DOI:10.20033/j.1003-7241.(2026)04-0169-05

基于规则模式的科技情报关键信息自动抽取算法

田春梅

(云南省临沧市科学技术情报研究所,云南 临沧 677000)

摘　要:非结构化、多模态的科技文本已成为情报分析与知识发现的主要挑战,科技情报以海量非结构化文本形式涌现,带来了信息过载、关键信息难以快速定位等问题,导致信息理解困难,利用率低下。为此,研究一种基于规则模式的科技情报关键信息自动抽取算法。该算法对原始科技情报文本进行分词与停用词去除等预处理,以规范文本结构、降低噪声;通过基于特征权重的关联规则挖掘方法,从预处理后的文本中提取出反映关键信息结构的规则模式,构建规则库;设计规则匹配机制,将待抽取文本与规则库进行多轮匹配,实现对科技情报中关键信息的自动抽取。实验选取多领域科技情报简报作为测试数据,结果显示,所研究算法的折扣累计增益更大,表明其在关键信息抽取的准确性与相关性方面表现更优,验证了该算法在提升科技情报分析效率与自动化水平方面的有效性与实用性。

关键词:数据挖掘;规则模式;科技情报;关键信息;自动抽取算法

中图分类号:TP391.3　　**文献标志码**:A　　**文章编号**:1003-7241(2026)04-0169-05

Automatic extraction algorithm for key information in scientific and technological intelligence based on rule pattern

TIAN Chunmei

(Yunnan Lincang Institute of Science and Technology Information, Lincang 677000, Yunnan, China)

Abstract: Unstructured and multimodal scientific and technological texts have become the main challenge of information analysis and knowledge discovery. The emergence of scientific and technological intelligence in the form of massive unstructured texts has brought about problems such as information overload and difficulty in locating key information quickly, which has led to difficulties in understanding information and low utilization rate. Therefore, this paper studies an automatic extraction algorithm of key information of scientific and technological intelligence based on rule pattern. The algorithm preprocesses the original sci-tech information text such as word segmentation and stop words removal, so as to standardize the text structure and reduce noise. Through the mining method of association rules based on feature weight, the rule pattern reflecting the key information structure is extracted from the preprocessed text, and the rule base is constructed. A rule matching mechanism is designed to match the text to be extracted with the rule base for multiple rounds, so as to realize the automatic extraction of key information in scientific and technological intelligence. The experiment selects multi-domain scientific and technological information briefings as test data, and the results show that the discount cumulative gain of the studied algorithm is larger, which shows that it performs better in the accuracy and correlation of key information extraction, and verifies the effectiveness and practicability of the algorithm in improving the efficiency and automation level of scientific and technological information analysis.

Keywords: data mining; rule pattern; scientific and technological intelligence; key information; automatic extraction algorithm

在信息爆炸的时代,科技情报往往以海量的文本数据形式存在,这些文本数据中潜藏着对技术发展、行业变革乃至全球经济趋势至关重要的关键信息。因此,如何从这些庞杂的数据中准确、高效地提取出有价值的关键信息,成为了科技情报领域亟待解决的问题[1]。科技情报中的关键信息,如技术趋势的预测、最新研究成果的发布、行业动态的实时更新等,对于科研工作者、企业决策者乃至政策制定者而言,都具有极其重要的参考价值。因此,开发一种能够自动、准确地从科技情报文本中抽取关键信息的算法,具有极高的实际应用价值和深远的战略意义。

Dellanzo A 等提出一种基于深度学习的关键信息自动抽取方法,首先将输入的文本转换为向量或矩阵形式的数值表示,然后使用标注好的数据(通常是关键信息的标记位置或类别)对深度学习模型进行训练,最后利用训练好的模型对新的、未见过的文本进行关键信息抽取[2]。深度学习模型通常基于已知的数据和模式进行训练,对于新出现的事件或未知的领域可能缺乏足够的泛化能力,这导致模型的抽取结果受到限制。陈勇等在研究中,BERT 层用于将输入的情报文本转化为词向量,并为每个词提供上下文表示,利用 BiLSTM 层对 BERT 层输出的词向量进行

收稿日期:2024-06-21;**录用日期**:2024-07-11

基金项目:云南省科技厅科技计划项目(202104AC100001-A15)

作者简介:田春梅(1976—),女,副研究员,研究方向:科技情报研究,科技信息咨询,科技项目管理与服务。

引用本文:田春梅. 基于规则模式的科技情报关键信息自动抽取算法[J]. 自动化技术与应用, 2026,45(4):169-172, 181. (TIAN Chunmei. Automatic extraction algorithm for key information in scientific and technological intelligence based on rule pattern[J]. Techniques of Automation and Applications, 2026,45(4):169-172, 181.)

