Available online at www.sciencedirect.com
ScienceDirect
Trans. Nonferrous Met. Soc. China 35(2025) 2790−2802

Transactions of
Nonferrous Metals
Society of China
www.tnmsc.cn

DNMKG: A method for constructing domain of nonferrous metals knowledge graph based on multiple corpus

Hai-liang LI[1], Hai-dong WANG[2]

1. Youke Publishing Co., Ltd., Beijing 100088, China;
2. School of Minerals Processing and Bioengineering, Central South University, Changsha 410083, China

Received 30 June 2023; accepted 29 April 2025

Abstract: To address the underutilization of Chinese research materials in nonferrous metals, a method for constructing a domain of nonferrous metals knowledge graph (DNMKG) was established. Starting from a domain thesaurus, entities and relationships were mapped as resource description framework (RDF) triples to form the graph's framework. Properties and related entities were extracted from open knowledge bases, enriching the graph. A large-scale, multi-source heterogeneous corpus of over 1×10^9 words was compiled from recent literature to further expand DNMKG. Using the knowledge graph as prior knowledge, natural language processing techniques were applied to the corpus, generating word vectors. A novel entity evaluation algorithm was used to identify and extract real domain entities, which were added to DNMKG. A prototype system was developed to visualize the knowledge graph and support human−computer interaction. Results demonstrate that DNMKG can enhance knowledge discovery and improve research efficiency in the nonferrous metals field.

Key words: knowledge graph; nonferrous metals; thesaurus; word vector model; multi-source heterogeneous corpus

1 Introduction

The research on nonferrous metals is expanding rapidly in China, and a large number of research materials, such as academic literature [1,2] and web pages, are produced. These represent rich scientific knowledge. Researchers mainly search and use study materials obtained through search engines. Although search can help users find knowledge from a large amount of academic research data to a certain extent, this approach has limitations. The usual search method is based on mechanical matching of keyword strings, which often results in not retrieving literally mismatched content that is actually semantically related. Because the semantic level of the words entered by the user cannot be understood, searching lacks a semantic hierarchical association between the search result and the input word, thus missing a lot of valuable content. Because a large amount of different types of scientific research knowledge on materials is not stored and organized based on semantic associations, knowledge in the field of materials science cannot be recognized and understood by machines. However, automated knowledge discovery and knowledge integration can be pursued. Otherwise, the rapidly increasing knowledge cannot be effectively disseminated and utilized.

In the era of big data, researchers' need for knowledge has shifted from simply gathering information to more automated knowledge acquisition. Automatic or semi-automatic extraction of knowledge from vast amounts of data and information can transform information retrieval into knowledge mining and provide researchers with intelligent knowledge services.

Corresponding author: Hai-liang LI, Tel: +86-010-82241914, E-mail: 27401268@qq.com
https://doi.org/10.1016/S1003-6326(25)66848-8

A knowledge graph is an effective method to solve the above problems. A knowledge graph describes concepts, entities, and their relations in the objective world in a structured form. It is an important form of knowledge representation. The knowledge hidden within the massive amount of existing information is expressed by a knowledge graph in a way that is closer to human cognition. A knowledge graph also facilitates the ability to better organize, manage, and understand vast amounts of information. The knowledge graph concept was introduced by Google in 2012. In recent years, a variety of large-scale open domain knowledge maps have been constructed, such as DBPEDIA [3], YAGO [4], and FREEBASE [5,6] (Google Knowledge Graph). In China, there are Baidu Knowledge Graph and Sogou ZhiLiFang among others. These knowledge graphs follow the Resource Description Framework (RDF) [7] data model and contain hundreds of millions of entities and billions of facts (i.e., attribute values and relations with other entities), and these entities are organized in thousands of conceptual structures of the objective world. They have served extensively to improve search results, smart Q & A, and other business scenarios.

The characteristics of academic and scientific knowledge regarding nonferrous metals in materials science are as follows: (1) It is mainly in the form of documents, books, patents, and web pages with scattered sources and inconsistent content systems; (2) The amount of research data is increasing, so it is difficult for ordinary retrieval methods to effectively extract knowledge from the massive amount of total information. Inadequate large-scale knowledge mining of materials science research has resulted in a small size of the existing knowledge graph.

In this paper, a general method to construct a large-scale knowledge graph from the research materials of nonferrous metals was proposed. The method is named as DNMKG (domain of nonferrous metals knowledge graph). DNMKG can be built as a big and comprehensive knowledge graph with the help of semantic techniques. The contributions of this paper are as follows.

(1) An effective method for building the framework of a nonferrous metals knowledge graph from a domain thesaurus was proposed. Based on this framework, entities and relations were extracted from the open knowledge base and stored into DNMKG in the form of RDF triples.

(2) To enrich the entities and relations in the DNMKG, a multi-source heterogeneous nonferrous metals corpus was constructed, including papers, reference books, and monographs, which together comprise more than 1×10^9 words.

(3) A pipeline based on natural language processing (NLP) was constructed to preprocess the corpus and generate word vector models through calculation. Then, a proposed entity evaluation algorithm was used to extract the related entities from the corpus. Finally, a prototype system was built to demonstrate the interface between researchers and the knowledge graph.

2 Related works

2.1 Materials science knowledge graph

A knowledge graph is a multi-relational graph composed of entities (nodes) and relations (different types of edges). The entities express an existing concept in the world. The relations show the connections between two entities and are usually represented as a triple of the form (head entity, relation and tail entity). For example, <Beijing, isCapitalOf, China> is a relation in the triple form.

