**Software Engineering and Applications** 软件工程与应用, 2022, 11(6), 1554-1560
Published Online December 2022 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2022.116160

Hans汉斯

# 电解加工知识本体中领域术语提取的研究与应用

**季　峰[1]，黄　萍[1]，Ali Abdullahi Moallim[2*]**

[1]南通大学杏林学院，江苏　南通
[2]南通大学机械工程学院，江苏　南通

收稿日期：2022年11月30日；录用日期：2022年12月23日；发布日期：2022年12月31日

## 摘　要

**为提高电解加工工艺知识本体中的概念提取的完整性，本文中构建了一种半自动化领域术语提取模型，该模型结合统计分析和数据挖掘的思想设计了N-Word算法，进行领域术语中词组的提取，3-Word构词性能最佳。为了提高领域术语的准确性，基于互信息(*MI*)和绝对词频对领域术语过滤得到2137个术语，进一步对术语修正和同义词合并处理，最终得到标准化的领域概念1894个。此模型满足对电解加工领域术语的提取，提高术语的领域覆盖度，保证本体构建的准确性。**

## 关键词

**领域术语，提取模型，N-Word算法，互信息，本体**

# Research and Application of Domain Term Extraction in Electrochemical Machining Knowledge Ontology

**Feng Ji[1], Ping Huang[1], Ali Abdullahi Moallim[2*]**

[1]School of Xinglin, Nantong University, Nantong Jiangsu
[2]School of Mechanical Engineering, Nantong University, Nantong Jiangsu

Received: Nov. 30th, 2022; accepted: Dec. 23rd, 2022; published: Dec. 31st, 2022

## Abstract

**In order to improve the integrity of concept extraction in ECM process knowledge ontology, this**

*通讯作者。

文章引用: 季峰, 黄萍, Ali Abdullahi Moallim. 电解加工知识本体中领域术语提取的研究与应用[J]. 软件工程与应用, 2022, 11(6): 1554-1560. DOI: 10.12677/sea.2022.116160

**paper constructs a semi-automatic domain term extraction model, which combines the idea of statistical analysis and data mining to design N-Word algorithm to extract phrases in domain terms. 3-Word has the best word formation performance. In order to improve the accuracy of domain terms, 2137 terms were filtered based on mutual information (*MI*) and absolute word frequency, and 1894 standardized domain concepts were finally obtained through further term modification and synonym merging. This model can extract terms in the field of electrochemical machining, improve the domain coverage of terms, and ensure the accuracy of ontology construction.**

## Keywords

**Domain Terminology, Extraction Model, N-Word Algorithm, Mutual Information, Ontology**



## 1. 引言

领域术语的提取是领域知识本体构建的基础，主要有规则法和统计法。规则法基于大量的文本分析的规则，文献[1] [2]中采用机器学习等方法提取术语，需要对大量的文本进行语义分析，基于足够数量的规则对术语进行提取。文献[3] [4]中统计法基于词性的统计特征对领域术语进行提取，不依赖于语义。电解加工工艺领域术语时，人工选取比例高，导致准确性和完备性程度不高。尤其是在对领域术语中的词组提取时，受到主观性干扰，造成领域核心术语的质量比较低，且术语提取不全面，本研究提出了一种半自动的领域术语提取模型，该模型结合统计分析和数据挖掘的思想，结合词频和术语所属的文本数进行筛选，采用 N-Word 算法进行构词后，基于互信息 *MI* 进行筛选、细化，得到领域本体的最终概念集合。

## 2. 领域术语资料

电解加工领域资料形式比较多，包括加工设备中存储的元数据，工件设计的图纸数据，加工过程产生的图片数据，各类传感器采集得到的数据。不同研究选择的数据各不相同，例如，加工缺陷检测的研究使用图片数据，电化学加工设备的故障检测中，收集传感器采集的数据。

