SCIENCE CHINA
Technological Sciences

•Article•

June 2023 Vol.66 No.6: 1815–1830
https://doi.org/10.1007/s11431-022-2283-7

# Domain knowledge discovery from abstracts of scientific literature on Nickel-based single crystal superalloys

LIU Yue[1,2,8], DING Lin[1], YANG ZhengWei[1], GE XianYuan[1], LIU DaHui[1], LIU Wei[4], YU Tao[5], AVDEEV Maxim[6,7] & SHI SiQi[3,4,8*]

[1] *School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China;*
[2] *Shanghai Engineering Research Center of Intelligent Computing System, Shanghai 200444, China;*
[3] *State Key Laboratory of Advanced Special Steel & Shanghai Key Laboratory of Advanced Ferrometallurgy & School of Materials Science and Engineering, Shanghai University, Shanghai 200444, China;*
[4] *Materials Genome Institute, Shanghai University, Shanghai 200444, China;*
[5] *Institute of Functional Materials, Central Iron & Steel Research Institute, Beijing 100081, China;*
[6] *Australian Nuclear Science and Technology Organisation, Locked Bag 2001, Kirrawee DC NSW 2232, Australia;*
[7] *School of Chemistry, The University of Sydney, Sydney 2006, Australia;*
[8] *Zhejiang Laboratory, Hangzhou 311100, China*

Received August 27, 2022; accepted December 12, 2022; published online April 27, 2023

Despite the huge accumulation of scientific literature, it is inefficient and laborious to manually search it for useful information to investigate structure-activity relationships. Here, we propose an efficient text-mining framework for the discovery of credible and valuable domain knowledge from abstracts of scientific literature focusing on Nickel-based single crystal superalloys. Firstly, the credibility of abstracts is quantified in terms of source timeliness, publication authority and author's academic standing. Next, eight entity types and domain dictionaries describing Nickel-based single crystal superalloys are predefined to realize the named entity recognition from the abstracts, achieving an accuracy of 85.10%. Thirdly, by formulating 12 naming rules for the alloy brands derived from the recognized entities, we extract the target entities and refine them as domain knowledge through the credibility analysis. Following this, we also map out the academic cooperative "Author-Literature-Institute" network, characterize the generations of Nickel-based single crystal superalloys, as well as obtain the fractions of the most important chemical elements in superalloys. The extracted rich and diverse knowledge of Nickel-based single crystal superalloys provides important insights toward understanding the structure-activity relationships for Nickel-based single crystal superalloys and is expected to accelerate the design and discovery of novel superalloys.

**Nickel-based single crystal superalloys, text mining, named entity recognition, credibility analysis, domain knowledge**

**Citation:** Liu Y, Ding L, Yang Z W, et al. Domain knowledge discovery from abstracts of scientific literature on Nickel-based single crystal superalloys. Sci China Tech Sci, 2023, 66: 1815–1830, https://doi.org/10.1007/s11431-022-2283-7

## 1 Introduction

In order to obtain novel materials with superior performance and low cost, the structure-activity relationships of Nickel-based single crystal (SC) superalloys are widely studied typically focusing on materials synthesis and characterization [1–3]. The cutting-edge findings are mainly published in scientific literature, which stores historical knowledge. That latent knowledge can help researchers design new materials more quickly and accurately. However, it is inefficient and laborious to manually retrieve information from the ever-increasing body of literature [4]. Information flow and

*Corresponding author (email: sqshi@shu.edu.cn)

extraction are further complicated by the fact that authors widely vary in writing style, used terms, and professional level. Therefore, an effective system to extract relevant information from diverse sources would substantially improve the efficiency of new material design.

The developments of text mining (TM) following natural language processing (NLP) technologies have enabled the automatic extraction of valuable information latent in unstructured texts [5,6]. Thereinto, named entity recognition (NER) [7], as a type of TM technology, has been widely employed, which aims to enable automatic identification and extraction of chemical attributes data, chemical synthesis conditions, organic compounds and their metabolites, chemical term spectra, and chemical structure descriptors from refs. [8–12]. Based on 800 abstracts jointly labeled by two materials experts, Ceder et al. [13–16] successfully employed Bi-LSTM (bidirectional long short-term memory) and CRF (conditional random field) model to find thermoelectric predictions and inorganic materials entities from 1.5 million abstracts of materials-science-related publications and further carried out research on solid-state synthesis materials and inorganic materials' synthesis processing. Islamaj et al. [17] doubly annotated 150 articles to tag chemical entities. Some TM tools toward materials literature were developed (e.g., ChemicalTagger [18] and ChemDataExtractor [19]) to obtain chemical identifiers. ChemDataExtractor 2.0 [20] was focused on the articles published in 26 authoritative journals to extract information on the microstructure of materials. Recently, Wang et al. [21] proposed an automated data extraction pipeline to mine compositions and properties of superalloys and generated a structural database for NLP, which was successfully mined based on labeled named entity. These studies indicate that NER can discover domain knowledge latent in texts for the research on the structure-activity relationships of Nickel-based SC superalloys.

However, NER models are driven by labeled text data, which requires the acquisition of enough data to perform manual annotation. This procedure can be laborious, time-consuming and its objectivity is affected by the different preferences of annotators. In addition, the quality of the texts affects the quality of the mined information to a certain extent. Namely, the authority of publications can determine the quality of the information mined, and then affect the information usability. Therefore, the quality of the text and the knowledge should be considered in the process of mining. For example, Gu [22] considered the source quality of bibliographic information during text mining. Nowadays, the impact factor (IF, from Journal Citation Reports) is used to evaluate the quality of publications [23]. The academic standing of authors is another important factor which can be used to gauge the quality of knowledge mined from literature. The above factors can reflect the quality of text or knowledge to some extent, but they have large differences in data formats and dimensions. How to effectively quantify the quality of text and express these factors in a unified way is one of the research focuses of credibility analysis. Further, how to integrate these factors into the process of knowledge discovery is also a critical task considered in this study.

To enable effective knowledge discovery for Nickel-based SC superalloys, we propose a knowledge discovery workflow based on a text mining framework (KD-TMF), able to extract credible knowledge and valuable data from meta-information and abstracts of Nickel-based SC superalloys literature. The distant supervision based on a dictionary is exploited to replace manual annotation and a scheme for detecting and analyzing the quality of acquired literature is proposed. This can quantify the credibility of texts and rank up-to-date and domain-related literature, so that screen out high-quality texts. Referring to *Web of Science* [24], we also map out a cooperative network of research in the area of Nickel-based single crystal superalloys based on the authors list in meta-information, to quantify the academic standing of the authors and find influential domain experts.

The main contributions of this paper are as follows. (1) A multi-dimensional credibility evaluation method is proposed from the perspective of the source timeliness, publication authority, and author's academic standing, and applied for credibility analysis of text and knowledge.

(2) Eight entity types are defined for materials NER and a dictionary of terms and twelve naming rules are established to identify and screen entities from texts, to facilitate the accuracy of NER.

(3) Based on the proposed method and its obtained results, an academic cooperative "Author-Literature-Institute" network is constructed and influential experts are identified. The overall research progress of Nickel-based SC superalloys is visualized, and contents and potential values of chemical compositions are mined and ranked for novel materials R&D.

## 2 Methods

A text mining framework, KD-TMF, is proposed to effectively extract credible and valuable domain knowledge from meta-information and abstracts of Nickel-based SC superalloys (Ni-SCS) literature, as shown in Figure 1. The knowledge mined and visualized in this study includes the chemical compositions of superalloys, the generations of developed superalloys as well as the influential experts in the academic cooperative network.

### 2.1 Multi-dimensional credibility evaluation method

To quantitively assess the credibility of literature, a

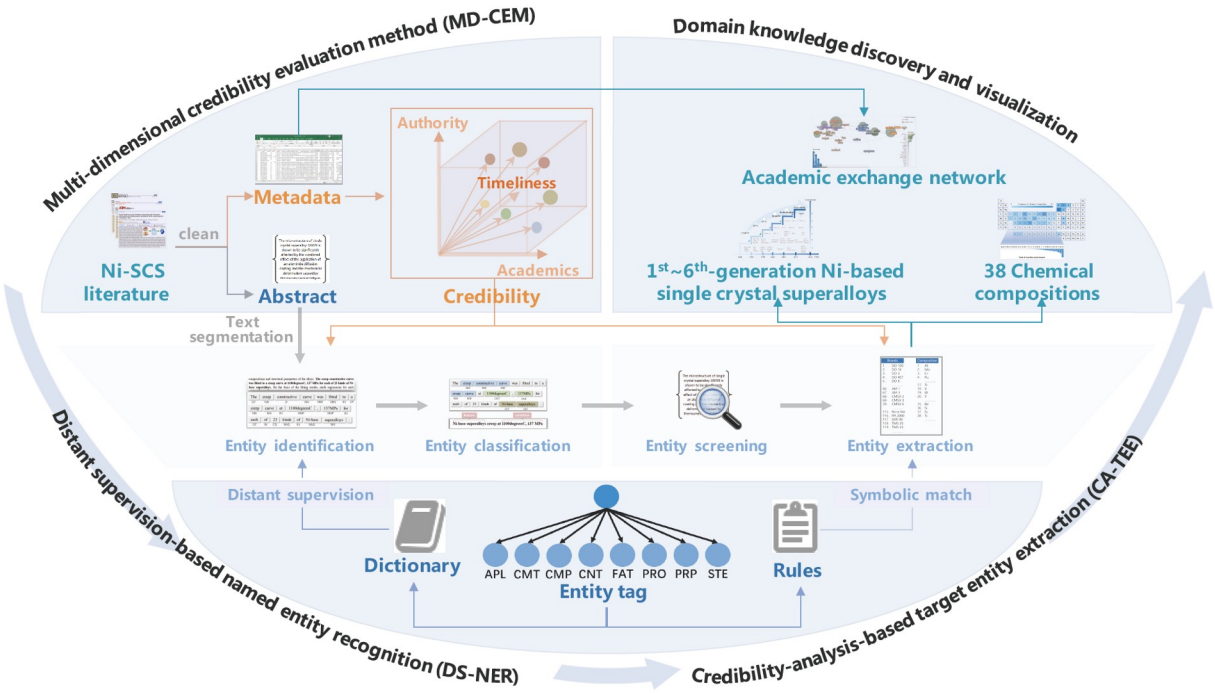


**Figure 1** (Color online) The framework of KD-TMF. It consists of (i) multi-dimensional credibility evaluation method (MD-CEM), (ii) distant-supervision-based named entity recognition (DS-NER), (iii) credibility-analysis-based target entity extraction (CA-TEE), and finally knowledge is mined and refined through MD-CEM, then visualized. The thick lines represent the main text mining process. The fine lines represent the credibility analysis, distant supervision and domain knowledge discovery process in each part, respectively.

multi-dimensional credibility evaluation method is proposed from the perspective of source timeliness, publication authority, and author's academic standing.