In addition to open-domain knowledge graphs such as DBpedia and Freebase, knowledge graphs in many professional fields have been constructed. Knowledge graphs have been developed rapidly in bioinformatics [8,9], and many large-scale usable systems have been built in that domain. However, in the field of materials science, including nonferrous metals, the progress of knowledge graphs has been relatively slow. Some works based on XML Schema have attempted to integrate materials information. For example, MATML [10] aims to facilitate exchange of materials properties information, and JRC-MATDB [11] has been built for interoperability of engineering materials testing databases. The disadvantage of XML is that its relationship types are very few, making it challenging to sufficiently capture and reflect the semantic information. Some other knowledge graphs are based on ontology [12]. Ontology models concepts and their relations, whereas knowledge graphs model entities and their relations according to concepts. The abstraction level of ontology is higher than that of knowledge graph. MatSeek [13] was designed for knowledge representation of materials science, integration of materials databases, modeling of origin data, and extraction of new knowledge through reasoning. SLACK [14] is based on a set of ontologies for laminate composite materials and designed for manufacturing and integrates them into a previously developed engineering design framework. The existing efforts mostly rely on manual construction by domain experts i.e., the classification is more accurate, but the time taken to build is longer and the update frequency is slower.

2.2 Proposed approach for knowledge graph construction

At present, the methods for constructing domain knowledge graphs are based on the use of a thesaurus, domain-independent open knowledge bases, and large-scale corpora. QIAO et al [15] established maps from an agricultural thesaurus to an agricultural knowledge graph schema layer and a data layer using rules for determining whether the thesaurus entry is a concept or an entity. ZHANG et al [16] built a metallic materials knowledge graph based on DBpedia and Wikipedia through algorithmic extraction of entities based on semantic similarity. WANG et al [17] used machine learning algorithms to auto-populate domain ontologies from online encyclopedias to compensate for the lack of a thesaurus description of the semantic relation between terms. The research of automated methods often stops at constructing knowledge graphs only from a thesaurus or only from an online encyclopedia from which entities and attributes can be extracted and fails to make full use of a large amount of domain-specific corpora, such as research materials.

3 Construction of DNMKG

3.1 Overview of approach

In this paper, we designed a method to extract entities and their properties and relations from a nonferrous metals thesaurus, open knowledge bases, and the scientific literature corpus. The knowledge was represented as RDF triples, which together formed the knowledge graph. The knowledge graph was stored in the Neo4j database, which is a graph-based database, resulting in a large-scale triple-based semantic knowledge base. Figure 1 shows the overview of this approach.

Definition 1: Knowledge graph

An actual statement can be expressed as a semantic web triple <subject, predicate, object>, which can formalize semantic data into a knowledge triple set K, $K \subseteq S \times P \times O$, where S is the subject set comprising entities, O is the object set comprising entities or entities' properties, and P is the predicates set usually comprising relations between entities or between entities and properties. A knowledge graph is the set of the triples, so it can be represented as DNMKG=$\{E, P, R\}$, where $E=\{e|e$ is the entity of nonferrous metals$\}$, $P=\{p|p$ is the property of the entity$\}$, and $R=\{r|r$ is the relation of e and $p\}$.

Step 1 We obtained and digitized thesaurus books related to nonferrous metals. Using the thesaurus as an entity, with various types of relations between thesauruses, we mapped them

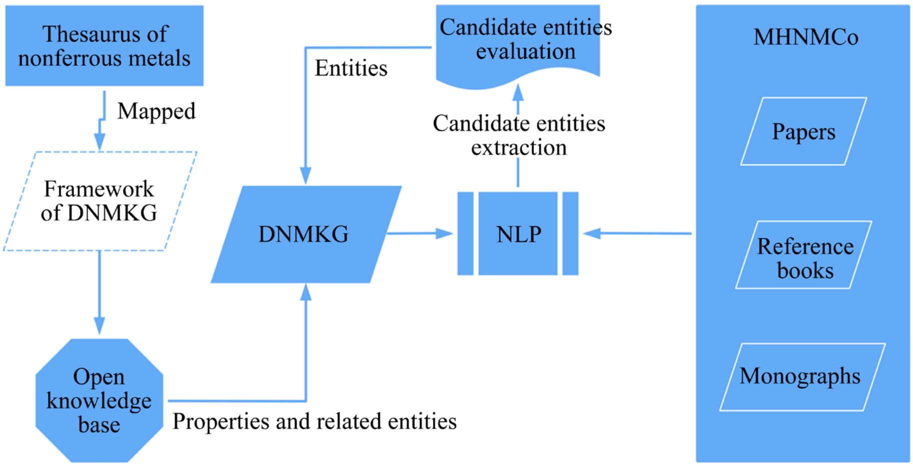

Thesaurus of nonferrous metals
Mapped
Framework of DNMKG
Open knowledge base
Properties and related entities
Entities
DNMKG
Candidate entities evaluation
Candidate entities extraction
NLP
MHNMCo
Papers
Reference books
Monographs

Fig. 1 Overview of construction of DNMKG

into RDF triples and stored them in the form of graph-based data to establish the framework of DNMKG.

Step 2 Using the entities of the nonferrous metals thesaurus as seed words, the entity attributes and related entities were extracted from multiple open knowledge bases. The obtained entities, attributes, and relations were converted into triples, which were merged and stored in DNMKG.

Step 3 The scientific literature related to the field of nonferrous metals was collected, including papers, reference books, monographs, and textbooks. These materials were classified and stored to establish a large multi-source heterogeneous nonferrous metals corpus (MHNMCo).

Step 4 Through a series of natural language processing and semantic analysis processes, MHNMCo was processed to extract candidate entities for nonferrous metals. Combined with the nonferrous metals word vector model, we applied a proposed candidate entity evaluation method to obtain a more complete DNMKG through multiple iterations.