本文选取电解加工领域的文本资料，不包括图片、音频和视频等格式的资源。研究中选取的领域资料是非结构化数据，分为中文资料和英文资料两种。在本体的半自动化构建中，选取的数据是英文资料，对英文资料提取领域术语，中文资料单独进行术语抽取，并对两者的结果进行对照和补充，通过领域研究人员对两种方式得到的领域术语进行互补，最终确定领域术语，作为本体概念的基础。

领域信息材料的选取直接决定电化学加工领域的概念质量，以科学性、标准性、权威性和学科理论性等原则，进行信息材料的选择。选取的信息材料主要包括：2010~2021 年电解加工领域发表的核心期刊文献和《电化学加工技术》英文版书籍。

## 3. 领域术语的提取模型

领域术语的提取是电解加工工艺领域本体构建的前提，目前，领域术语常用的有两种：一是基于语言学的提取方法，二是基于统计的方法。

本文选用基于统计的方法，利用电解领域中特定词或词组的统计特征来抽取领域术语，相比于基于语言学的方法更适合电解加工领域，不受语法和语义的影响，易于实现和扩展。构建了一种半自动领域

术语的构建模型。如图 1 所示，模型首先对领域文本进行预处理，对领域词汇或词组进行词频计算，进行同义词消除等操作，得到英文的领域概念。再结合中文文献里提取的关键词，两部分进行比对，结合 wordnet 和领域专家的建议，进行修正和补充，最终得到领域概念。

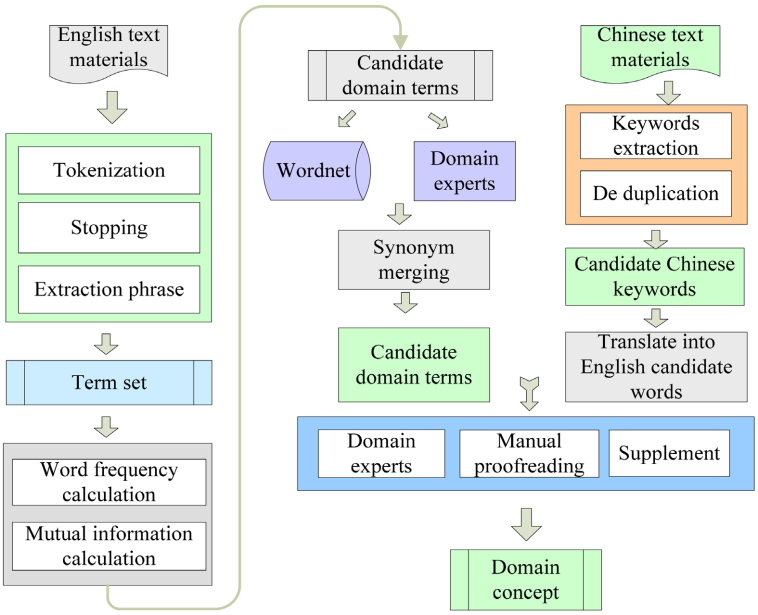


**Figure 1.** Domain term extraction model
**图 1.** 领域术语提取模型

## 4. 基于 N-Word 法的领域术语提取

在电解加工工艺本体中，不仅存在单词类型的术语，还存在词组形式的术语，而且短语往往比单词更能反映领域特征。经过文本分词初步切分后只能形成一个个单词，因此需要对这些单词进行处理形成有效的词组[5] [6]。

基于规则的提取方式，由于规则的约束性大，会漏掉领域词汇，因此，本研究利用 N-Word 算法进行词组概念提取。N-Word 能够提取从 1 词到 N 词的所有组合，不需要词典和规则的支撑，但是会产生一些噪音词组。领域中以 2 词型、3 词型和 4 词型词组居多，本研究中只计算 2-Word、3-Word 和 Word。

