arg3) (arg2, pred2, arg3)
	arg1、conj1、arg2、pred1、arg3、conj3、arg4	SBV、COO、LAD、VOB	(arg1, pred1, arg3) (arg2, pred1, arg3) (arg1, pred1, arg4) (arg2, pred1, arg4)
	arg1、pred1、conj2、pred2、arg3、conj3、arg4	SBV、COO、LAD、VOB	(arg1, pred1, arg3) (arg1, pred1, arg4) (arg1, pred2, arg3) (arg1, pred2, arg4)
ORE_Model3	arg1、pred、prep、arg2	SBV、CMP、POB	(arg1, pred－prep, arg3)
	pred、prep、arg2、arg3	ATT、CMP、POB	(arg3, pred－prep, arg2)
	pred、prep、arg2、arg3、arg4、arg5	ATT、CMP、POB、COO	(arg5－arg3, pred－prep, arg2) (arg5－arg4, pred－prep, arg2)
	arg3、arg5、arg6	ATT	(arg3, arg5, arg6)

组成三元组的主语或者谓语,“/”和图 2 中的“/”均表示“或者”的含义,即两种情况均可能出现(但不可能同时出现)。从表 4 中可以发现模型 ORE_Models 覆盖了多种句子形式,能够处理具有多变的语法表达方式的新闻文本。

例如,句子“上海市食药监局查封了一批毒鸡蛋”的依存分析结果如图 3 所示。从图 3 中可以得到候选实体有“上海市食药监局”(机构名称)、“毒鸡蛋”,关系词为“查封”,它们之间的依存关系符合模型 ORE_Model1 的分析,最后可抽取出实体关系三元组(上海食药监,查封,一批毒鸡蛋)。

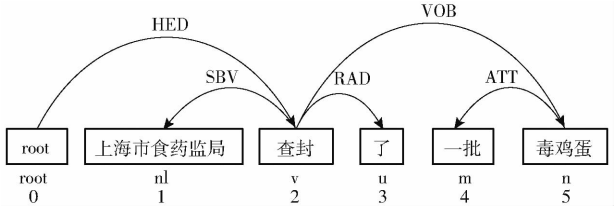

图 3 实例 1 的句子依存分析结果

Fig. 3 Sentence dependency parsing results of example 1

再如句子“上海市食药监局发布最新一期食品安全抽检信息,通报了 5 批次不合格的食用性农产品。”的依存分析结果如图 4 所示。从图 4 中可以得到候选实体有“上海市食药监局”(机构名称)、“信息”和“农产品”;关系词为“发布”和“通报”,且在句子中是并列关系。“上海市食药监局”作为句子的主语分别通过“发布”和“通报”2 个关系词与作为句子宾语的“信息”和“农产品”连接,“5 批次”、“不合格”和“食用性农产品”之间依次存在定中关系。实体和关系词之间的依存关系符合模型 ORE_Model1、ORE_Model2 和 ORE_Model3 的分析,最后可抽取出实体关系三元组(上海市食药监局,发布,最新一期食品安全抽检信息)、(上海市食药监局,通报,5 批次不合格的食用性农产品)和(5 批次,不合格,食用性农产品)。

上述 2 个句子均是关于“上海市食药监局”相关的信息,基于实体关系抽取模型 ORE_Models 从不同的描述文本中抽取出了不同的实体关系三元组,这些三元组共同表述了同一主体的信息且不同三元组之间也存在关联关系。文本中一般包含较多数量的句子,能够抽取出大量的实体关系三元组。这些三元组高度概括了文本的主要内容且形式精炼,梳理后能帮助快速了解文本的知识脉络,得到目标信息。