进一步编码,以获取每个词的上下文表示,CRF 层用于对 BiLSTM 层输出的特征进行分类,为每个词分配一个实体标签,从而实现关键信息的抽取[3]。对于过长的序列,BiLSTM 模型存在信息丢失的问题,导致在抽取关键信息时,对于跨越多个句子的复杂关系处理不够准确。赵玉媛等通过一种隐私数据结构知识增强机制,将结构先验知识融入隐私信息抽取过程中,这种机制能够加强模型对句子语义结构的理解,从而更加准确地识别隐私信息的边界,然而结构先验知识需要不断更新和完善以适应新的隐私数据类型和场景[4]。然而,知识更新的速度和准确性可能受到限制,导致算法在某些情况下表现不佳。许鑫等利用 BERT 模型获取文本中词语的语义表示,为样本增加扰动,捕捉文本中的时序依赖关系,确保输出标签序列的合法性,从而提高信息抽取的准确率[5]。对于不同的信息抽取任务,可能需要调整模型的结构和参数,以适应不同的应用场景。

科技情报中的关键信息往往隐藏在复杂的文本结构中,这使得传统算法在科技情报关键信息抽取时,处理和分析能力较弱。为此,研究一种基于规则模式的科技情报关键信息自动抽取算法。通过本研究以期能够更准确地从复杂的文本结构中抽取科技情报中的关键信息,减少误报和漏报。

1 科技情报关键信息自动抽取

在竞争激烈的科技领域,及时获取并分析最新的科技情报对于保持竞争优势至关重要。通过自动化抽取技术,可以实现对科技情报的实时监控,一旦有新的关键信息出现,便能立即进行抽取和分析,为科技创新和决策支持提供及时、准确的信息支持[6]。为此,研究一种基于规则模式的科技情报关键信息自动抽取算法。该研究主要分为三部分,即科技情报文本预处理、规则模式挖掘、关键信息自动抽取。下面针对这 3 个步骤进行具体分析。

1.1 科技情报文本预处理

科技情报文本预处理不仅是整个流程的第一步,更是为后续关键信息抽取提供高质量数据基础的关键环节。在科技情报领域,信息的获取、整合和分析对于企业和研究机构至关重要。然而,原始的科技情报文本往往包含大量 3 噪声数据、复杂的文本结构和多样的文本格式,这使得直接从中提取关键信息变得困难[7]。因此,文本预处理成为了一个不可或缺的前置步骤。

1) 科技情报分词

在科技情报文本中,一个句子通常是由连续的字符序列组成,这些字符序列按照特定的语言规则和语法结构组合成有意义的词汇单元。分词的目标就是将这些连续的字符序列切分成独立的词汇单元,以便于后续的文本分析和处理[8]。基于词典的分词方法基本思想是将待分析的文本与一个预先构建的词典中的词条进行匹配,如果词典中存在某个字符串,则将其作为一个词切分出来。

2) 科技情报停用词去除

停用词,如“的”“是”“在”“了”“和”等中文常用词,以及英文中的“the”“a”“an”“in”“on”等冠词、连词和介词。这些词汇在句子中起到了语法或结构上的作用,但在文本分析和挖掘时,它们往往不携带太多有意义的信息[9]。停用词去除的主要目的是减少文本的数据量,同时提高后续分析的效率和准确性。当从文本中提取关键信息、进行主题建模、情感分析或文本分类等任务时,停用词的存在可能会干扰算法的判断,因为它们几乎不携带与文本内容相关的有价值信息。因此,去除停用词可以简化文本数据,使算法更容易聚焦于那些真正有意义的词汇,从而提高分析的效率和准确性。停用词去除的实现方法相对简单,通常包括以下几个步骤:首先需要准备一个包含科技情报常见停用词的列表,将分词后的科技情报文本与停用词表进行匹配,如果某个词汇在停用词表中出现,则将其从文本中删除。去除停用词后,得到的科技情报文本将只包含那些有意义的词汇,可以进一步用于后续的分析和挖掘任务[10]。