**Table 1.** Partial phrases extracted based on 2-Word
**表 1.** 基于 2-Word 提取的部分词组

| 2-Word | 提取的词组 |
|---|---|
| Electrochemical/machining/ | Electrochemical machining |
| machining/process/ | machining process |
| process/engineering/ | process engineering |
| engineering/remove/ | engineering remove |
| remove/materials | remove materials |
| materials/anodic/ | materials anodic |
| anodic/dissolution/ | anodic dissolution |
| dissolution/metal/ | dissolution metal |
| metal/workpiece/ | metal workpiece |

**Table 2.** Partial phrases extracted based on 3-Word
**表 2.** 基于 3-Word 提取的部分词组

| 3-Word | 提取的词组 |
|---|---|
| Electrochemical/machining/process/ | Electrochemical machining process |
| machining/process/engineering | machining process engineering |
| process/engineering/remove/ | process engineering remove |
| engineering/remove/materials/ | engineering remove materials |
| remove/materials/anodic | remove materials anodic |
| materials/anodic/dissolution | materials anodic dissolution |
| anodic/dissolution/metal/ | anodic dissolution metal |
| dissolution/metal/workpiece/ | dissolution metal workpiece |

**Table 3.** Partial phrases extracted based on 4-Word
**表 3.** 基于 4-Word 提取的部分词组

| 4-Word | 提取的词组 |
|---|---|
| Electrochemical/machining/process/engineering/ | Electrochemical machining process engineering |
| machining/process/engineering/remove/ | machining process engineering remove |
| process/engineering/remove/materials/ | process engineering remove materials |
| engineering/remove/materials/anodic/ | engineering remove materials anodic |
| remove/materials/anodic/dissolution/ | remove materials anodic dissolution |
| materials/anodic/dissolution/metal/ | materials anodic dissolution metal |

N-Word 的效率不高，它的优势是不会遗漏领域词汇，但也会提取出明显错误的合成词，例如，dissolution metal workpiece、materials anodic dissolution、materials anodic dissolution metal 等，采取直接删除的方式进行消除。表 1~3 列举了部分结果。从结果看 4-Word 提取的词组中错误率最高。基于 N-Word 算法的提取结果相对粗糙，需要对其进行过滤处理。

## 5. 基于互信息的领域术语过滤

本文利用互信息对概念进行过滤。在信息理论中，互信息(*MI*)是一种定量衡量信息相关性的方法，用来度量词语之间的关联性，互信息越大，关联性越强，越能组成一个真正的领域词汇。文献[7] [8]中基于互信息的过滤，起到很好的效果。因此，本研究假设两个词汇分别用 $\alpha$ 和 $\beta$ 表示，$\alpha\beta$ 表示新的合成词，$P(\alpha)$表示词语 $\alpha$ 在文本中出现的概率，$P(\beta)$表示词语 $\beta$ 在文本中出现的概率，$P(\alpha\beta)$表示词语 $\alpha\beta$ 在文本中出现的概率，$f(\alpha)$表示词语 $\alpha$ 出现的频率。$f(\beta)$表示词语 $\alpha$ 出现的频率。$f(\alpha\beta)$表示词语 $\alpha\beta$ 出现的频率。

定义互信息为 *MI*， $MI_{\alpha\beta}$ 表示词 $\alpha$ 和 $\beta$ 的互信息。

$$MI_{\alpha\beta} = \frac{P(\alpha\beta)}{P(\alpha)+P(\beta)-P(\alpha\beta)} \tag{1}$$

定义词汇 $\alpha$ 和 $\beta$，合成词 $\alpha\beta$ 在文本中出现的概率为 $P$。

$$P(\alpha) \approx \frac{f(\alpha)}{w_t},\ P(\beta) \approx \frac{f(\beta)}{w_t},\ P(\alpha\beta) \approx \frac{f(\alpha\beta)}{w_t} \tag{2}$$

由(1) (2)联合推理得到公式(3)

$$MI_{\alpha\beta} \approx \frac{f(\alpha\beta)}{f(\alpha)+f(\beta)-f(\alpha\beta)} \tag{3}$$