3.2 Building knowledge graph based on thesaurus

3.2.1 Understanding structure of thesaurus

A thesaurus is a collection of these concepts and their relations and thus contains rich semantic relations. It was first used in traditional literature indexing. There are three kinds of relations in a thesaurus, namely equivalence, hierarchical, and associative relations.

The equivalence relation is also known as the identity relation: Y (use) and D (alternate). It includes synonymous and alternative relations that imply the same concept or the same usage.

The hierarchical relation is also known as the subordinate relation: S (genus), F (sub) and Z (family head). This relation includes genus-species, whole-part, and multi-level relations. The hyponym of each level must be the same as the conceptual type of the hypernym and belong to the same category.

The associative relation is also known as the phylogenetic relation: C (reference). It enables indexing and retrieval. The associative relation is important for expanding the scope of concepts.

A thesaurus has a clear semantic structure and can be used to build a knowledge graph. Some researchers in other fields have already done related work. YUAN et al [18] compared existing ontology construction methods and proposed a thesaurus-based oil field ontology construction method. GU et al [19] built a Chinese ancient medical literature ontology. However, there is no relevant research in the field of nonferrous metals.

3.2.2 Building thesaurus of nonferrous metals domain

The *China Thesaurus for Nonferrous Metals Industry* [20] is the first large-scale comprehensive retrieval reference book in the field of nonferrous metals in China. It is also a reference book for standardizing the terminology of the nonferrous metals industry. It serves as bridge between the natural language used in written literature and the standardized language used in the knowledge base system. It contains 14224 entries in 29 categories, including nonferrous metal mining, mineral processing, smelting, pressure processing, metal materials, powder metallurgy, metallurgy and heat treatment, metal corrosion and protection, analytical testing, metallurgical automation, energy, metallurgical construction, technical economy, environmental protection, and labor safety. We digitized this book to enable processing by computational methods.

A typical example of a nonferrous metals thesaurus entry is as follows:

Tungsten steels

S	Alloy steels
Z	Steel

WCdCo alloys

S	Tungsten alloys
Z	Refractory metal alloys

Tungsten alloys

D	Tungsten containing alloys
S	Transition metal alloys
S	Refractory metal alloys
F	Tungsten base alloys
F	Tungsten additions
F	WCdCo alloys

3.2.3 Transforming thesaurus into knowledge graph

The foundation of the knowledge graph is the entities and their relationships. The thesaurus already contains these elements. Through processing of the thesaurus, we can acquire the concepts and entities related to the nonferrous metals discipline and the relationships between them, which can be used to build a knowledge graph.

In this paper, we used "is" to indicate the equivalence relation, "isA" to indicate the hierarchical relation, and "related" to indicate the associative relation. We used the following process: Thesaurus−Relationship annotation−Relationship mapping−Building RDF triples—Storing the triples. The relation mapping is shown in Table 1. After the transformation, thousands of triples were generated and stored in the graph-based database Neo4j. This is the framework of DNMKG. Table 1 shows examples of entities and relations. Table 2 shows the 29 categories of entities in the framework of DNMKG. A classical example of framework of DNMKG is shown in Fig. 2.

Table 1 Transformation of thesaurus to RDF triples

Relations in thesaurus	Relation mapping	Before transformation	After transformation (RDF triples)
Equivalence relation (Y, D)	is	Tungsten alloys D Tungsten containing alloys	⟨Tungsten alloys, is, Tungsten containing alloys⟩
Hierarchical relation (S, F, Z)	isA	Tungsten alloys F tungsten base alloys	⟨Tungsten alloys, isA, Tungsten base alloys⟩
Associative relation (C)	related	Coating C Metal surface protection	⟨Coating, relate, Metal surface protection⟩

3.3 Enriching knowledge graph using open knowledge bases

3.3.1 Open knowledge bases

At present, there are some knowledge bases that can be easily accessed online (wikis), such as DBpedia, YAGO, and FreeBase. In China, Baidu Baike, Hudong Baike, and Chinese Wikipedia (zh.wikipedia.org) are the three largest encyclopedia sites. Generally, a page of online encyclopedia corresponds to an entity. This page often contains a lot of information, such as the entity's name, an abstract that summarizes the most important information, the description (text that provides detailed information in various sections of the page), links (references to other pages), an infobox (structured information about the page in a table format), and a category (the topic of the page). The attributes and related entities can be extracted from the online encyclopedia pages [21]. A typical page is shown in Fig. 3.

Table 2 29 categories of entities in framework of DNMKG

Category	Number of entity
Mining	2129
Mineral processing	2141
General metallurgical issues	1393
Metallurgy	1242
Powder metallurgy	113
Nonferrous metals smelting	1653
Metallography	651
Performance of metal	882
Physical analysis tests on metals	139
Heat treatment	228
Metal corrosion and protection, surface treatment	578
Metal casting	258
Metal welding, cutting, bonding	300
Metal pressure working	853
Metals and alloys	1020
Alloyology, various alloys	315
Metal material	113
Carbon	113
Refractories	485
Chemical testing of metals	423
Metallurgical automation	196
Computer technology	22
Environmental science	501
Industrial safety, labor protection	473
Industrial economy	857
Library and information work	193
Elements and compounds	863
Generic entries	2504
Test equipment and instruments	88
Total	20726*

*As an entity can be classified into multiple categories, this total exceeds 14224, which was the actual total number of entities