这些步骤能够有效地去除文本中的噪声数据,将文本拆分成更小的词汇单元,并为每个词汇单元标注词性,从而便于后续的分析和理解[11]。最后,预处理还能够使文本数据更加符合算法的输入要求,从而提高算法的通用性和可扩展性。

1.2 规则模式挖掘

在科技情报关键信息自动抽取算法中,规则模式挖掘是一个重要的步骤。规则模式挖掘是指从数据中提取出具有普遍性和规律性的模式,这些模式可以用规则的形式进行描述[12]。在科技情报关键信息自动抽取中,规则模式挖掘主要用于发现文本数据中隐藏的关键信息结构或规律,以便更准确地提取关键信息。基于特征权重的关联规则挖掘算法是一种结合了特征权重信息来改进关联规则生成的方法。这类算法不仅考虑了项集在数据集中的出现频率(即支持度),还考虑了项集中每个特征的重要性(即权重),从而能够更准确地挖掘出有意义的关联规则。以下是关于基于特征权重的规则挖掘的步骤。

步骤 1 设置支持度和置信度阈值。

步骤 2 确定待抽取的科技情报参考样本并按照章节 1.1 进行预处理。

步骤 3 计算每个科技情报参考样本中每个词的特征权重,即

$$Q_i = A_i \times B_i \tag{1}$$

其中,

$$A_i = \frac{N_{ik}}{\sum_{k=1}^{K} n_{ik}} \tag{2}$$

$$B_i = \log \frac{M}{M_i + 1} \tag{3}$$

式中,M 为科技情报文本总数;M_i 代表科技情报文本集中包含第 i 个词语的文本数目;n_{ik} 代表第 i 个词语在科技情报文本 k 中出现的次数;Q_i 代表参考样本第 i 个词语的词频-逆文档频率值;A_i 代表第 i 个词语的词频;B_i 代表第 i 个词语的逆文档频率。

步骤 4 特征权值归一化。

$$w_i = \frac{Q_i}{\sqrt{(Q_i)^2}} \tag{4}$$

式中，w_i 代表第 i 个词语的特征权重。

步骤 5　根据 A_i，计算支持度 C_i。

$$C_i = \frac{A_i}{\sum_{k=1}^{K} n_{ik}} \tag{5}$$

步骤 6　对比 C_i，找出所有大于等于最小支持度的词语，得出样本的最大频繁 1–项关键词集 L_1。

步骤 7　将 1–项集 L_1 按照特征权重和支持度的组合进行排序。

步骤 8　从排序后的 1–项集中选择前 N 个项目，并使用它们来生成 2–项关键词集 L_2。

步骤 9　遍历样本集，计算每个 2–项关键词集的支持度。

步骤 10　重复上述过程，删除支持度低于阈值的候选项关键词集，最终得到具有关联性的候选项关键词集 L_k。

步骤 11　对于每个候选项关键词集（假设它是一个 k – 项集），提取出多个（$k-1$）项集作为前件，并将剩余的一个项目作为后件，从而生成多个可能的关联规则。对于候选项关键词集 $\{A, B, C\}$，可以提取出规则 $A, B \Rightarrow C$；$A, C \Rightarrow B$；$B, C \Rightarrow A$。

步骤 12　对于每个关联规则，计算其置信度 D。

$$D = \frac{N(\alpha \cup \beta)}{N(\alpha)} \tag{6}$$

式中，$N(\alpha \cup \beta)$ 代表关键词项集 α 和 β 同时出现在一个样本中的频率；$N(\alpha)$ 代表关键词项集 α 出现在一个样本中的频率。

步骤 13　将计算出的置信度 D 与阈值进行比较。如果置信度高于该阈值，则认为该规则是重要的或“强”的关联规则。否则，可以将其视为不重要的或“弱”的规则。

步骤 14　将删选出来的规则整理成规则模式库，便于后续的信息抽取过程。

规则模式挖掘能够发现科技情报中的有趣、隐藏和非平凡的关联规则，这些规则能够揭示数据背后的潜在模式和规律。通过应用这些规则，能够更准确地识别并抽取科技情报中的关键信息，从而提高信息抽取的准确性和效率[13-14]。