**Definition 1** Text credibility: The credibility $C$ of a text from literature $p$ depends on its source timeliness, publication authority, and author's academic standing, and can be expressed by

$$C = w_{c1}C_{\text{time}} + w_{c2}C_{\text{print}} + w_{c3}C_{\text{author}}, \tag{1}$$

where the weights $w_{c1}$, $w_{c2}$, and $w_{c3}$ are adjustment factors, which are determined through the entropy weight method (EWM) [25]. $C_{\text{time}}$, $C_{\text{print}}$, and $C_{\text{author}}$ represent the credibility on timeliness, authority, and academics, respectively, and the details are as follows.

### 2.1.1 *Timeliness credibility evaluation*

Knowledge stylishness depends on the timeliness of sources. Here, the timeliness credibility ($C_{\text{time}}$) for each text is calculated through eq. (2) according to the maximum time span of all available literature.

$$C_{\text{time}}(p) = \frac{y - Y_{\min}}{Y_{\max} - Y_{\min}}, \tag{2}$$

where $y$ is the publication date of $p$. $Y$ is the set of all publication dates. $Y_{\min}$ and $Y_{\max}$ are the minimum and maximum of $Y$.

### 2.1.2 *Authority credibility evaluation*

Since text sources vary, e.g., journals vs. conference proceedings, the materials-science-related literature needs to be classified and quantified for effective evaluation of authority credibility. Here, the authority credibility ($C_{\text{print}}$) of a paper depends on its type and the impact of its corresponding publication, which is calculated by

$$C_{\text{print}}(p) = C_{\text{publication}}(p) + C_{\text{type}}(p), \tag{3}$$

$$C_{\text{publication}}(p) = \begin{cases} \text{IF}, & \text{if publication is journal,} \\ \text{Level}, & \text{if publication is conference,} \end{cases} \tag{4}$$

$$C_{\text{type}}(p) = \begin{cases} 1, & \text{if type is Review,} \\ 0.8, & \text{if type is Article,} \\ 0.6, & \text{if type is Proceedings Paper,} \\ 0.2, & \text{if type is others,} \end{cases} \tag{5}$$

where IF represents the normalized value of impact factor of journals, and Level represents the class of conferences. $C_{\text{type}}(p)$ originates from the meta-information for the publication type, which is quantified as [0, 1]. More details are available in Supplementary information S.2.

### 2.1.3 *Academics credibility evaluation*

The academics credibility of a text is closely related to the domain influence of the author(s) who published it.

Therefore, quantifying the academics credibility should take the domain influence of authors into account. Supposing that literature $p$ is completed by $N$ authors, author $a$ ranks $r_i$ in the co-author list $A$, and the academics credibility ($C_{\text{author}}$) can be expressed by

$$C_{\text{author}}(p) = \sum_{a \in A} [\text{Acc}_a g_a(r_i, N, p)], \tag{6}$$

$$g_a(r_i, N, p) = \frac{1}{r_i \times \sum_{n=1}^{N} \frac{1}{n}} \ (N \leqslant 3), \tag{7}$$

$$\sum_{a \in A} g_a(r_i, N, p) = 1, \tag{8}$$

where $A$ is a matrix of $|N| \times 1$. $g_a(r_i, N, p)$ represents the contribution degree [26] of $a$ on $p$, which is calculated through eq. (7) and satisfies eq. (8). $\text{Acc}_a$ is the acceptability of $a$, which is defined below to measure the domain influence of the author.

**Definition 2** Researcher's acceptability: The acceptability of researchers depicts their academic standing in the domain, including domain ranking, relevance, leadership, and closeness.

In order to effectively evaluate the researcher's acceptability, a complex network is constructed through the cooperative relationship between authors. Thereinto, the measurements (including PageRank, Betweenness, Central point dominance, and Closeness) in complex networks [27,28] are introduced and detailed as follows.

(a) $\text{Rank}_a$ represents the author's importance and is calculated by the following formula:

$$\text{Rank}_a = \frac{1-q}{N_A} + q \sum_{i \in M(a)} \frac{\text{Rank}_i}{L(i)}, \tag{9}$$

where $N_A$ represents the number of authors in the cooperative network. $q$ is the damping factor, which is generally taken as 0.85. $M(a)$ is the set of authors that link to author $a$. $\text{Rank}_i$ represents the rank value of author $i$. $L(i)$ is the number of links of author $i$.

(b) $\text{Bet}_a$ represents the relevance of one researcher to that of others, which is calculated by

$$\text{Bet}_a = \sum_{A_i, A_j \in A, i \neq j} \frac{\sigma(A_i, a, A_j)}{\sigma(A_i, A_j)}, \tag{10}$$

where $\sigma(A_i, a, A_j)$ is the number of the shortest paths (calculated by Dijkstra algorithm) between authors $A_i$ and $A_j$ passing through author $a$, $A_i \neq A_j \neq a$. $\sigma(A_i, A_j)$ is the total number of shortest paths between $A_i$ and $A_j$. The more paths the author is in, the more important the author is in the cooperative network.

(c) $\text{CPD}_a$ represents the researcher's leadership in cooperation and is calculated by

$$\text{CPD}_a = \frac{1}{N_A - 1} \sum_{a \in A} (\text{Bet}_{\max} - \text{Bet}_a), \tag{11}$$

where $\text{Bet}_{\max}$ is the maximum value of Bet in the subnetwork which contains author $a$.

(d) $\text{Cls}_a$ represents the researcher's closeness to others and is calculated by

$$\text{Cls}_a = \left[ \sum_{A_i, A_j \in A, i \neq j}^{N_A} \frac{d(A_i, A_j)}{(N_A - 1)} \right]^{-1}, \tag{12}$$

where $d(A_i, A_j)$ represents the distance (the length of the shortest path) between authors $A_i$ and $A_j$ in a cooperative network.

Then, the researcher's acceptability is calculated as follows:

$$\text{Acc}_a = w_{a1}\text{Rank}_a + w_{a2}\text{Bet}_a + w_{a3}\text{CPD}_a + w_{a4}\text{Cls}_a, \tag{13}$$

where $w_{a1}$, $w_{a2}$, $w_{a3}$ and $w_{a4}$ represent the weights of 4 indexes, which are determined by the EWM.

The processes of constructing a cooperative network and calculating acceptability are shown in Algorithm 1. Thereinto, researchers with high values of Rank, Bet, CPD, and Cls are found as Hubs in a network.

## 2.2 Distant-supervision-based named entity recognition

### *2.2.1 Construction of dictionary based on materials entity types*

In NER tasks, the entity types commonly need to be predefined, e.g., seven entity types have been proposed by Weston et al. [14] for tackling the NER task for inorganic materials. Though some types of entities have been proposed for specific materials [15], there is no label representing the compositions of materials for mining the structure-activity relationships of superalloys. Therefore, in this study, comprehensive materials entity types are defined, which are detailed in Definition 3 and Table 1.

**Definition 3** Materials entity types: Materials entities are special terms used to express materials information. Their types are defined as follows:

**Algorithm 1** Acceptability evaluation based on a cooperative network

**Input**:
The set of authors of all literature: Author.
**Output**: The acceptability of all authors: $N$.
1. Build a network through the cosign relationship among Author;
2. Initialize the Rank based on the number of links owned by each author;
3. **do**

① Calculate the Rank by eq. (9) for all authors;
② Update the Rank of each author based on the results of step ①;

**Until** the Rank converges;
4. Calculate the Bet, CPD, and Cls by eqs. (10)–(12);
5. Normalize Rank, Bet, CPD, and Cls;
6. Calculate the weight of 4 indexes by EWM: $w_{a1}$, $w_{a2}$, $w_{a3}$, and $w_{a4}$;
7. Calculate the Acc of each author according to eq. (13).

Type⟨APL,CMT,CMP,CNT,FAT,PRO,PRP,STE⟩,

where APL, CMT, CMP, CNT, FAT, PRO, PRP, and STE represent application, characterization, composition, condition, feature, processing, property, and structure, respectively.

Based on the defined entity types, a dictionary of Nickel-based SC superalloys is built for distant supervision [29] to alleviate the effect of manual annotation, whose building process is shown in Figure 2. The terms related to Nickel-based SC superalloys are collected from *Terms of Materials Science and Technology* [30] and structured and stored via Trie tree [31]. Following this, domain-related phrases are selected from terms under the guidance of materials experts and classified according to 8 materials entity types manually. On this basis, the dictionary is constructed via the obtained instances (i.e., domain entities with tags). This can reduce the workload of manual annotation on large amounts of text and improve the recognition rates of common words.

### 2.2.2 *Distant supervision based on dictionary and CRF*

Note that the coverage of the proposed dictionary may limit the performance of NER models. Besides, off-the-shelf models show poor performance on unlabeled texts and identify long multiword entities as multiple shorter phrases with distorted meanings. To this end, a NER method based on the distant supervision, combining the dictionary and the CRF model, is employed to recognize entities. With the help of predefined instances in the dictionary, the distant-supervision-based NER model can identify target entities from unlabeled texts effectively.

Figure 3 illustrates the results of the main steps of NER. Concretely, the raw text is cleaned and segmented for lemmatization and POS (part of speech) tagging. As shown, the POS tags of named entities are mainly "NN", "NNP", "NNS", etc. Note that the target entities are usually in the form of nouns, therefore, the POS tags have less impact on our NER tasks. Then, by utilizing NLTK (the natural language toolkit based on Python), the text is segmented into phrases (marked by colored boxes) and stop words (uncolored boxes). Finally, domain phrases are classified into different entity classes (e.g., marked by "PRP") according to the instances in the dictionary.