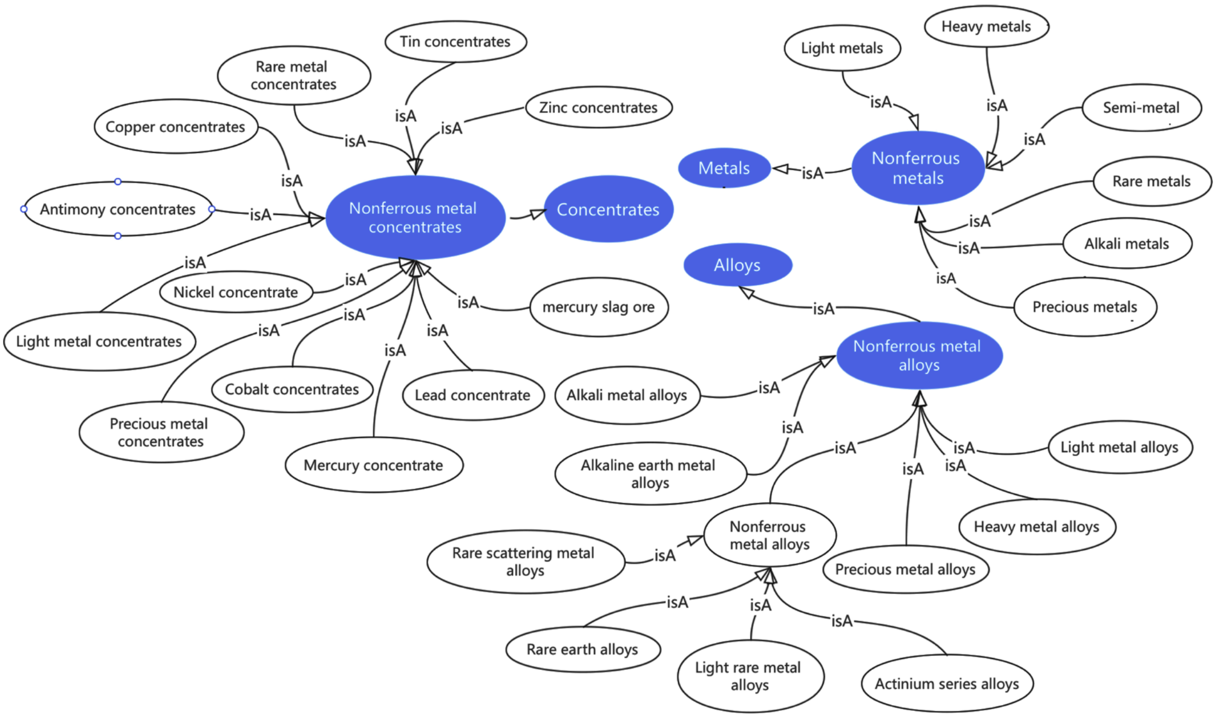


Fig. 2 Classical example of framework of DNMKG

WANG et al [22] presented an English–Chinese bilingual knowledge graph named XLore, which harvested 2466956 concepts, 16284901 instances, and 446236 relations from four online wikis, namely English Wikipedia, Chinese Wikipedia, Baidu Baike, and Hudong Baike. In this paper, XLore was used to complete the properties and relations and extract more related entities of DNMKG.

3.3.2 Extracting entities and relations from knowledge using DNMKG

The entities in the framework of DNMKG were matched to the entities in XLore one by one. Their attributes, classifications, and related entities were extracted from the matched entity page. Then, triples were generated and stored in DNMKG. Figure 4 shows an example of an entity's enrichment. Through this stage of processing, DNMKG was substantially expanded. The number of entities increased, attributes became richer, and relations were strengthened.

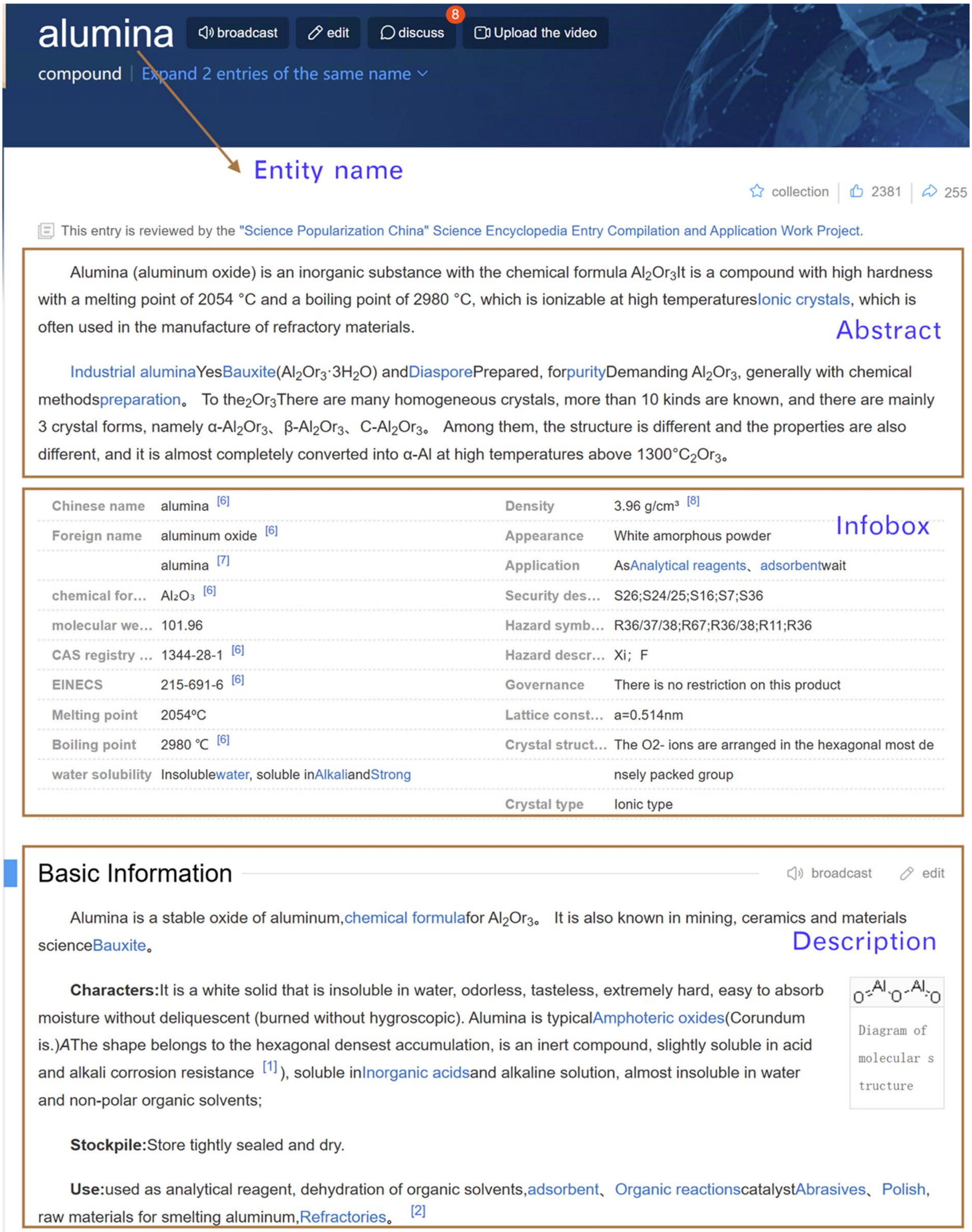
4 Enriching DNMKG using large-scale text corpus