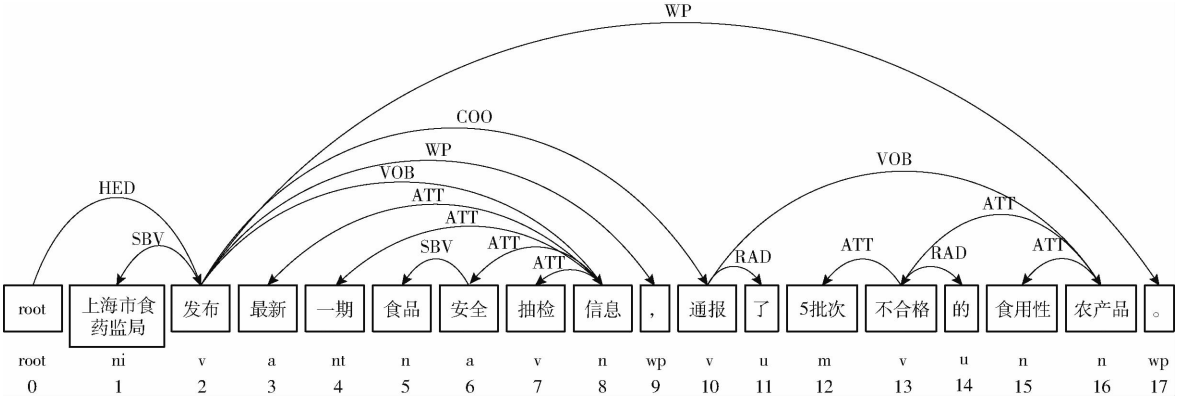

图 4 实例 2 的句子依存分析结果

Fig. 4 Sentence dependency parsing results of example 2

2 结果与分析

2.1 实验设计

实验所用的数据是利用爬虫技术爬取的近 5 年全国范围内各大新闻门户网站(包括腾讯新闻中心、搜狗新闻中心、百度新闻中心和新浪新闻中心等多个网站)上与食品相关的中文新闻文本,共 75 214 篇。这些中文新闻文本包含食品安全事件、与食品相关的非事件性新闻文本和其他领域的各类新闻文本,共同构成了新闻文本语料库,且不同类型文本的数量统计结果为:食品安全事件新闻文本 40 427 篇,与食品相关的非事件性新闻文本 31 086 篇,其他领域的各类新闻文本 3 701 篇。

(1)利用分类模型对语料库中的所有文本进行分类。虽然 PU 学习是在含少量标记的积极样例和大量未标记样例情况下训练分类器,但是为了与其他分类方法进行比较,仍需要额外做如下标记:手动标注了 1 000 篇食品安全事件新闻文本和 1 000 篇非食品安全事件的其他混合类型的新闻文本,将这 2 000 篇已标注的新闻文本作为数据集。随机抽取其中的 300 篇食品安全事件新闻文本和 300 篇非食品安全事件新闻文本共 600 篇文本作为测试集,其余的 1 400 篇文本作为训练集来训练分类器。在测试过程中,更多关注的是准确率,其计算公式为

$$p_c = \frac{N_r}{N_{\text{classifier}}} \times 100\% \tag{9}$$

式中　p_c——分类器的准确率

N_r——正确分类的文本数量

$N_{\text{classifier}}$——分类器分类的文本数量

(2)从分类得到的食品安全事件类别中随机抽取 1 000 篇文本,用于测试模型 ORE_Models 在食品安全事件新闻文本上的实体关系抽取性能。由于自然语言处理工具的处理对象是完整的句子,所以利用正则表达式方法[22]按照“。”、“?”、“!”、“……”、“:”、“;”6 种标点符号将这 1 000 篇文本分割成独立的句子。

(3)从分割 1 000 篇文本获得的句子中随机选择 1 000 个句子作为数据集 news_dataset1 进行实体关系抽取。注意,采用两次随机抽取,是为了在具有可操作性的数据量下降低新闻编辑者的语法习惯对抽取模型性能的影响,使结果具有更高的可靠性,从而更好的对食品安全事件进行实体关系抽取,有效地解决难以快速获取事件主要内容、脉络联系不明确等问题。

(4)再次从语料库中随机抽取 1 000 篇文本,这 1 000 篇文本中包含食品安全事件在内的多种混合类型的新闻。采用与得到数据集 news_dataset1 同样的方法得到包含 1 000 个句子的数据集 news_dataset2,该数据集用来评估模型 ORE_Models 对开放领域混合类型的新闻文本的抽取性能,从而验证模型 ORE_Models 的可移植性,使其能够应用于更多的研究领域。

在本实验中,由两名专业人员根据文献[12]的标注策略分别标注句子中的实体关系元组,然后经过汇总、纠正后,最终确定数据集应该被正确抽取的结果。本文的评估侧重于句子级别的抽取,实验后,将实验抽取结果与手动标注的结果进行比较,并通过 3 个度量标准对实体关系抽取结果进行评估,分别是准确率(P)、召回率(R)和 F 值(F)。P、R、F 的计算公式为