The process of domain entity recognition is shown in Algorithm 2. Through matching similar instances in the dictionary, extraneous words are filtered out from the phrase set. Then, the phrases are categorized into 8 predefined types

**Table 1** The detailed information of eight materials entity types

| Type | Tag | Description | Example |
|---|---|---|---|
| Application | APL | Any advanced applications or any specific devices such as superalloys, etc. | CMSX-6, DD6, SRR99 |
| Characterization | CMT | Any methods used to characterize a material, such as equations, experimental models, theories, etc. | XRD, TEM, DFT, Rabotnov-Kaehanov's formulation |
| Composition | CMP | Anything related to chemical formulas or any descriptions of what is inside materials related to contents, etc. | Content of Al, γ phase mole fraction |
| Condition | CNT | Any descriptions of the external and service conditions of materials. | 1050°C, 100 MPa, 5 h |
| Feature | FAT | Special descriptions of the type or shape of samples. | Single crystal |
| Processing | PRO | Any techniques or processes for synthesizing a material. | Pulsed laser deposition, solid-state reaction, annealing, etching |
| Property | PRP | Any measurable values with units, any qualitative properties, behavior or phenomenon exhibited by a material in a physical/chemical process or mechanism. | Creep behavior, stacking fault energy, phase transition temperature, elastic property |
| Structure | STE | Names used to describe crystal structures, phases, defects, etc. | FCC phase, liquid crystalline, dislocation, line defect |

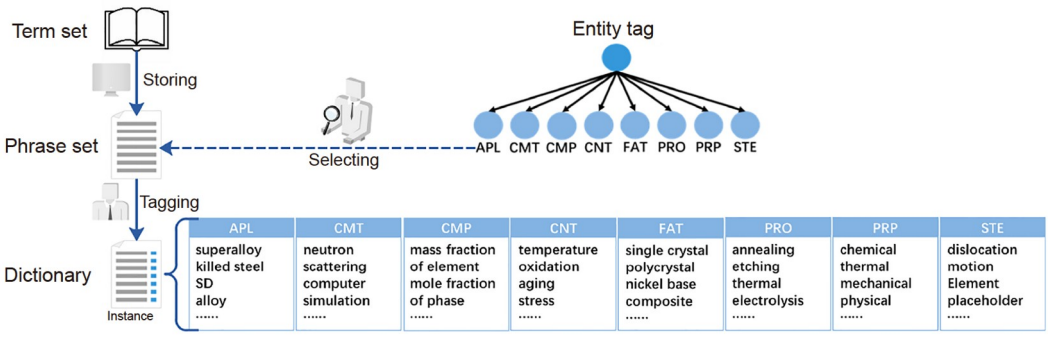


**Figure 2** (Color online) The dictionary is constructed based on eight entity types.

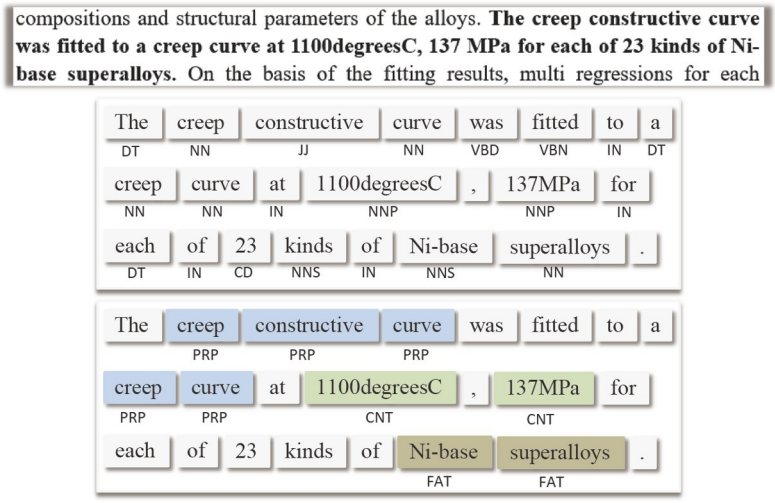


**Figure 3** (Color online) Example of text segmentation, POS tagging, and named entity recognition.

**Algorithm 2** Distant-supervision-based named entity recognition

**Input**:
The text set: Abstract;
The materials entity types: Type;
The set of materials terms with type: Dictionary.
**Output**: The set of entities with type: entity.
1. Text cleaning and structured processing;
2. **do**

① Word segmentation by space/symbol;
② Text filtering to exclude Stop Word;
③ POS tagging for words through NLTK;

**Until** All abstracts are processed
3. Filter extraneous phrases according to Dictionary;
4. Categorize phrases into 8 types and carry out synonym disambiguation by Dictionary;
5. Calculate the features of each word and combine with labels;
6. Perform training and inference on the CRF by supervised data;
7. Output labeled phrases as entity.

with the domain dictionary to acquire supervised entity information. During this stage, singular and plural, abbreviations of words in the dictionary are used to disambiguate synonymous entities. On this basis, the features of each word in the abstract are calculated. Combined with tags, the CRF model is trained and reasoned on supervised data. Finally, the entity set is obtained for credible entity extraction.

## 2.3 Credibility-analysis-based target entity extraction

### *2.3.1 Target entity extraction based on naming rules*

The information contained in brands reflects the research progress of different superalloys in certain time periods. Under the guidance of existing research [32], the types of superalloys can be divided into single-crystal, polycrystal, multicomponent system, etc. Based on the convention of the manufacturers around the world and the standard of Classification & Designation [33] of superalloys, twelve symbolic rules for naming materials are formulated to further screen the alloy brands from the entity, shown in Table 2.

It is generally known that the chemical compositions affect the microstructure and properties of materials and are in turn used to establish the structure-activity relationships of superalloys. To this end, we here employ the co-occurrence rule of descriptors representing chemical compositions and brands signifying superalloys to extract their chemical compositions. Thereinto, the co-occurrence rule is derived from the condition that the entities of "composition" and "application" from the same text correspond to the target descriptors and brands. Following this, the abbreviations and full names of descriptors are employed, and the chemical compositions are selected from the entity.

### *2.3.2 Entity credibility analysis based on information tracing*

In order to mine the latent domain knowledge accumulated in literature, it is necessary to validate it in terms of authenticity and usability. In this context, entity credibility is proposed based on Definition 1 and integrated into the extracted information to obtain reliable domain knowledge.

**Definition 4** Entity credibility: Assuming that entity $E$ is obtained from text set $L$, its credibility $C_E$ can be calculated by the following formula:

$$C_E = \frac{1}{|P|'} \sum_{L_i \in L}^{|P|'} C(L_i), \tag{14}$$

where $|P|'$ represents the number of texts containing entity $E$. $L_i$ represents the $i$-th element of $|P|' \times 1$ matrix $L$. $C(L_i)$ is the credibility of $L_i$ calculated by eq. (1).

The sources of entities are obtained through information tracing [34] based on the DOIs of texts to calculate the $C_E$.

**Table 2** Naming rules of Nickel-based SC superalloys

| Form | Rule |
|---|---|
| 1 | "DD" $x$ \| $x$∈{2, 3, 4, …} |
| 2 | "Haynes" $x$ \| $x$∈{230, 282, …} |
| 3 | "Inconel" & "IN" $x$ \| $x$∈{625, 690, 706, …} |
| 4 | "Mar-M" $x$ \| $x$∈{200, 247, …} |
| 5 | "Rene" $x$ \| $x$∈{80, 88, 95, …} & "Rene N" $x$ \| $x$∈{4, 5, …} |
| 6 | "AM" $x$ \| $x$∈{1, 3, …} |
| 7 | "CMSX" $x$ * \| $x$∈{2, 3, 4, …}, *∈{R, K} |
| 8 | "PWA" $x$ \| $x$∈{1422, 1480, 1484, 1497, …} |
| 9 | "RR" $x$ \| $x$∈{1000, 2000, 2072, …} |
| 10 | "SRR" $x$ \| $x$∈{99, …} |
| 11 | "TMS" $x$ \| $x$∈{75, 138, 162, …} |
| 12 | Ni-$X$ \| $X$∈{Al, Cr, Fe, …} |

Following this, the timelines of different superalloys can be visualized based on the publication dates of $L$.

The ranges for the content of different chemical elements are established by analyzing the minimum and maximum of the same chemical composition in "composition" entities. Then, the range of chemical compositions for each element can be further fine-tuned by credibility analysis when the values of $C_E$ of target entities are in the range of the pre-specified confidence interval. Significantly, the entity importance degree is introduced based on Definition 4, to quantify the research value of chemical elements in materials and rank the importance of different transition metal elements.

**Definition 5** Entity importance degree: Assuming that the credibility of entity $E$ is $C_E$, its importance degree $I_E$ can be calculated by the following formula:

$$I_E = C_E f_E, \tag{15}$$

where $f_E$ represents the frequency of entity $E$ in the Abstract.

## 3 Experiments

### 3.1 Experimental dataset and preprocessing

The experimental dataset contains 22711 meta-information entries from 3154 abstracts, collected from *Web of Science*. Thereinto, the meta-information consists of title, author, keywords, publication type, research direction, research institute, publication date, publication, IF/Level, and DOI, which are detailed in Supporting Information S.1.

Because the meta-information can be incomplete or incorrect, e.g., the "publication date" contains only "year", the "author" is wrong, or the "publication type" is not unique, etc., the preprocessing steps of the dataset need to be carried out.

(1) Cleaning meta-information. The records with the same "DOI" are merged. Then, the operations of numeralization and normalization are performed for the structured meta-information.

(2) Cleaning abstracts. Auto marking and manual correcting are performed to obtain structured text set.

### 3.2 Experimental setups

The proposed text mining framework, KD-TMF, consists of MD-CEM, DS-NER, and CA-TEE, of which details are set as follows.

**MD-CEM.** The Acc of each author is calculated by 4 indexes based on 5136 objects, and the $C$ of each text is calculated through $C_{\text{time}}$, $C_{\text{print}}$, and $C_{\text{author}}$. Here, the first corresponding author in the co-author list is selected for the calculation of $C_{\text{author}}$. According to EWM, the weights of eqs. (1) and (13) can be obtained, which are $w_{c1}$=0.3328, $w_{c2}$=0.3356, $w_{c3}$=0.3316, $w_{a1}$=0.3982, $w_{a2}$=0.3878, $w_{a3}$=0.2197, and $w_{a4}$=−0.0057, respectively. The calculation formulas of weights are displayed as follows:

$$f_{mn} = \frac{x'_{mn}}{\sum_{n=1}^{N} x'_{mn}}, \tag{16}$$

$$E_m = -K \sum_{n=1}^{N} \left(f_{mn} \ln f_{mn}\right), \tag{17}$$

$$w_m = \frac{1 - E_m}{M - \sum_{m=1}^{M} E_m}, \tag{18}$$

where $M$ is the number of evaluation indexes, and $N$ is the number of evaluation objects. $x_{mn}$ is the value of the $m$-th evaluation index of the $n$-th evaluation object. $x'_{mn}$ represents the normalized value of $x_{mn}$, which is standardized by the Min-max normalization. $f_{mn}$ is the contribution of the $n$-th object to the $m$-th index. It is assumed that when $f_{mn} = 0$, $E_m = 0$. $E_m$ is the entropy of the $m$-th index. $K$ is a positive constant, which is set as $\ln n$ in this study. $w_m$ is the weight of the $m$-th index, and $\sum_{m=1}^{M} w_m = 1$ $(m = 1, 2, \ldots, M)$.