The above discussion showed how the framework of DNMKG was built using a thesaurus. Next, entities and related properties were supplemented using open knowledge bases. However, there are two shortcomings: (1) because of the low frequency of updates to the nonferrous metals thesaurus, new concepts and terms were absent, which resulted in a limited number of entities extracted from the open knowledge bases; (2) Online open knowledge bases are usually intended for public use, and therefore containing limited specialized knowledge of professional fields. This is especially true for the nonferrous metals domain, which also results in extracting a limited number of entities. To address these limitations, we developed a method to use a large-scale corpus to enrich DNMKG.

4.1 MHNMCo: Multi-source heterogeneous nonferrous metals corpus

Scientific literature in the field of nonferrous metals includes research papers, reference books, and monographs, among other materials. The content in the literature is relatively new and updated quickly. It can be considered the de facto standard corpus for knowledge mining.

One of the important contributions of this article is the MHNMCo that we have built for enriching DNMKG (Fig. 5). We gathered thousands



This entry is reviewed by the "Science Popularization China" Science Encyclopedia Entry Compilation and Application Work Project.

Alumina (aluminum oxide) is an inorganic substance with the chemical formula Al_2Or_3It is a compound with high hardness with a melting point of 2054 °C and a boiling point of 2980 °C, which is ionizable at high temperaturesIonic crystals, which is often used in the manufacture of refractory materials.

Abstract

Industrial aluminaYesBauxite($Al_2Or_3 \cdot 3H_2O$) andDiasporePrepared, forpurityDemanding Al_2Or_3, generally with chemical methodspreparation。 To the$_2Or_3$There are many homogeneous crystals, more than 10 kinds are known, and there are mainly 3 crystal forms, namely α-Al_2Or_3、β-Al_2Or_3、C-Al_2Or_3。 Among them, the structure is different and the properties are also different, and it is almost completely converted into α-Al at high temperatures above 1300°C_2Or_3。

Infobox

Chinese name	alumina [6]	Density	3.96 g/cm³ [8]
Foreign name	aluminum oxide [6]	Appearance	White amorphous powder
	alumina [7]	Application	AsAnalytical reagents、adsorbentwait
chemical for...	Al_2O_3 [6]	Security des...	S26;S24/25;S16;S7;S36
molecular we...	101.96	Hazard symb...	R36/37/38;R67;R36/38;R11;R36
CAS registry ...	1344-28-1 [6]	Hazard descr...	Xi； F
EINECS	215-691-6 [6]	Governance	There is no restriction on this product
Melting point	2054°C	Lattice const...	a=0.514nm
Boiling point	2980 ℃ [6]	Crystal struct...	The O2- ions are arranged in the hexagonal most de nsely packed group
water solubility	Insolublewater, soluble inAlkaliandStrong		
		Crystal type	Ionic type

Basic Information

Alumina is a stable oxide of aluminum,chemical formulafor Al_2Or_3。 It is also known in mining, ceramics and materials scienceBauxite。