$$P = \frac{r}{a} \times 100\% \tag{10}$$

$$R = \frac{r}{W} \times 100\% \tag{11}$$

$$F = \frac{2PR}{P+R} \times 100\% \tag{12}$$

式中　r——模型 ORE_Models 抽取出的正确元组的数量

a——模型 ORE_Models 抽取出的所有元组的数量

W——语料库中实际存在的元组的数量

2.2　结果及对比分析

2.2.1　食品安全事件新闻文本的分类结果

为了验证 PU 学习方法的食品安全事件新闻文本的分类结果,首先只保留训练集中的 200 个标注的食品安全事件标签,其余数据的标签均隐藏(即相当于未标记数据)。然后在训练集中训练分类模型。最后,将得到的分类模型在测试集中进行测试,得到最终的分类结果。为了进行实验对比,在所有数据均保留了完整标注的相同数据集下,分别采用支持向量机(SVM)、逻辑回归算法(Logistic regression)、随机森林(Random forest)[27-28] 3 种监督分类方法进行训练,将得到的结果进行比较分析。实验结果为:本文的分类器准确率达到 82.35%,SVM 准确率为 75.94%,Logistic regression 准确率为 82.88%,Random forest 准确率为 83.49%。

上述结果显示 SVM 的准确率在 4 个分类器中是最低的,Random forest 分类器的准确率是最高的,但是仅比本文的分类器高出 1.14 个百分点。其次是 Logistic regression 分类器,比本文的分类器高出 0.53 个百分点。从这些数据中可以发现,本文构建的分类器准确率尽管不是最高的,但是达到了与其余 3 种监督方法相似的效果,相比于这 3 种监督方法需要完成的大量标注所需要的人力、时间的损耗,且在将大规模网络文本全部进行手动标注几乎不可能实现的前提下,半监督分类方法更能满足大规模数据分类研究的需要,并且降低了监督方法中由于人的主观因素引起的误差,因此更适合应用于大规模网络文本的食品安全事件的分类。

将本文的分类器应用于语料库,共得到了 37 901 篇食品安全事件新闻文本。

2.2.2　实体关系抽取的性能评估

从分类得到的 37 901 篇食品安全事件新闻文本随机抽取 1 000 篇文本并分割成句子后,共得到 24 015 个完整句子。再按照 2.1 节中描述的步骤构建数据集 news_dataset1 和 news_dataset2。

为了评估食品安全事件新闻文本的实体关系抽取结果和混合类型新闻文本的实体关系抽取结果的质量,得到 ORE_Models 抽取数据集 news_dataset1 和 news_dataset2 时的性能如表 5 所示。

表 5　ORE_Models 抽取不同数据集时的性能

Tab.5　Performance of ORE_Models when extracting different datasets　%

数据集	P	R	F
news_dataset1	80.59	63.78	71.21
news_dataset2	76.34	60.37	67.42

从表 5 可以看出,ORE_Models 模型的准确率相对较高,很难有更大的改进余地,但是获得高准确率的同时牺牲了部分召回率,使得召回率没有达到与准确率接近的性能。

news_dataset1 和 news_dataset2 数据集上的抽取性能相比,ORE_Models 模型在食品安全事件新闻文本数据集 news_dataset1 上的准确率、召回率、F 值均高于混合类型新闻文本数据集 news_dataset2 上的值,这说明 ORE_Models 更适用于食品安全事件新闻文本的实体关系抽取。但是在混合类型的新闻文本上的抽取性能也达到了较高的水平,与在食品安全事件新闻文本相比仅在准确率上降低了4.25 个百分点,召回率上降低了 3.41 个百分点,F 值上降低了 3.79 个百分点,与食品安全事件新闻文本的抽取效果之间的差距控制在了 5 个百分点之内,均未出现较大差异,表明了 ORE_Models 也可以应用于开放领域的新闻文本抽取。

2.2.3 实体关系抽取的性能对比

为了验证模型 ORE_Models 的性能能够满足新闻文本关系抽取的需要,设计 2 组对比实验:①ZORE 系统、CORE 系统与 ORE_Models 同时处理数据集 news_dataset1。②ZORE 系统、CORE 系统与 ORE_Models 同时处理数据集 news_dataset2。2 组实验的评估均对照同一标准结果进行判定。2 组实验结果如表 6 所示。

表 6 ZORE 系统、CORE 系统抽取 news_dataset1 和 news_dataset2 的性能

Tab. 6 Performance of ZORE system and CORE system to extract news_dataset1 and news_dataset2

%

方法	数据集	*P*	*R*	*F*
ZORE 系统	news_dataset1	58.37	49.96	53.84
	news_dataset2	62.13	53.85	57.69
CORE 系统	news_dataset1	50.66	41.79	45.80
	news_dataset2	52.41	40.89	45.94