**DS-NER.** In order to improve the credibility of obtained entities, the confidence interval of $C$ is set as 95%. In this context, the refined $C$ of remained texts becomes 0.35, indicating that there are only a few texts with low credibility. All the texts were used for our NER experiments, since deleting texts with low credibility has little influence on the performance of the NER model. The details of the dictionary are shown in Supporting Information S.3. The epochs of CRF are set at 100.

In order to validate the performance of DS-NER effectively, an evaluation system containing Precision, Recall, and $F1$-Score is employed. The equations in detail are as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{19}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \tag{20}$$

$$F1\text{-Score} = \frac{2\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}, \tag{21}$$

where TP, TN, FP, and FN denote positive samples classified correctly, positive samples classified incorrectly, negative samples classified incorrectly, and negative samples classified correctly, respectively. Precision refers to the ratio of the correct identifications to all identifications. Recall denotes the ratio of the correct identifications to the total number of entities. *F*1-Score represents the harmonic average of Precision and Recall.

**CA-TEE.** Based on the text credibility evaluated by MD-CEM, the credibility and the importance degree of entities can be calculated through eqs. (14) and (15). In order to evaluate the effectiveness of credibility analysis on knowledge discovery, text credibility with different dimensions is calculated and set as the threshold for entity screening. The contrast experiments are employed in Subsection 3.3.3.

## 3.3 Results and discussion

### *3.3.1 Text credibility analysis*

Through the evaluation of $C_{\text{time}}$, $C_{\text{print}}$, and $C_{\text{author}}$, the text credibility distributions of timeliness, authority, and academics are obtained and shown in Figure 4. It can be seen from Figure 4(a) that most of the credibility values evaluated by $C_{\text{time}}$ are high, which are in the range of [0.35, 1.0]. This indicates that only employing timeliness credibility fails to reveal the authority and academics of texts. Also, Figure 4(b) and (c) show that the credibility distribution on $C_{\text{print}}$ or $C_{\text{author}}$ is too concentrated, which is not conducive to text credibility evaluation. This indicates that authority or academics credibility has the same situation as timeliness credibility.

In order to validate the effectiveness of MD-CEM, the indexes evaluating text credibility from different perspectives are compared. The credibility distributions of 3154 texts calculated through timeliness with authority, authority with academics, and timeliness with academics are shown in Figure 5(a)–(c), respectively. It can be seen from Figure 5(a)–(c) that the credibility values of many texts are in the range of [0.0, 0.1], which is unlikely and suggests that appropriate weighting should be introduced. Weighted by EWM, the text credibility is calculated based on three credibility indexes, and the credibility distribution is shown in Figure 5(d). 87.38% of *C* values are in the range of [0.4, 1.0] while 37.88% of *C* values are in the range of [0.6, 0.7], which indicates that most texts have high credibility and only the ones from literature published a long time ago and/or by authors with low Acc have the low credibility.

Following this, the literature is ranked according to corresponding text credibility, and we here select the top 20 sources with the highest *C*. The meta-information of the literature is listed and visualized in Figure 6. As shown, "Liu L", "Zhang J", and "Meng J" have more literature with high IF, which can assist readers in finding active experts. The publication rate shows an increasing trend, especially in 2005–2008 and 2015–2018. It indicates that great attention has been paid to the research on Ni-based SC superalloys in these years.

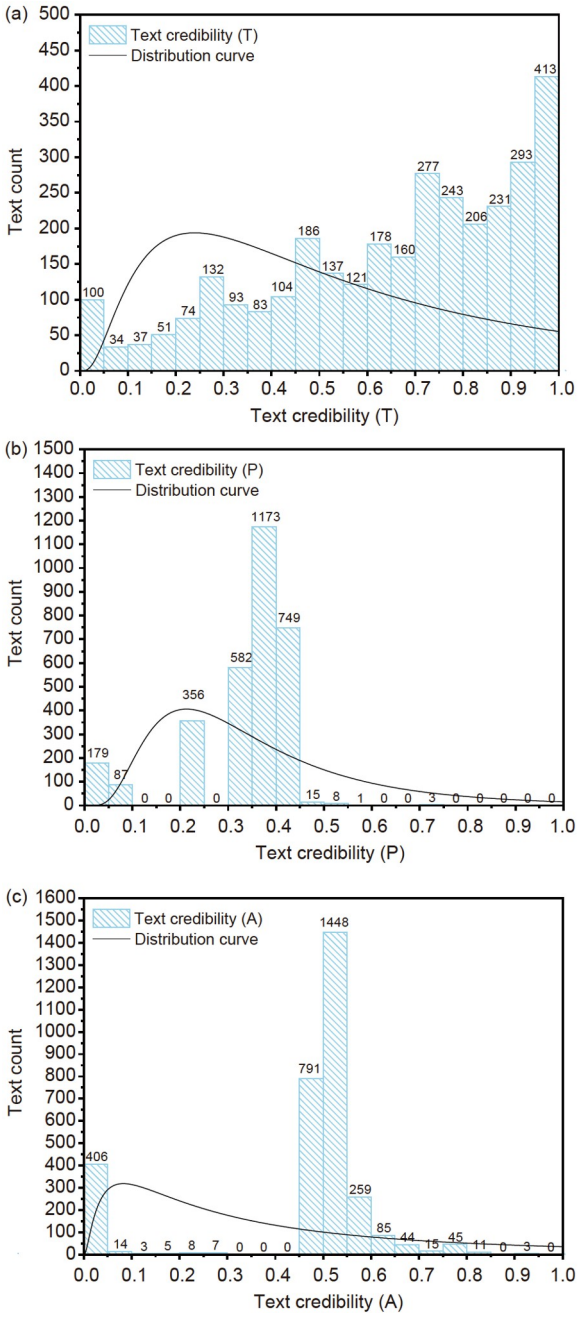


**Figure 4** (Color online) The credibility distributions of 3154 texts evaluated by single index. The text credibility calculated by (a) timeliness, (b) authority, and (c) academics.

### *3.3.2 Domain entity recognition*

Aiming to validate the effectiveness of DS-NER, a series of

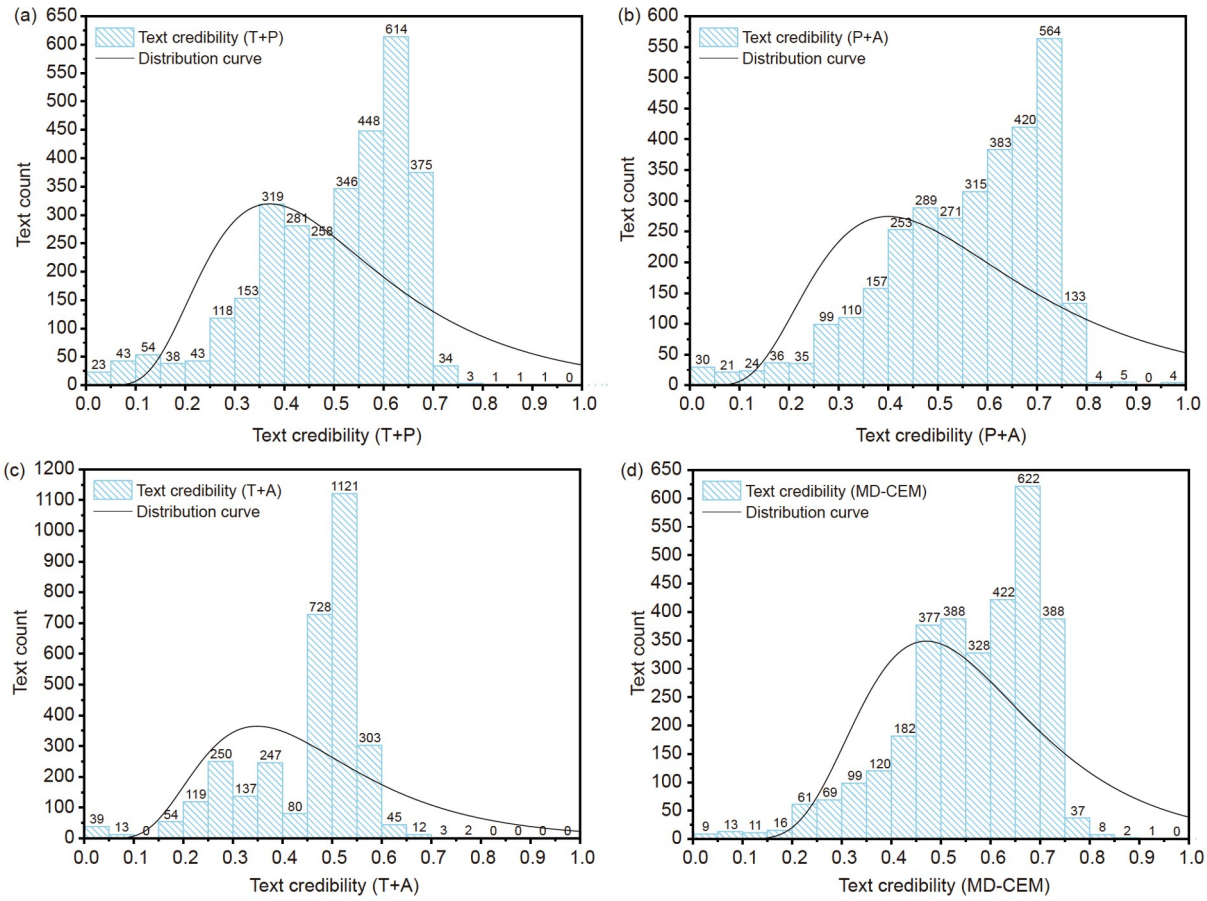


**Figure 5** (Color online) The credibility distribution of 3154 texts under different evaluation indexes. The text credibility calculated by (a) timeliness with authority, (b) authority with academics, (c) timeliness with academics. (d) The text credibility evaluated by MD-CEM.

comparative experiments are performed with comparative methods and the eight types of entities. Table 3 shows the experimental results. Only relying dictionary on identifying entity has poor performance (accuracy of 55%). However, DS-NER (CRF+Dictionary) obtains the best ability to recognize entities and further classify them to eight types with accuracy of 85%. This is because CRF has excellent identification performance (more than 87%) for the sequence tags, i.e., B (begin), I (inside) and O (outside), and the combination of dictionary and CRF can improve the recognition ability for long and multi-class entities. The details of entities recognized by DS-NER are shown in Supplementary Information S.3. The experimental results for different entity types reflect the performance of DS-NER on the entities of "application", "characterization", "structure", etc., which are detailed in Supplementary Information S.4.