Description

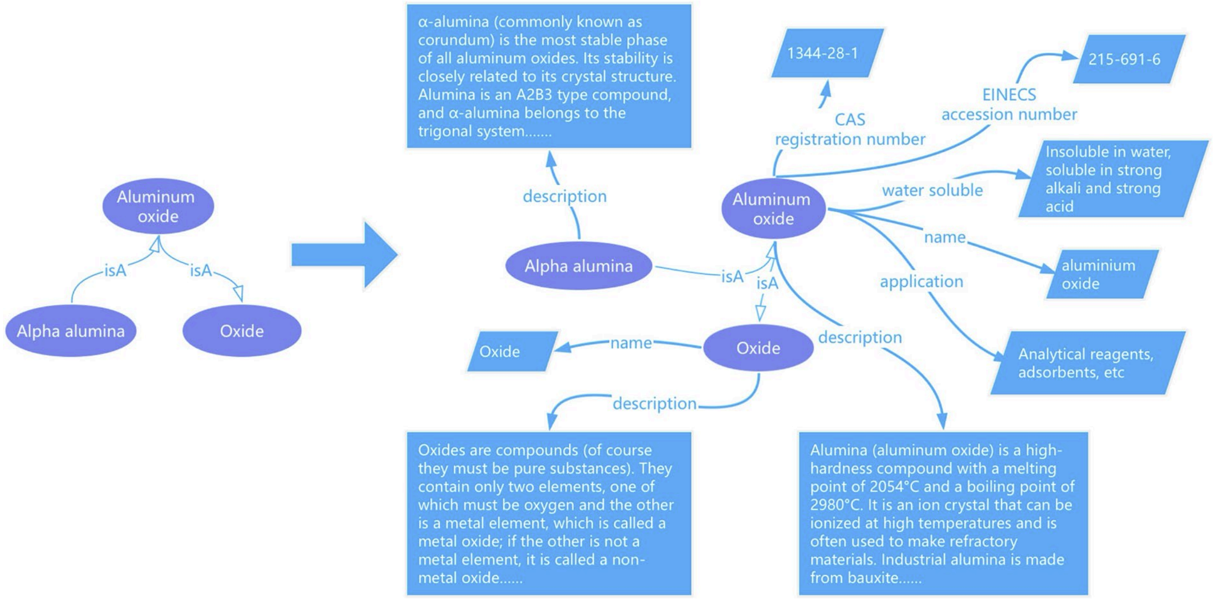
Characters: It is a white solid that is insoluble in water, odorless, tasteless, extremely hard, easy to absorb moisture without deliquescent (burned without hygroscopic). Alumina is typicalAmphoteric oxides(Corundum is.)*A*The shape belongs to the hexagonal densest accumulation, is an inert compound, slightly soluble in acid and alkali corrosion resistance [1]), soluble inInorganic acidsand alkaline solution, almost insoluble in water and non-polar organic solvents;

Stockpile: Store tightly sealed and dry.

Use:used as analytical reagent, dehydration of organic solvents,adsorbent、Organic reactionscatalystAbrasives、Polish, raw materials for smelting aluminum,Refractories。 [2]

Fig. 3 Typical page of Baidu encyclopedia (https://baike.baidu.com/item/%E6%B0%A7%E5%8C%96%E9%93%9D/2849623)

of documents from the last five years from 167 nonferrous metals journals with the China National Knowledge Infrastructure (CNKI). Dozens of reference books and monographs were also digitized into structured or semi-structured form, including professional dictionaries, encyclopedias, metallographic maps, and handbooks, among others. With respect to research papers, metadata

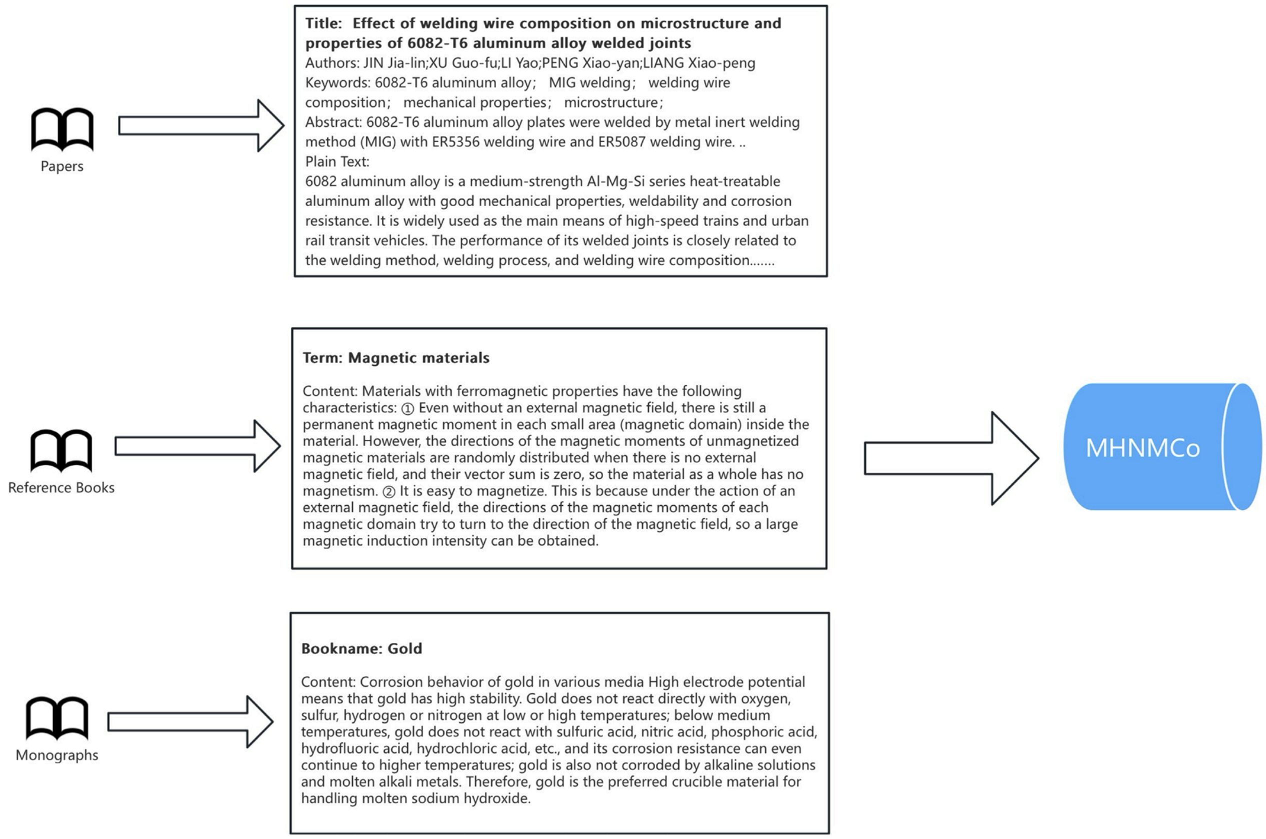


α-alumina (commonly known as corundum) is the most stable phase of all aluminum oxides. Its stability is closely related to its crystal structure. Alumina is an A2B3 type compound, and α-alumina belongs to the trigonal system.......