按照公式(1)、(2) (3)，将三个词的互信息用 $MI_{\alpha\beta\gamma}$ 表示，分子为三个词的交集，用 $f(\alpha\beta\gamma)$表示。分母 $f(\alpha)+f(\beta)+f(\gamma)-f(\alpha\beta)-f(\beta\gamma)-f(\alpha\gamma)+f(\alpha\beta\gamma)$ 为三个词的并集。因此，三个词的互信息数学模型如公式(4)所示。

$$MI_{\alpha\beta\gamma} \approx \frac{f(\alpha\beta\gamma)}{f(\alpha)+f(\beta)+f(\gamma)-f(\alpha\beta)-f(\beta\gamma)-f(\alpha\gamma)+f(\alpha\beta\gamma)} \tag{4}$$

推理得到，四个词的互信息计算模型为公式(5)。

$$MI_{\alpha\beta\gamma\theta} \approx \frac{f(\alpha\beta\gamma\theta)}{\begin{array}{l} f(\alpha)+f(\beta)+f(\gamma)+f(\theta)-f(\alpha\beta)-f(\beta\gamma)-f(\alpha\gamma)-f(\alpha\theta)- \\ f(\gamma\theta)-f(\beta\theta)+f(\alpha\beta\gamma)+f(\beta\gamma\theta)+f(\alpha\gamma\theta)+f(\alpha\beta\theta)-f(\alpha\beta\gamma\theta) \end{array}} \tag{5}$$

在基于互信息的过滤中，以 $MI_{2,\text{threshold}}$ 为互信息阈值基准，则得到计算 $MI_{3,\text{threshold}}$ 的公式(6)。

$$MI_{3,\text{threshold}} = 0.75MI_{2,\text{threshold}},\ MI_{4,\text{threshold}} = 0.5MI_{2,\text{threshold}} \tag{6}$$

基于互信息的过滤算法中，首先，初始化 $MI_{2,\text{threshold}}$ 值和构词算法的实现；然后，循环计算 $MI_{i,\text{threshold}}$、$ns_i$ 和 $cw_i$；最后，基于 $MI$ 进行过滤。算法步骤如表 4 所示。

**Table 4.** Filtering algorithm table based on mutual information
**表 4.** 基于互信息的过滤算法

| | |
|---|---|
| Input | Domain text; |
| Output | Domain word set *dw*[]; |
| Step1 | Initializations:spaces number *ns*; $MI_{2,\text{threshold}}$; compound word *cw*[]; |
| Step2 | Get $MI_{i,\text{threshold}} = MI_{2,\text{threshold}}$; |
| Step3 | Calculate $MI_{3,\text{threshold}}$ and $MI_{4,\text{threshold}}$ with formula (7); |
| | For (int $i = 1$; $I < n$; $i$++) |
| Step4 | Get *cw* set [*n*]; |
| Step5 | Calculate $ns_i$; |
| Step6 | Calculate $MI_i$ with formula (3~5); |
| Step7 | If ($MI_i > MI_{i,\text{threshold}}$) |
| | {*dw*[] = $cw_i$;} |
| | End |

为了提高词汇过滤的准确性，文献[9] [10]中加入了绝对词频进行协作分析，本文中也引入绝对词频进行协作过滤。过滤结果统计显示，绝对词频为 3，互信息值取 0.45 时，过滤结果最佳。过滤后得到总词数 2137 个，部分结果见表 5。

“Electrochemical machining”的绝对词频和互信息值都满足条件，所以加入到合成词库中。“machining process”和“metal workpiece”需要进一步确定互信息阈值。“process engineering remove”和“Electrochemical machining process engineering”直接删除。

**Table 5.** Partial results based on mutual information filtering
**表 5.** 基于互信息过滤后的部分结果

| 词汇 | 绝对词频 | 互信息值 |
|---|---|---|
| Electrochemical machining | 116 | 0.7243 |
| machining process | 61 | 0.5635 |
| metal workpiece | 78 | 0.6237 |
| process engineering remove | 0 | 0 |
| Electrochemical machining process engineering | 2 | 0.1718 |