Then, to intuitively express the importance degree of different entities, the frequency of entities is analyzed and shown in Figure 7. As shown, "composition" ("Ni", "Al", "Re", etc.), "feature" (e.g., "Nickel-based"), and "characterization" ("SEM", etc.) types of entities have a higher frequency. This indicates that chemical compositions, types, and synthesis & characterization of materials are relatively important in the research of Nickel-based SC superalloys, followed by the applications, microstructures, properties, processing, and conditions.

### 3.3.3 *Credible entity extraction*

Following entity recognition, the entity credibility analysis is performed based on naming rules and information tracing. The distributions of $C_E$ on brands and generations of alloys are shown in Figure 8. It can be found from Figure 8(a) that more than 95% of the $C_E$ values are in the range of [0.35, 0.85], and the ratio in the range of [0.5, 0.55] is the highest, accounting for 20.80% of the total. Compared with the text credibility (Figure 5(d)), the $C_E$ of brands has a similar distribution, revealing the reliability of extracted information. Figure 8(b) shows that the $C_E$ values of generations are in the range of [0.2, 0.85], and the fraction in the range of [0.6, 0.7] is the highest, accounting for 34.29%. The brands of Nickel-based SC superalloys are available in Supporting Information S.5.

To mine credible domain knowledge of Nickel-based SC superalloys, entity credibility obtained above is utilized to screen credible entities and analyze the structure-activity relationships from the perspective of generations of brands

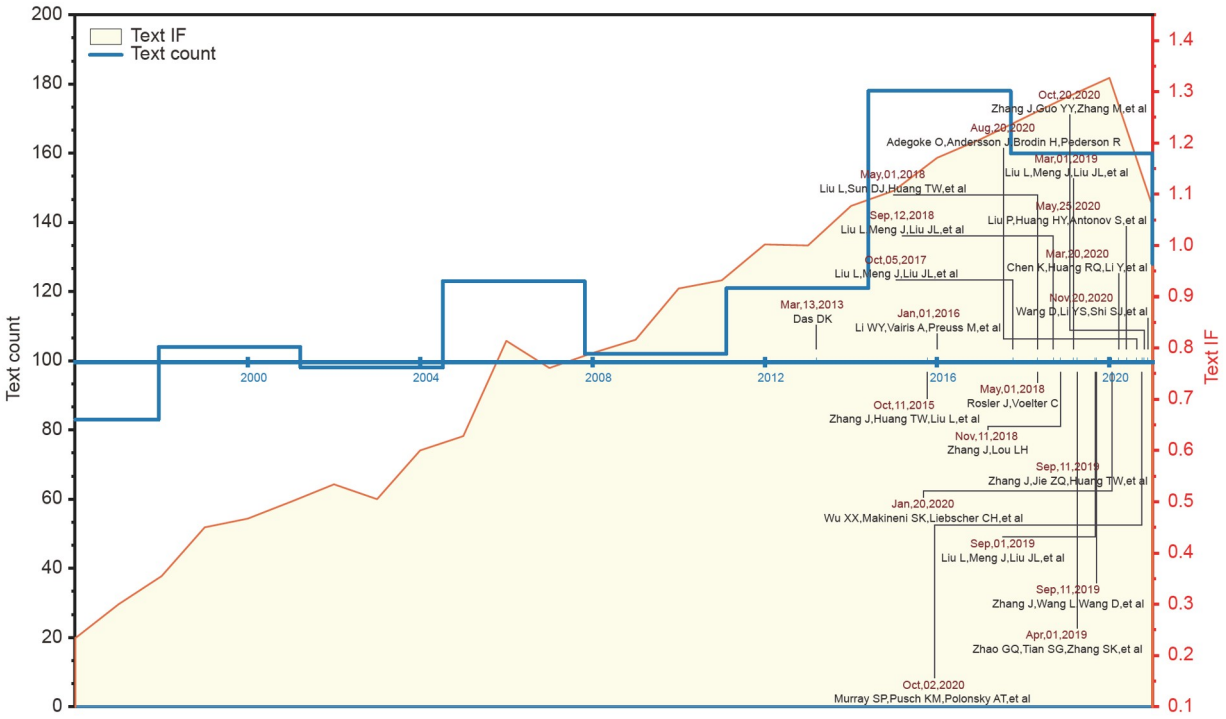


**Figure 6** (Color online) The overview of the literature from 1996 to 2020 in the field of Ni-based SC superalloys. The fine lines mark the publication date and authorship of the top 20 pieces of literature.

and chemical compositions.

**Alloy brands.** The brands within 95% confidence interval are screened out through their $C_E$. Then, the distribution of brands in different generations is analyzed based on their research dates, which is visualized in Figure 9. The results indicate that Nickel-based SC superalloys have reached the 6th generation. Thereinto, the studies on 2nd-generation superalloys (highlighted by the green curve) are the most extensive, accounting for 48.99%. This is mainly because the 2nd-generation superalloys have good properties and low cost, which is consistent with what has been reported in [35]. From 1996 to 2020, CMSX 4 has been studied the most, with more than 300 instances, followed by TMS 75 and DD 6. The composition formulas of functionally applied superalloys can be further used to guide the new materials R&D.

**Chemical compositions.** Similarly, the chemical compositions within 95% confidence interval are screened out through their $C_E$. Following this, the thresholds of the contents of 20 chemical elements are refined to provide effective parameter constraints for materials machine learning, which are illustrated in Table 4. We can see that the thresholds with credibility (such as Al, Nb, W, etc.) are in a reasonable range, compared with that without credibility. The corresponding distributions of contents are shown in Supporting Information S.6. The contents of Al and W with credibility are more well-defined compared with the raw values. Especially, the contents of V and Y in materials can be set as 4.0 wt.% and 1.0 wt.%, respectively, while this is uncertain in the raw data without credibility analysis. The metal elements (Al, Co, Cr, etc.) account for relatively high contents, non-metallic elements (B, C, etc.) make only a small contribution, while rare earth elements (Re, Ru, etc.) are studied relatively more. The elements with small contents (such as Nb, V, Y, Zr, etc.) may be expected to be present only in a small number of new materials.

**Table 3** The NER performance of comparative methods

| Method | $P$ (%) | $R$ (%) | $F1$-Score (%) |
|---|---|---|---|
| CRF | 87.89 | 87.74 | 87.77 |
| Dictionary | 67.85 | 49.33 | 55.27 |
| **DS-NER** | **84.94** | **85.51** | **85.10** |

To validate the credibility analysis of knowledge, the modes and thresholds of contents of different chemical compositions are compared, as illustrated in Table 5. The refined thresholds of contents (such as Al, Co, Cr, etc.) are generally consistent under different $C$, indicating that the credibility analysis is effective for parameter constraints. However, the mode of the content of Re is usually set as 0 wt.% when the $C_E$ is analyzed through T+P, P+A, or T+A. This situation seems to be inconsistent with the laboratory preparations of materials. It can be found that the modes of low-content chemical compositions (such as Nb, Y, and Zr), analyzed by MD-CEM, can be obtained successfully. These low-content elements also play important roles in designing new superalloys, e.g., as was shown for Nb and reported in