1.3　规则匹配抽取关键信息

在科技情报领域，信息的准确性和时效性对于决策者、研究人员和企业来说至关重要。因此，能够自动、高效地从海量的科技文献、专利、研究报告等文本中抽取关键信息，成为了科技情报分析的重要一环。在这一章节中，将探讨如何利用先前挖掘出来的规则，对科技情报中的关键信息进行自动抽取[15]。需要将挖掘和制定的规则进行整合和分类，以便在自动抽取过程中能够有效地利用。这些规则可能包括基于关键词、短语、句法结构、语义关系等不同层面的规则，它们共同构成了科技情报关键信息抽取的规则库[16]，具体过程如下。

步骤 1　输入待抽取的科技情报测试文本。

步骤 2　按照章节 1.1 研究处理待抽取的科技情报文本。

步骤 3　用规则模式库中第 i 条规则与样本中每个句子进行匹配，匹配公式如下：

$$H_i = \sqrt{\sum_{j=1}^{m} (x_i - r_{ij})^2} \tag{7}$$

式中，H_i 代表匹配度；x_i 代表第 i 条规则；r_{ij} 代表第 i 个句子第 j 个属性。

步骤 4　根据匹配结果，按照下述公式判断是否匹配成功？若成功，保存该句子；否则，退出该条规则，进行下一步。

$$\begin{cases} \text{匹配成功}, H_i \geqslant \gamma \\ \text{匹配未成功}, H_i < \gamma \end{cases} \tag{8}$$

式中，γ 代表阈值。

步骤 5　利用第 $i+1$ 条规则再次与样本中每个句子进行匹配，保存匹配成功的句子。

步骤 6　重复上述过程，遍历规则模式库所有规则。

步骤 7　将所有规则匹配出来的结果组合在一起，组成抽取出来的关键信息列表。

经过上述过程，完成科技情报关键信息抽取。

2　实验分析

2.1　测试文本

科技情报是指通过公开信息渠道获取的、具有知识性、传递性和效用性的、关于科学发展、技术创新和最新动态的有用知识。这些情报信息是科技工作的基础，也是重要组成部分。实验分析中摘取三份科技情报简报中内容作为所研究算法的测试文本，如图 1 所示。

喷气燃料脱色精制中颗粒白土失活加快的原因分析

针对大连石化分公司喷气燃料脱色精制中出现的颗粒白土失活快的问题，着重剖析比较了生产中采用的 3 种不同颗粒白土的结构性质包括比表面积、孔径分布和酸性质等，并用喷气燃料馏分油对比了不同白土的脱色性能，发现不同生产厂的白土自身性质和相应的喷气燃料馏分油的脱色性能差异很大，具有高比表面积、孔体积和最大酸量的 A 厂白土表现出最好的脱色性能。颗粒白土的本身性质决定脱色性能。不同性质的白土所表现出的不同脱色效果可能是引起颗粒白土脱色不稳定的因素之一。

（a）文本 A

柴油逆流加氢超深度脱硫芳烃技术的研究和开发

介绍了柴油逆流加氢超深度脱硫芳烃技术。气液逆流操作可克服常规并流工艺的劣势，提高氢分压并降低液相中 H_2S 浓度，可以获得更高的脱硫、脱芳烃程度。试验结果表明，无论是单级反应器还是二级反应器串联，都体现了逆流操作的优势，采用一级并流，二级逆流串联操作方式，可生产硫含量低于 10μg/g 的超低硫清洁柴油。所开发的催化剂具有三级孔道结构，可防止液泛，在逆流反应器的循环氢中 H_2S 浓度很低的情况下仍能保持较高的活性和稳定性。

（b）文本 B

用动态剪切流变试验评价硫磺改性沥青的高温性能

对 2 种基质沥青及其产品进行动态剪切流变学分析，并分别对硫化改性前后沥青的高温性能进行对比。结果表明，相对于传统环球软化点方法，动态剪切流变学评价能够较好地反映沥青的高温性能，并且能够较好地反映出沥青硫化前后抗车辙性能和弹性性能的变化。沥青硫化改性后，弹性性能变好，抗车辙能力增强，使用寿命延长。