从表 5 和表 6 中可以看到,在数据集 news_dataset1 和 news_dataset2 上 CORE 系统的准确率、召回率和 F 值均是最低的,其次是 ZORE 系统,各个性能最好的是 ORE_Models。在 news_dataset1 数据集上,ZORE 系统和 CORE 系统的各个指标均表现出了类似的性能,几乎没有差异,这说明这 2 个系统都未对食品安全事件进行更加深入的抽取研究。虽然 ZORE 系统和 CORE 系统面向的是开放领域各类别的实体关系抽取,但是在 news_dataset2 数据集上,它们的性能仍低于 ORE_Models,这表明 ORE_Models 虽然主要面向食品安全事件新闻文本,但是它同样可以很好地处理开放领域的文本,体现了 ORE_Models 的有效性与可移植性。

对于抽取过程中出现的抽取错误问题或者未抽取出句子中存在的元组问题,主要是由以下几方面引起的:NLP 工具在分词、词性标注或者命名实体识别等过程中出现错误,存在未覆盖的领域导致无法正确处理句子,不能与模型匹配或匹配错误;新闻文本中存在复杂度很高或者口语化、不规范的句子,该类句子的依存解析在模型中未涉及到。

3 结束语

提出一种基于依存分析的食品安全事件新闻文本的实体关系抽取方法 FSE_ERE,根据中文语法特性和句子的依存分析结果构建了关系抽取模型,实现了无监督的食品安全事件新闻文本的实体关系抽取。为了在高质量的食品安全事件新闻文本上进行抽取工作,引入结合文本相似度算法和改进的特征加权方法的 PU 学习半监督分类方法,对大规模网络文本进行分类,准确率达到 82.35%。FSE_ERE 方法能够从大规模的网络文本中准确得到食品安全事件类别的新闻文本,且无需标记大量数据的类别;同时,实体关系抽取过程也打破了标注语料库、预先定义关系类型等限制,可快速准确地抽取出文本中包含的各种信息,在食品安全事件新闻文本数据集上 F 值达到 71.21%,在多类型混合新闻文本数据集上 F 值达到 67.42%。FSE_ERE 方法节省了大量的人力和时间,对于大规模网络文本的信息统计分析具有重要意义,为中文的开放式实体关系抽取提供了新的思路。

参 考 文 献

[1] 江美辉, 安海忠, 高湘昀, 等. 基于复杂网络的食品安全事件新闻文本可视化及分析[J]. 情报杂志, 2015, 34(12): 121 - 127.
JIANG Meihui, An Haizhong, GAO Xiangyun, et al. The visualization and analysis of news texts about food safety incidents based on complex networks[J]. Journal of Intelligence, 2015, 34(12): 121 - 127. (in Chinese)

[2] 陈瑛, 陈昂轩, 董玉博, 等. 基于 LSTM 的食品安全自动问答系统方法研究[J/OL]. 农业机械学报, 2019, 50(增刊): 380 - 384.
CHEN Ying, CHEN Angxuan, DONG Yubo, et al. Methods of food safety question answering system based on LSTM[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(Supp.): 380 - 384. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no=2019s058&flag=1&journal_id=jcsam. DOI: 10.6041/j.issn.1000-1298.2019.S0.058. (in Chinese)

[3] 李明耀，杨静. 基于依存分析的开放式中文实体关系抽取方法[J]. 计算机工程，2016，42(6)：201－207.
LI Mingyao，YANG Jing. Open Chinese entity relation extraction method based on dependency parsing [J]. Computer Engineering，2016，42(6)：201－207.（in Chinese）

[4] XU Jing，GAN Liang，DENG Lu，et al. Dependency parsing based Chinese open relation extraction [C] // 2015 4th International Conference on Computer Science and Network Technology (ICCSNT). IEEE，2015：552－556.

[5] KOZAREVA Z，HOVY E. Not all seeds are equal：measuring the quality of text mining seeds [C] // Human Language Technologies：The Conference of the North American Chapter of the Association for Computational Linguistics，2010：618－626.

[6] MINTZ M，BILLS S，SNOW R，et al. Distant supervision for relation extraction without labeled data[C] // Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP，2009：1003－1011.

[7] 张硕望. 一种基于远程监督的中文实体关系抽取方法[D]. 衡阳：南华大学，2018.