Oxides are compounds (of course they must be pure substances). They contain only two elements, one of which must be oxygen and the other is a metal element, which is called a metal oxide; if the other is not a metal element, it is called a non-metal oxide......

Alumina (aluminum oxide) is a high-hardness compound with a melting point of 2054°C and a boiling point of 2980°C. It is an ion crystal that can be ionized at high temperatures and is often used to make refractory materials. Industrial alumina is made from bauxite......

Fig. 4 Example of entity enrichment



Title: Effect of welding wire composition on microstructure and properties of 6082-T6 aluminum alloy welded joints
Authors: JIN Jia-lin;XU Guo-fu;LI Yao;PENG Xiao-yan;LIANG Xiao-peng
Keywords: 6082-T6 aluminum alloy; MIG welding; welding wire composition; mechanical properties; microstructure;
Abstract: 6082-T6 aluminum alloy plates were welded by metal inert welding method (MIG) with ER5356 welding wire and ER5087 welding wire. ..
Plain Text:
6082 aluminum alloy is a medium-strength Al-Mg-Si series heat-treatable aluminum alloy with good mechanical properties, weldability and corrosion resistance. It is widely used as the main means of high-speed trains and urban rail transit vehicles. The performance of its welded joints is closely related to the welding method, welding process, and welding wire composition.......

Term: Magnetic materials

Content: Materials with ferromagnetic properties have the following characteristics: ① Even without an external magnetic field, there is still a permanent magnetic moment in each small area (magnetic domain) inside the material. However, the directions of the magnetic moments of unmagnetized magnetic materials are randomly distributed when there is no external magnetic field, and their vector sum is zero, so the material as a whole has no magnetism. ② It is easy to magnetize. This is because under the action of an external magnetic field, the directions of the magnetic moments of each magnetic domain try to turn to the direction of the magnetic field, so a large magnetic induction intensity can be obtained.

Bookname: Gold

Content: Corrosion behavior of gold in various media High electrode potential means that gold has high stability. Gold does not react directly with oxygen, sulfur, hydrogen or nitrogen at low or high temperatures; below medium temperatures, gold does not react with sulfuric acid, nitric acid, phosphoric acid, hydrofluoric acid, hydrochloric acid, etc., and its corrosion resistance can even continue to higher temperatures; gold is also not corroded by alkaline solutions and molten alkali metals. Therefore, gold is the preferred crucible material for handling molten sodium hydroxide.

Fig. 5 Example of building MHNMCo

(titles, authors, keywords, abstracts, etc.) were extracted. For encyclopedias or dictionaries, the terms and their explanations were extracted separately. Structured and semi-structured corpora can express more semantic information and are more amenable to subsequent processing.

A large amount of plain text content was extracted and cut into sentences and then stored in a database for processing. Table 3 provides an overview of MHNMCo.

Table 3 Overview of MHNMCo

Type	Source	Number of documents	Number of words
Papers	CNKI	242222	1067410000
Reference books and handbooks	www.ukus.com.cn	93	77900000
Monographs	www.ukus.com.cn	28	13650000
Total			1158960000

4.2 Extracting candidate entities

For MHNMCo, a pipeline based on NLP was constructed to create a more extensive network of inter-word relations, which enabled richer knowledge of nonferrous metals entities as a supplement. Figure 6 shows the workflow.

4.3 Corpus preparation

The first step of text corpus processing is word segmentation. Professional domain dictionaries can aid NLP to segmenting words more accurately. We used a nonferrous metals domain dictionary, which contained a collection of domain specific nouns. We added the research paper keywords and reference book items extracted from MHNMCo into the dictionary.

Firstly, we performed word segmentation processing on the documents in MHNMCo. A large-scale Chinese dictionary was constructed, which contained the research paper keywords and reference book items extracted from MHNMCo to ensure the accuracy of the word segmentation. Secondly, the corpus text was labeled with parts of speech. The verbs, adverbs, and nouns in the sentences were identified, but only kept all nouns. Stop words were also removed. We used Stanford University's CoreNLP [23] to preprocess all the documents and sentences. It is a widely used integrated NLP toolkit that supports Chinese language. After the processing, the corpus was used to calculate the word vector model.

4.4 Calculation of word vector model

Word vector technique converts words in a corpus into dense vectors. Words with similar semantics have similar vector representations. There are several ways to generate word vectors starts, including a statistics-based language model and a neural network-based language model. Some of the classical language models are Word2Vec, GloVe, ELMo, and BERT. We used the Word2Vec model.

Word2vec [24] is a word vector representation with a supervised method. Word2vec provides two neural network models, namely continuous bag-of-words and skip-gram, for computing continuous vector representations of words from very large data sets. We adopted skip-gram. Skip refers to the words in a certain window. Even if they are separated by some words, the probability of their co-occurrence can still be calculated.

In the training stage, if the processed corpus is regarded as a sequence of words w_1, w_2, ⋯, w_T, the goal of the skip-gram model during training is to maximize the average value of the logarithm of the probability $p(w_{j+1}\,|\,w_t)$ of co-occurrence:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-k\le j\le k}\lg p(w_{j+1}\,|\,w_t) \tag{1}$$

where k is the size of the training window, which is the distance between the words before and after the current word, and in the following work, we set