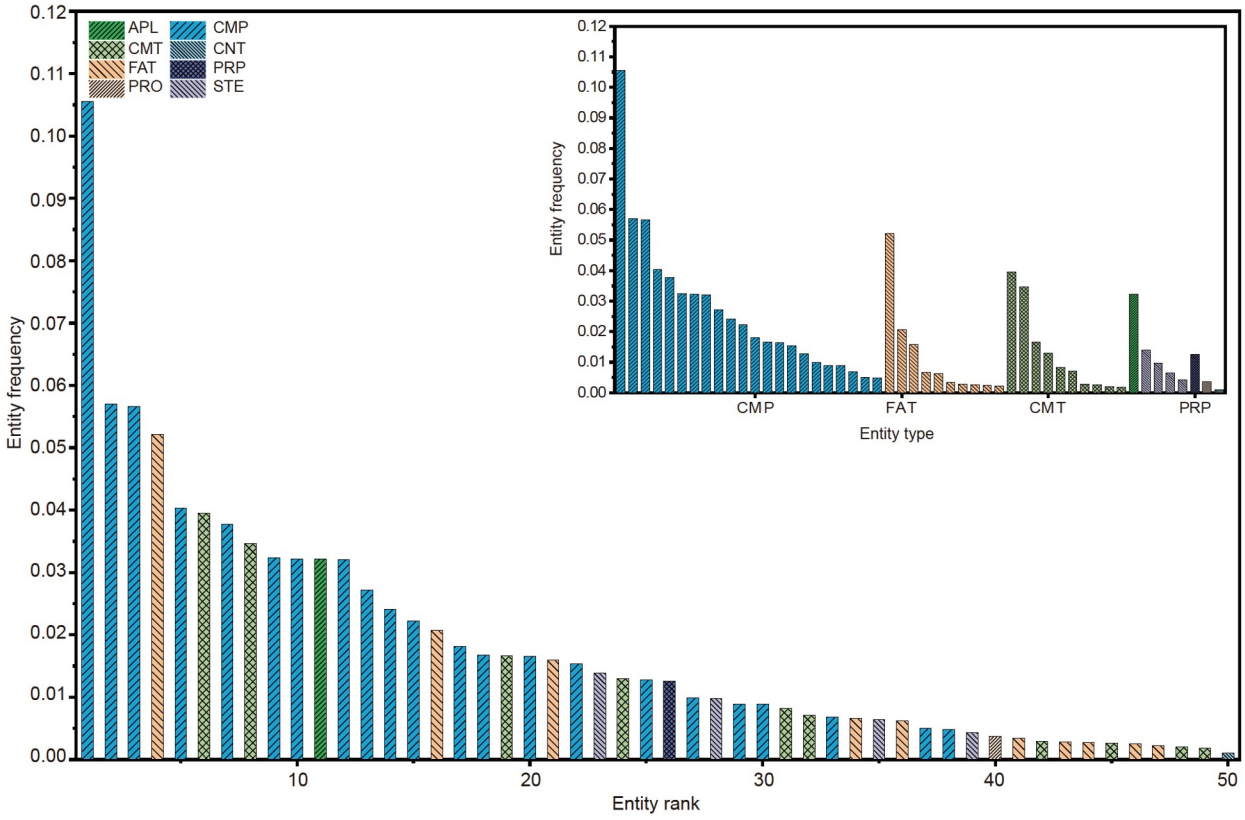


**Figure 7** (Color online) The occurrence frequency of 8 types of entities.

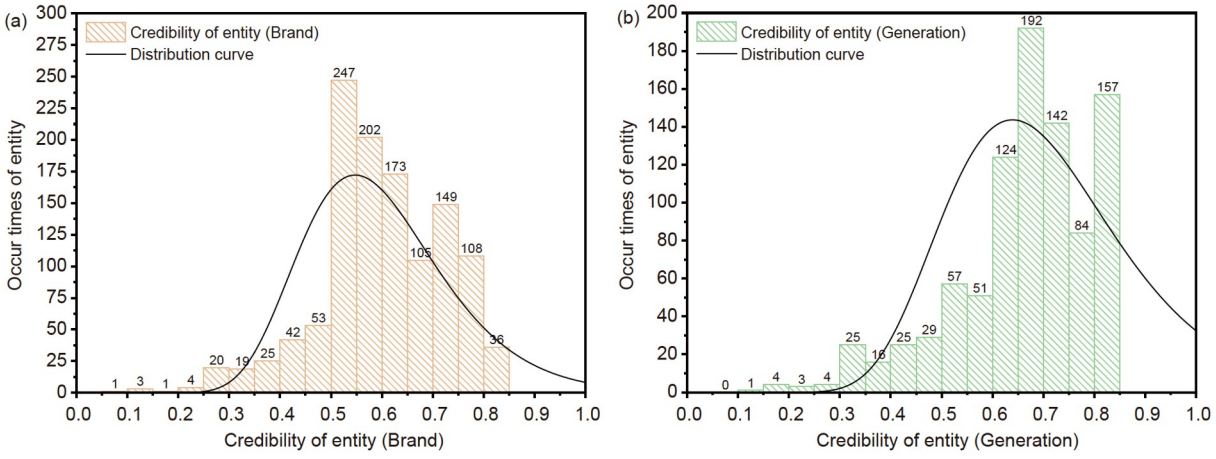


**Figure 8** (Color online) The credibility distribution of brands (a) and generations (b) of Nickel-based SC superalloys.

[36].

The potential impact of different transition metal elements in materials R&D is further explored, i.e., the frequency of entities in the texts is counted and the importance degree of entities is calculated through eq. (15). On this basis, the values of $I_E$ of 38 transition metal elements are ranked. The results are detailed in Supporting Information S.7, and the implications are discussed below in Subsection 3.4.3.

### 3.4 Applications

To validate the proposed method, we here introduce information extraction, information tracing, and information visualization [37–39] in a fashion similar to that employed in other sectors [40]. Specifically, knowledge graph visualization is combined with the MD-CEM to construct an academic cooperative network to identify academic

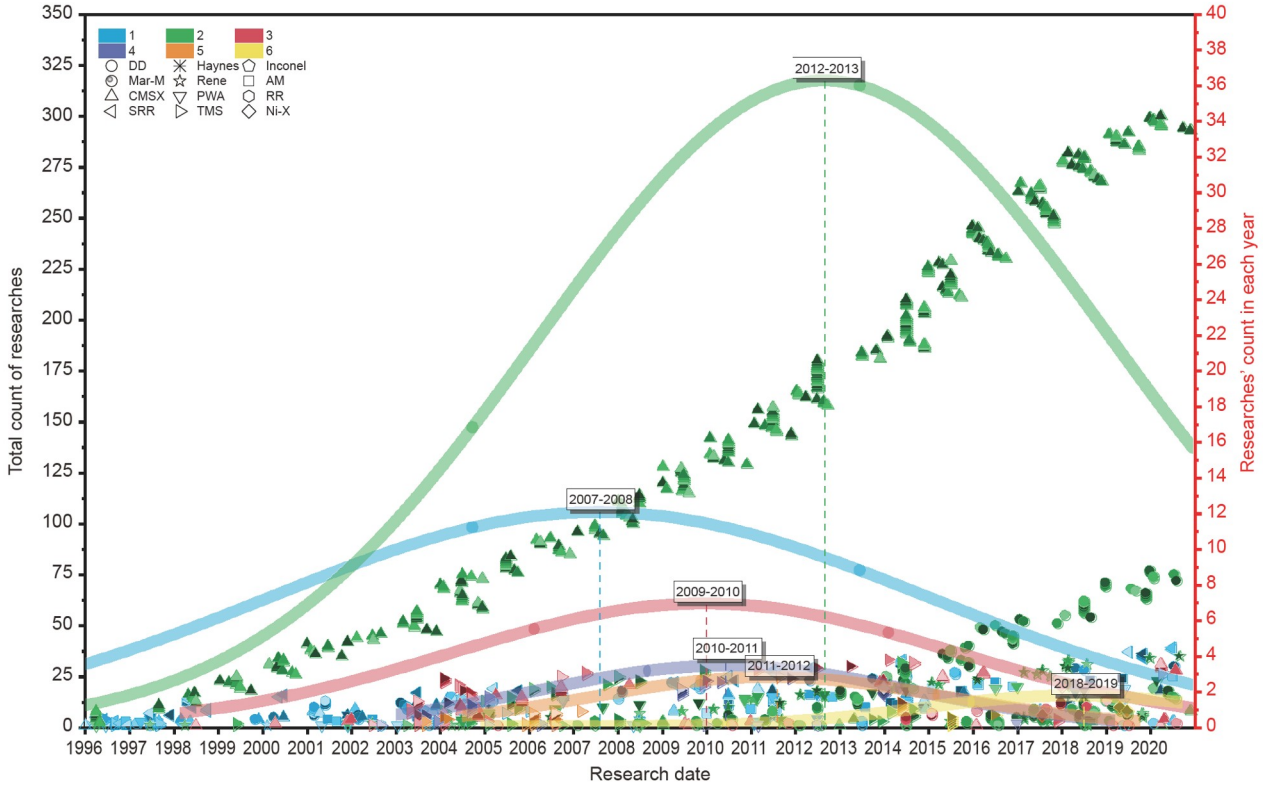


**Figure 9** (Color online) The distribution of 1st–6th-generation Nickel-based SC superalloys on timeline.

**Table 4** The contents of 20 chemical compositions in Nickel-based SC superalloys

| Chemical composition | Content (wt.%) | | | | | |
|---|---|---|---|---|---|---|
| | Mode | First quartile | Median | Third quartile | Threshold without credibility | Threshold with credibility |
| Al | 6 | 2.05 | 3.5 | 5.4 | [0.2, 22] | [0.2, 8.5] |
| B | 0 | 0.00375 | 0.015 | 0.765 | [0, 3.5] | [0, 3.5] |
| C | 0.06 | 0.05 | 0.09 | 0.14 | [0, 0.6] | [0, 0.6] |
| Ce | 0.022 | 0.022 | 0.022 | 0.022 | – | – |
| Co | 0 | 5 | 9.43 | 12.1425 | [0, 20.6] | [0, 20.6] |
| Cr | 2 | 4 | 5.1 | 6.2 | [0, 42.0] | [0, 42.0] |
| Dy | 0.042 | 0.042 | 0.042 | 0.042 | – | – |
| Hf | 0.1 | 0.1 | 0.13 | 0.2 | [0, 1.5] | [0, 1.5] |
| Mo | 1 | 0.8 | 1.2 | 2 | [0, 8.5] | [0, 8.5] |
| Nb | 1 | 0.5 | 1 | 1 | [0, 1.5] | [0, 1.0] |
| Re | 4.5 | 3 | 4.2 | 4.8 | [0, 10.0] | [0, 10.0] |
| Ru | 3 | 1.3 | 2.55 | 3.4 | [0, 6.0] | [0, 6.0] |
| Si | – | 1.0875 | 1.925 | 2.7625 | [0.25, 3.6] | [0.25, 3.6] |
| Ta | 0 | 4.82 | 6.7 | 7.3 | [0, 14.5] | [0, 14.5] |
| Ti | 1 | 0.9 | 1.28 | 2.225 | [0, 5.4] | [0, 5.4] |
| V | 4 | 2.5 | 4 | 4 | [1.0, 4.0] | – |
| W | 6 | 5 | 5.6 | 6 | [0.5, 11.0] | [0.5, 9.0] |
| Y | 0.017 | 0.017 | 0.025 | 0.5 | [0.011, 1.0] | – |
| Zr | 0.25 | 0.25 | 0.25 | 0.25 | – | – |
| Ni | 75 | 60.95 | 64 | 73.75 | [1.5, 86.1] | [1.5, 86.1] |