[8] 秦兵，刘安安，刘挺. 无指导的中文开放式实体关系抽取[J]. 计算机研究与发展，2015，52(5)：1029－1035.
QIN Bing，LIU Anan，LIU Ting. Unsupervised Chinese open entity relation extraction[J]. Journal of Computer Research and Development，2015，52(5)：1029－1035.（in Chinese）

[9] 陈志泊，李钰曼，许福，等. 基于 TextRank 和簇过滤的林业文本关键信息抽取研究[J/OL]. 农业机械学报，2020，51(5)：207－214，172.
CHEN Zhibo，LI Yuman，XU Fu，et al. Key information extraction of forestry text based on TextRank and clusters filtering[J/OL]. Transactions of the Chinese Society for Agricultural Machinery，2020，51(5)：207－214，172. http：//www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? flag=1&file_no=20200523&journal_id=jcsam. DOI：10.6041/j.issn.1000-1298.2020.05.023.（in Chinese）

[10] BANKO M，CAFARELLA M J，SODERLAND S，et al. Open information extraction from the web[C] // International Joint Conference on Artifical Intelligence. Morgan Kaufmann Publishers Inc.，2007：68－74.

[11] BANKO M，ETAIONI O. The tradeoffs between open and traditional relation extraction[C] // Proceedings of ACL-HLT，2008：28－36.

[12] FADER A，SODERLAND S，ETAIONI O. Identifying relations for open information extraction[C] // Proceedings of EMNLP，2011：1535－1545.

[13] SCHMITZ M，BART R，SODERLAND S，et al. Open language learning for information extraction[C] // Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning，2012：523－534.

[14] QIU Likun，ZHANG Yue. ZORE：a syntax-based system for Chinese open relation extraction[C] // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)，2014：1870－1880.

[15] TSENG Y H，LEE L H，LIN S Y，et al. Chinese open relation extraction for knowledge acquisition[C] // Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics，2014：12－16.

[16] WANG Yue，ZHOU Gang，TIAN Fei，et al. GCORE：a gravitation-based approach for Chinese open relation [C] // International Conference on Computer Science & Mechanical Automation. IEEE，2016：86－91.

[17] WU Xiaoyang，WU Bin. The CRFs-based Chinese open entity relation extraction[C] // Proceedings of 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC)，2017：405－411.

[18] CHE Wanxiang，LI Zhenghua，LIU Ting. Ltp：a Chinese language technology platform [C] // Proceedings of the 23rd International Conference on Computational Linguistics：Demonstrations. Association for Computational Linguistics，2010：13－16.

[19] BEKKER J，DAVIS J. Learning from positive and unlabeled data：a survey[J]. arXiv preprint arXiv：1811.04820，2018：1－42.

[20] 任亚峰，姬东鸿，张红斌，等. 基于 PU 学习算法的虚假评论识别研究[J]. 计算机研究与发展，2015，52(3)：639－648.
REN Yafeng，JI Donghong，ZHANG Hongbin，et al. Deceptive reviews detection based on positive and unlabeled learning [J]. Journal of Computer Research and Development，2015，52(3)：639－648.（in Chinese）

[21] 蒋健. 文本分类中特征提取和特征加权方法研究[D]. 重庆：重庆大学，2010.

[22] 王露瑶，张涛，陈才，等. 基于卡方统计改进的 TF－IDF 的文本分类的研究[J]. 电子世界，2019(6)：24－25.

[23] 周源，刘怀兰，杜朋朋，等. 基于改进 TF－IDF 特征提取的文本分类模型研究[J]. 情报科学，2017，35(5)：111－118.

[24] 李俊峰. 多特征融合的新闻聚类相似度计算方法[J]. 软件，2017，38(12)：170－174.

[25] WANG Y，YANG Y，ZHAO R. The Chinese open relation extraction based on dependency parsing[C] // International Conference on Frontiers of Manufacturing Science & Measuring Technology，2017：552－556.

[26] JIA Shengbin，SHIJIA E，LI Maozhen，et al. Chinese open relation extraction and knowledge base establishment[J]. ACM Transactions on Asian and Low-Resource Language Information Processing，2018，17(3)：1－22.

[27] 何清，李宁，罗文娟，等. 大数据下的机器学习算法综述[J]. 模式识别与人工智能，2014，27(4)：327－336.
HE Qing，LI Ning，LUO Wenjuan，et al. A survey of machine learning algorithms for big data[J]. Pattern Recognition and Artificial Intelligence，2014，27(4)：327－336.（in Chinese）

[28] MITCHELL T M，CARBONELL J G，MICHALSKI R S. Machine learning[M]. USA：McGraw-Hill，2003.