（c）文本 C

图 1　科技情报测试文本

Fig. 1　Test text of scientific and technological intelligence

图 1 中这 3 份科技情报来源不同，所介绍的科技信息不同，为方法验证提供了多样化测试数据。

2.2 规则挖掘测试

在文本预处理的基础上,基于章节1.2研究,挖掘规则模式,以文本C为例,挖掘结果如表1所示。

表1关联规则表列出了硫磺改性沥青与其性能评价之间的主要关联点,为后续关键信息抽取提供了重要参考。

表1 文本C的规则模式

Tab. 1 Rule pattern of text C

规则编号	规则前项(A)	规则后项(B)	解读
规则1	硫磺改性沥青(A1)	动态剪切流变学评价(B1)	硫磺改性沥青的高温性能可通过动态剪切流变学评价进行评估
规则2	硫磺改性沥青(A1)	弹性性能变好(B2)	硫磺改性后,沥青的弹性性能得到了提升
规则3	硫磺改性沥青(A1)	抗车辙能力增强(B3)	硫磺改性沥青具有更强的抗车辙能力
规则4	硫磺改性沥青(A1)	使用寿命延长(B4)	硫磺改性沥青的使用寿命相对于未改性沥青有所延长
规则5	动态剪切流变学评价(B1)	较好地反映高温性能(C1)	动态剪切流变学评价能够较好地反映沥青的高温性能
规则6	动态剪切流变学评价(B1)	反映硫化前后性能变化(C2)	动态剪切流变学评价能够反映沥青硫化前后抗车辙性能和弹性性能的变化

2.3 关键信息抽取结果

基于表1挖掘出来的规则,进行科技情报关键信息自动抽取,抽取结果如表2所示。基于表2抽取结果,证明了所研究算法的有效性,完成了关键信息抽取任务。

表2 关键信息抽取结果

Tab. 2 Key information extraction results

文本	关键信息类别	抽取结果
A	公司名称	大连石化分公司
	问题描述	喷气燃料脱色精制中颗粒白土失活快
	分析对象	3种不同颗粒白土
	结构性质	比表面积、孔径分布、酸性质
	脱色性能对比	A厂白土表现出最好的脱色性能
	性能差异因素	不同生产厂的白土自身性质
	最佳性能特点	高比表面积、孔体积和最大酸量
	脱色性能决定因素	颗粒白土的本身性质
	脱色不稳定因素	不同性质的白土所表现出的不同脱色效果
B	技术名称	柴油逆流加氢超深度脱硫芳烃技术
	技术特点	气液逆流操作
	优点	提高氢分压、降低液相中H_2S浓度、获得更高的脱硫脱芳烃程度
	试验结果	可生产硫含量低于10 μg/g的超低硫清洁柴油
	反应器类型	单级反应器、二级反应器串联
	操作方式	一级并流,二级逆流串联
	催化剂特性	三级孔道结构、防止液泛、在H_2S浓度低时保持高活性和稳定性
C	沥青类型	2种基质沥青及其产品
	改性方式	硫化改性
	评估方法	动态剪切流变学分析
	对比方法	传统环球软化点方法
	评估性能	高温性能、抗车辙性能、弹性性能
	改性效果	弹性性能变好、抗车辙能力增强、使用寿命延长

2.4 对比分析

以“折扣累计增益”作为衡量抽取算法性能的指标。计算公式为

$$Z = \sum_{i=1}^{M} \frac{2^{u_i}}{\log_2(u_i)} \tag{9}$$

式中,u_i代表关键信息列表位置上第i个信息的信息熵;M代表信息数量;Z代表折扣累计增益。Z值越高,说明抽取关键信息与文本的相关度越高,抽取结果越准确。将验证指标与3种传统方法进行对比,结果如图2所示。

从图2中可以看出,所研究算法的折扣累计增益大于3种传统方法,说明所研究算法的抽取结果更为准确,方法的抽取性能更好。

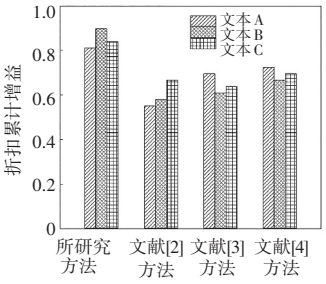

图2 折扣累计益对比图

Fig. 2 Comparison chart of discount cumulative gain

3 结论

通过信息抽取技术,可以从海量的科技情报中快速提取出有价值的信息,为科研、产业发展和政策制定提供有

(下转第181页)