**Table 5** The modes and thresholds of contents of 20 chemical compositions under different *C*, i.e., timeliness with authority, authority with academics, timeliness with academics, and MD-CEM

| Chemical composition | Mode of content (wt.%) | | | | Threshold of content (wt.%) | | | |
|---|---|---|---|---|---|---|---|---|
| | T+P | P+A | T+A | MD-CEM | T+P | P+A | T+A | MD-CEM |
| Al | 6 | 6 | 6 | 6 | [0.2,8.5] | [0.2,8.5] | [0.2,8.5] | [0.2,8.5] |
| B | 0.02 | 0.02 | 0 | 0 | [0,3.5] | [0,3.5] | [0,3.5] | [0,3.5] |
| C | 0.06 | 0.06 | 0.06 | 0.06 | [0.012,0.6] | [0.012,0.6] | [0,0.6] | [0,0.6] |
| Ce | 0.022 | 0.022 | 0.022 | 0.022 | – | – | – | – |
| Co | 0 | 0 | 0 | 0 | [0,20.6] | [0,20.6] | [0,20.6] | [0,20.6] |
| Cr | 2 | 2 | 2 | 2 | [1.0,42.0] | [1.0,42.0] | [1.0,42.0] | [0,42.0] |
| Dy | 0.042 | 0.042 | 0.042 | 0.042 | – | – | – | – |
| Hf | 0.1 | 0.1 | 0.1 | 0.1 | [0,0.4] | [0,1.5] | [0,1.5] | [0,1.5] |
| Mo | 1 | 1 | 1 | 1 | [0,8.5] | [0,8.5] | [0,8.5] | [0,8.5] |
| Nb | – | – | – | 1 | [0,1.0] | [0,1.0] | [0,1.0] | [0,1.0] |
| Re | 0 | 0 | 0 | 4.5 | [0,10.0] | [0,10.0] | [0,10.0] | [0,10.0] |
| Ru | 3 | 3 | 3 | 3 | [0,6.0] | [0,6.0] | [0,6.0] | [0,6.0] |
| Si | – | – | – | – | [0.25,3.6] | [0.25,3.6] | [0.25,3.6] | [0.25,3.6] |
| Ta | 0 | 0 | 0 | 0 | [0,14.5] | [0,14.5] | [0,14.5] | [0,14.5] |
| Ti | 0 | 0 | 0 | 1 | [0,5.4] | [0,5.4] | [0,5.4] | [0,5.4] |
| V | 4 | 4 | 4 | 4 | – | – | – | – |
| W | 8 | 8 | 6 | 6 | [0.5,9.0] | [0.5,9.0] | [0.5,9.0] | [0.5,9.0] |
| Y | – | – | – | 0.017 | – | – | – | – |
| Zr | – | – | – | 0.25 | – | – | – | – |
| Ni | 75 | 75 | 75 | 75 | [1.5,86.1] | [1.5,86.1] | [1.5,86.1] | [1.5,86.1] |

collaborations and influential experts. In addition, the information tracing is integrated with the CA-TEE based on the "application" entities to obtain the research progress of Nickel-based SC superalloys and discover their characteristics. Finally, statistical analysis and information tracing are employed based on the "composition" entities to explore and visualize the relative research value of different chemical elements.

### *3.4.1 Cooperative network of researchers*

The academic links between academic teams centered in 24 countries are visualized in Figure 10. It is found that the Matthew Effect is reflected between different academic unions, which is determined by the number of institutes. The institutes are mainly concentrated in Germany, followed by USA and China.

The distribution of researchers in different regions is shown at the bottom of Figure 10. The number of researchers follows a power-law distribution (highlighted by the yellow curve, with a power of 1.972). This reflects that the academic cooperative network belongs to a scale-free network and has the characteristics of preferential connectivity. Specifically, researchers in larger institutes are more inclined to collaborate. For example, "Wang CY" and "Li X" with a background in "Materials Science" and "Chemistry", respectively, cooperated in "Metallurgical Engineering", being affiliated with the "Central Iron and Steel Research Institute" and "Institute of Microstructure and Property of Advanced Materials", respectively. As a result, researchers usually set foot in different fields. As a result, interdisciplinary intersections are formed between academic teams. More details on academic cooperation (including Hubs with their research directions and institutes) are available in Supporting Information S.8.

### *3.4.2 Research progress of superalloys*

The characteristics of 1st–6th-generation Nickel-based SC superalloys in terms of chemical composition are summarized according to the distribution of components in materials. It can be found that the changes of rare earth elements are significant, which is shown in Figure 11. As shown, Re is hardly present in the 1st and 2nd-generation superalloys. The rare earth elements gradually increase contents in the 3rd–6th-generation superalloys. However, consistent with the analysis in ref. [41], the increases of elements in materials also result in a more complex microstructure of superalloys, which in turn affects their microstructural stability.

### *3.4.3 Relative research value of chemical elements*

Focusing on the structure-activity relationships, statistical analysis is performed on chemical elements in materials. The

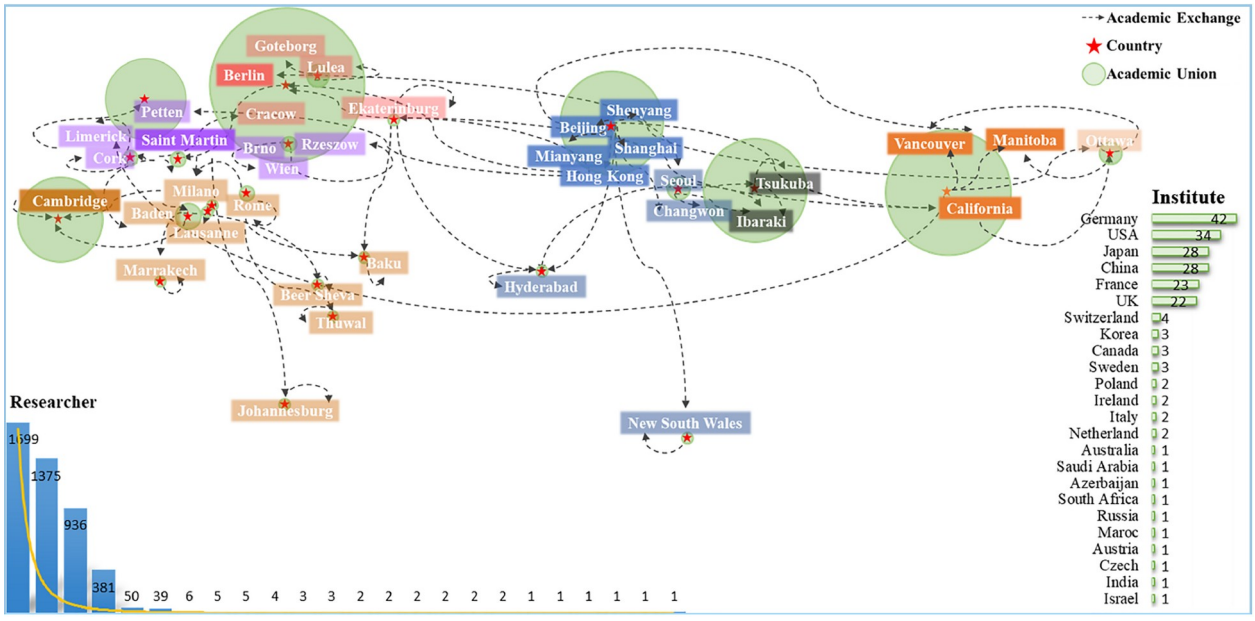


**Figure 10** (Color online) The academic cooperative network of 5136 researchers from 24 countries.

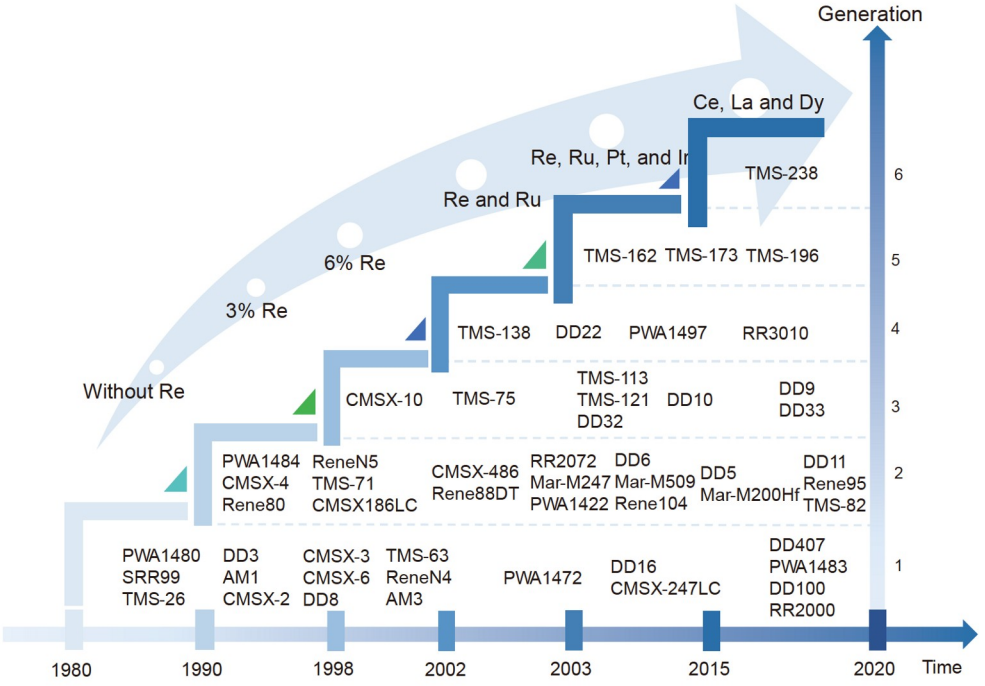


**Figure 11** (Color online) The research progress and characteristics of different generations of superalloys.

importance degrees of "composition" entities that have been applied in 1st–6th-generation Ni-based SC superalloys are calculated through eq. (15) and shown at the bottom of Figure 12. The recently added elements (Pt, Ir, La, Ce, and Dy, most are Lanthanum group elements) in the 5th and 6th generations are close to the Re and Ru in the 3rd and 4th-generation superalloys in the periodic table of elements. Besides, influenced by $C_E$, the values of $I_E$ of Ce and Dy change clearly compared with Re and Ru. As aforementioned, rare-earth elements (especially Re and Ru) play major roles in the structure-activity relationship of Nickel-based SC superalloys. The findings indicate that the research trends for Lanthanum group elements might change significantly in the future, and these elements will be tested to improve the performance and reduce the cost of new materials R&D.