从表中可以发现，4-Word 的构词效果不是很好，而且占有比较长的时间和资源，可以采取领域专家和研究人员直接给定的方式。

对电解加工领域概念进行同义词合并，降低概念的同义率，避免相同意思的概念出现在核心领域概念集当中，因此，在提取领域概念时，需要合并其中意义相同的术语，以保证每个概念有且只有一个形式化的表示，使得领域内术语具有一致性，最终得到领域概念集合。本研究采用了基于 WordNet 词典的方法将多个同义术语合并为一个领域概念，以保证电解加工领域本体中的每一个概念只有一种形式化的表示，基于同义词性和词频进行，电化学加工领域概念集的同义词合并，最后得到 1984 个领域术语。

## 6. 总结

本文构建了一种半自动化领域术语提取模型对电解加工领域术语进行提取，利用 N-Word 算法进行词组概念提取。N-Word 能够提取从 1 词到 N 词的所有组合，但是会产生一些噪音词组。领域中以 2 词型和 3 词型提取效果最佳。基于 N-Word 算法的提取结果相对粗糙，利用互信息 *MI* 和绝对词频对其进行过滤，大大提高了领域术语的关联性，通过测试，发现绝对词频为 3，互信息值取 0.45 时，提取结果最佳。对过滤得到的概念进行人工干预和修正，进行同义词合并，最终得到标准化的领域概念 1894 个。此模型够减少本体构建的工作量，降低结果的主观性，提高本体构建的准确性。

## 基金项目

南通市基础研究项目(JC2021064)；江苏省高校基础科学(自然科学)面上项目(22KJD520006)。

## 参考文献

[1] 任飞亮, 沈继坤, 孙宾宾. 从文本中构建领域本体技术综述[J]. 计算机学报, 2019, 42(3): 654-675.

[2] 白宁超, 唐聃, 王亚强. 基于主动学习的传统中医症状本体构建方法研究综述[J]. 电子技术与软件工程, 2016, 13(7): 162-163+222.

[3] 余丰民, 林彦汝. 基于关键词词频统计的学科研究热点漂移程度模型构建及实证分析[J]. 情报理论与实践, 2020, 43(2):100-105.

[4] 陈辰, 王璐, 郝晓雪. 基于词频统计与语义关联的京津冀协同发展研究热点与前沿监测研究[J]. 河北科技图苑, 2018, 31(1): 31-37.

[5] Orhan, U. and Tulu, C.N. (2021) A Novel Embedding Approach to Learn Word Vectors by Weighting Semantic Relations: SemSpace. *Expert Systems with Applications*, **180**, 115-146. https://doi.org/10.1016/j.eswa.2021.115146

[6] Srinath, A.N., López, L.P., Fashandi, S., *et al*. (2022) Thermal Management System Architecture for Hydrogen-Powered Propulsion Technologies: Practices, Thematic Clusters, System Architectures, Future Challenges, and Opportunities. *Energies*, **15**, 304-314. https://doi.org/10.3390/en15010304

[7] 肖宇, 邓正宏. 信噪比约束下基于互信息的雷达波形设计[J]. 系统工程与电子技术, 2021, 43(7): 1775-1780.

[8] 程玉胜, 宋帆, 王一宾. 基于专家特征的条件互信息多标记特征选择算法[J]. 计算机应用, 2020, 40(2): 503-509.

[9] 张曼婷. 基于互信息的不完备信息系统属性约简算法研究[D]: [硕士学位论文]. 西安: 西安科技大学, 2020.

[10] Mohammadi, S.J., Fashandi, S.A.M., Jafari, S. and Nikolaidis, T. (2021) A Scientometric Analysis and Critical Review of Gas Turbine Aero-Engines Control: From Whittle Engine to More-Electric Propulsion. *Measurement and Control*, **54**, 935-966. https://doi.org/10.1177/0020294020956675