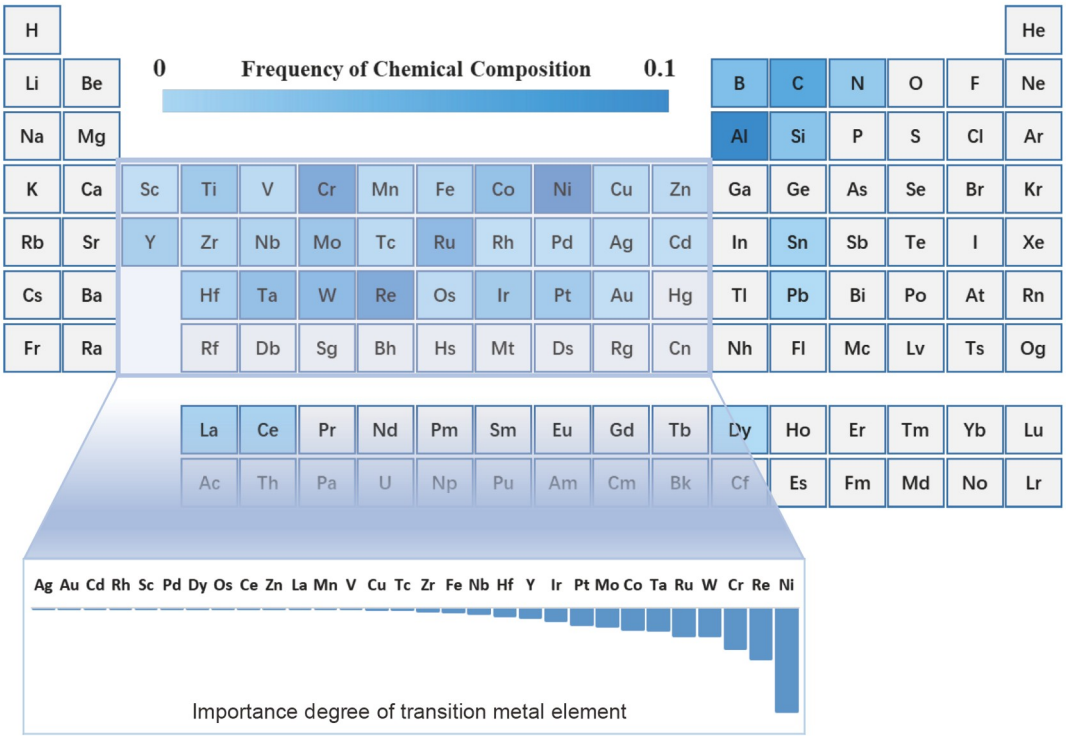


**Figure 12** (Color online) The research frequency of 38 chemical compositions in superalloys. The importance degree of transition metal element is normalized via Max Abs Normalization.

## 4 Conclusions

In the course of materials research on the structure-activity relationship of Nickel-based SC superalloys, a massive amount of scientific literature has been accumulated. It is of great significance to automatically extract valuable domain knowledge from the published unstructured text, which can help researchers design new materials. In this study, a text mining framework for Nickel-based SC superalloys based on distant supervision is proposed, aiming at domain knowledge discovery from credible abstracts. In order to guarantee the credibility of domain knowledge obtained from abstracts, MD-CEM is proposed to evaluate the text credibility from the perspective of source timeliness, publication authority, and author's academic standing. In the NER task, materials entity types and domain dictionary gain an obvious effect when labeling is unavailable. Experimental results show that the $F1$-Score of DS-NER is higher and achieves 87.24%. In CA-TEE, assisted by naming rules, reliable domain knowledge is discovered from fragmented information through screening, summarizing, and credibility analysis.

In conclusion, the proposed framework takes the text credibility into account, integrates the statistic-based models, dictionary, and rules for the NER task, and finally visualizes the domain knowledge of interest. In addition, the academic cooperative network of experts, research progress on Nickel-based SC superalloys, and chemical compositions in materials were successfully extracted. The follow-up work will be extended to low-content chemical elements, which also play a prominent role in superalloys. Note that the parts of this framework such as the definition of entity types, naming rules, and the model construction can be modified by the domain knowledge and thus extended to other fields. Overall, text mining provides valuable information on materials composition, processing, and properties, which can guide future R&D.

*This work was supported by the National Natural Science Foundation of China (Grant No. 52073169), the National Key Research and Development Program of China (Grant No. 2021YFB3802101), and the Key Research Project of Zhejiang Laboratory (Grant No. 2021PE0AC02). We also appreciate the High Performance Computing Center of Shanghai University, and the Shanghai Engineering Research Center of Intelligent Computing System for providing the computing resources and technical support.*

### Supporting Information

The supporting information is available online at tech.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

1 Mostafaei M, Abbasi S M. Designing and characterization of Al-and Ta-bearing Ni-base superalloys based on d-electrons theory. Mater Des, 2017, 127: 67–75
2 Mostafaei M, Abbasi S M. Prediction of incipient melting map and γ' features of Ni-base superalloys using molecular orbital method. In: TMS Annual Meeting & Exhibition. Cham: Springer, 2018. 453–466

3 Luo L, Ma Y, Li S, et al. Evolutions of microstructure and lattice misfit in a γ'-rich Ni-based superalloy during ultra-high temperature thermal cycle. Intermetallics, 2018, 99: 18–26
4 Krallinger M, Rabal O, Lourenço A, et al. Information retrieval and text mining technologies for chemistry. Chem Rev, 2017, 117: 7673–7761
5 Olivetti E A, Cole J M, Kim E, et al. Data-driven materials research enabled by natural language processing and information extraction. Appl Phys Rev, 2020, 7: 041317
6 Kononova O, He T J, Huo H Y, et al. Opportunities and challenges of text mining in materials research. iScience, 2021, 24: 102155
7 Eltyeb S, Salim N. Chemical named entities recognition: A review on approaches and applications. J Cheminform, 2014, 6: 17
8 Vaucher A C, Zipoli F, Geluykens J, et al. Automated extraction of chemical synthesis actions from experimental procedures. Nat Commun, 2020, 11: 3601–3611
9 Tarasova O A, Biziukova N Y, Rudik A V, et al. Extraction of data on parent compounds and their metabolites from texts of scientific abstracts. J Chem Inf Model, 2021, 61: 1683–1690
10 Kim E, Huang K, Saunders A, et al. Materials synthesis insights from scientific literature via text extraction and machine learning. Chem Mater, 2017, 29: 9436–9444
11 Jensen Z, Kim E, Kwon S, et al. A machine learning approach to zeolite synthesis enabled by automatic literature data extraction. ACS Cent Sci, 2019, 5: 892–899
12 Mahbub R, Huang K, Jensen Z, et al. Text mining for processing conditions of solid-state battery electrolytes. Electrochem Commun, 2020, 121: 106860
13 Tshitoyan V, Dagdelen J, Weston L, et al. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature, 2019, 571: 95–98
14 Weston L, Tshitoyan V, Dagdelen J, et al. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. J Chem Inf Model, 2019, 59: 3692–3702
15 He T J, Sun W H, Huo H Y, et al. Similarity of precursors in solid-state synthesis as text-mined from scientific literature. Chem Mater, 2020, 32: 7861–7873
16 Huo H Y, Rong Z Q, Kononova O, et al. Semi-supervised machine-learning classification of materials synthesis procedures. npj Comput Mater, 2019, 5: 62
17 Islamaj R, Leaman R, Kim S, et al. NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. Sci Data, 2021, 8: 91
18 Hawizy L, Jessop D M, Adams N, et al. ChemicalTagger: A tool for semantic text-mining in chemistry. J Cheminform, 2011, 3: 17
19 Swain M C, Cole J M. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. J Chem Inf Model, 2016, 56: 1894–1904
20 Mavračić J, Court C J, Isazawa T, et al. ChemDataExtractor 2.0: Autopopulated ontologies for materials science. J Chem Inf Model, 2021, 61: 4280–4289
21 Wang W R, Jiang X, Tian S H, et al. Automated pipeline for superalloy data by text mining. npj Comput Mater, 2022, 8: 9
22 Gu X Y. Study on quality index of bibliographic information (in Chinese). Library, 2007, 1: 73–75
23 Yu J R. Impact factor: Calculation, application, and limitations (in Chinese). Chin Bulletin Life Sci, 2002, 14: 2
24 Garfield E. Citation indexes for science: A new dimension in documentation through association of ideas. Int J Epidemiol, 2006, 35: 1123–1127
25 Kumar R, Singh S, Bilga P S, et al. Revealing the benefits of entropy weights method for multi-objective optimization in machining operations: A critical review. J Mater Res Tech, 2021, 10: 1471–1492
26 Wang Y, Guo J L. A comprehensive evaluation method for author influence based on grey relational analysis. J Intell, 2017, 36: 185–190+184
27 Kuznetsov O P. Complex networks and activity spreading. Autom Remote Control, 2015, 76: 2091–2109
28 Gleich D F. PageRank beyond the web. SIAM Rev, 2015, 57: 321–363
29 Yang Y S, Chen W L, Li Z H, et al. Distantly Supervised NER with Partial Annotation Learning and Reinforcement Learning. In: Proceedings of International Conference on Computational Linguistics, Santa Fe, 2018. 2159–2169
30 China National Committee for Terminology in Science and Technology. Chinese Terms in Materials Science and Technology (in Chinese). Beijing: Science Press, 2011. 1–199
31 Shi C Q, Tang M, Zhang D F, et al. Hash table based on Trie-tree. J Comput Appl, 2010, 30: 2193–2196
32 Shi C X, Zhong Z Y. Forty years of superalloy R&D in China (in Chinese). Acta Metallurgica Sinica, 1997, 33: 1–8
33 Yuan Y, Yan P, Zhuang J Y, et al. Classification and Designation for Superalloys and High Temperature Intermetallic Materials (in Chinese). Standards Press of China, 2005, GB/T 14992-2005
34 Chen X, Zou X Z, Qiu Y T. The application of resource discovery system in information tracing service for scientific research (in Chinese). Library Tribune, 2015, 5: 68–74,43
35 Zhang J, Wang L, Wang D, et al. Recent progress in research and development of Nickel-based single crystal superalloys (in Chinese). Acta Metall Sin, 2019, 55: 1077–1094
36 Shi Z X, Liu S Z, Yue X D, et al. Effect of Nb content on microstructure stability and stress rupture properties of single crystal superalloy containing Re and Ru. J Cent South Univ, 2016, 23: 1293–1300
37 Hiszpanski A M, Gallagher B, Chellappan K, et al. Nanomaterial synthesis insights from machine learning of scientific articles by extracting, structuring, and visualizing knowledge. J Chem Inf Model, 2020, 60: 2876–2887
38 Nie Z W, Liu Y J, Yang L Y, et al. Construction and application of materials knowledge graph based on author disambiguation: Revisiting the evolution of $LiFePO_4$. Adv Energy Mater, 2021, 11: 2003580
39 El-Bousiydy H, Lombardo T, Primo E N, et al. What can text mining tell us about lithium-ion battery researchers' habits? Batteries Supercaps, 2021, 4: 758–766
40 Chen B, Xie Y B. Functional knowledge integration of the design process. Sci China Tech Sci, 2017, 60: 209–218
41 Liu T Y, Zhang S, Wang Q, et al. Composition formulas of Ti alloys derived by interpreting Ti-6Al-4V. Sci China Tech Sci, 2021, 64: 1732–1740