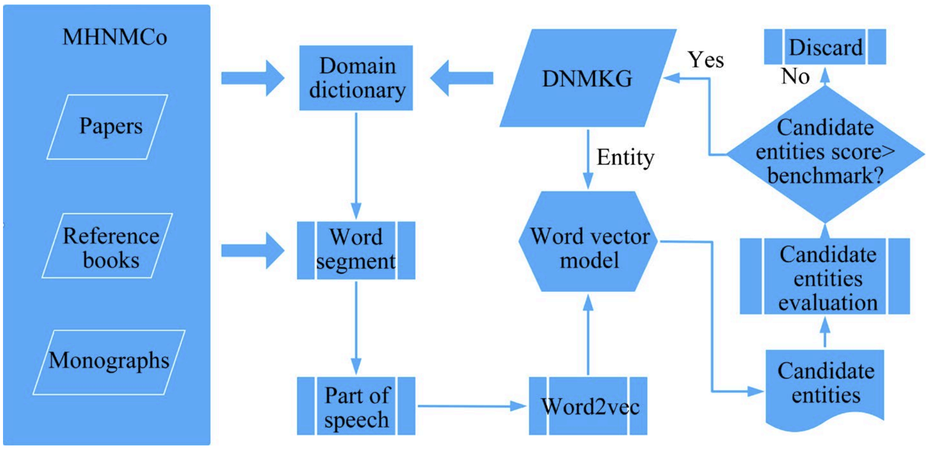


Fig. 6 Pipeline for enriching domain of DNMKG based on MHNMCo

k=10; T is the total number of entities.

In the prediction stage, the skip-gram model associates two parameter vectors $\boldsymbol{u}_w$ and $\boldsymbol{v}_w$ for learning for each input w. The two parameter vectors are the input and output vectors of w, respectively. For a given word w_j, the probability that the word w_i is correctly predicted is as follows:

$$p(w_i|w_j)=\frac{\exp(\boldsymbol{u}_{w_i}^{\mathrm{T}}\boldsymbol{v}_{w_j})}{\sum_{i=1}^{V}\exp(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{v}_{w_j})} \tag{2}$$

where V is the number of words in the dictionary, and T denotes the transpose of a vector.

We used gensim [25] as our word2vec software. The processed corpus from the previous step was input into gensim. After training, we obtained the neural network model as the output.

4.5 Extraction of candidate entities

Using the nonferrous metals corpus and the word vector model, we built a candidate entity extraction algorithm to identify real entities. With the entities in DNMKG as input, the related entities and their relevance were calculated through the word vector model. Some of the entities appeared in DNMKG with the label "Entity", and the others were candidate entities labeled "Candidate Entity". We calculated the average relevance between these entities and the input entity and used it as the benchmark to determine whether the candidate entity is a real entity. Figure 7 shows a sample view of the calculation result. It is an undirected weighted graph.

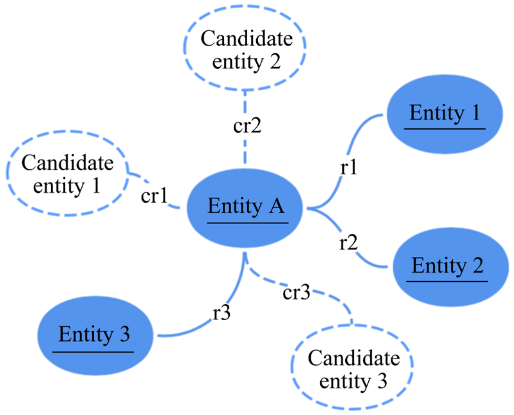


Fig. 7 Sample view of entity and its candidate entities

Definition 2: Entities and candidate entities set

$$\text{EntitySet}(e_\text{a})=\{e_1, e_2, e_3, \cdots|r_1, r_2, r_3, \cdots\} \tag{3}$$

$$\text{CandidateEntitySet}(e_\text{a})=\{ce_1, ce_2, ce_3, \cdots|cr_1, cr_2, cr_3, \cdots\} \tag{4}$$

where e_a refers to the input entity from DNMKG; e_1, e_2 and e_3 refer to the entities calculated from word vector model; r_1, r_2 and r_3 are the respective probabilities between e and e_a; ce_1, ce_2 and ce_3 refer to the candidate entities calculated from the word vector model, and cr_1, cr_2 and cr_3 are the respective probabilities between ce and e_a. We believe that compared with setting a constant as the benchmark, this average value dynamically reflects the minimum correlation that the current entity and its related entities should have. The correlation of candidate entities must be greater than the benchmark to be considered a true entity.

Definition 3: Benchmark of entity's relevance

$$\text{Benchmark}(e)=\sum_{i=1}^{T}r_i\bigg/T \tag{5}$$

where Benchmark(e) is the average of the relevance between these entities and the input entity.

Definition 4: Candidate entities evaluation

$$\text{EntityEvaluation}(e,cr)=\begin{cases}\text{True}, & cr\geq\alpha\cdot\text{Benchmark}(e)\\ \text{False}, & cr<\alpha\cdot\text{Benchmark}(e)\end{cases} \tag{6}$$

where α is an adjustment factor. We set α=1. If a candidate entity's relevance is greater than $\alpha\cdot$Benchmark(e), it is an entity, otherwise it is not and entity.

The detailed steps are as follows:

Step 1 Extract an entity from DNMKG.

Step 2 Calculate the top 30 entities with the highest correlation with the entity using the word vector model, retain the correlation, and calculate Benchmark(e) and EntityEvaluation(e,cr). Filter out new entities.

Step 3 Extract the properties and related entities of the new entity from the open knowledge bases and store them in DNMKG.

Step 4 Repeat Steps 1 to 3 until all entities in DNMKG are traversed.

5 Experiment and discussion

After applying our proposed pipeline to the thesaurus of nonferrous metals and MHNMCo, DNMKG had 51852 distinct entities and 226857 RDF triples.

Table 4 lists the number of entities and the number of RDF triples in each stage during the construction of DNMKG.

Firstly, the more complex part of constructing the knowledge graph was the discovery and recognition of domain concepts and entities. With the integration of domain expert knowledge, the thesaurus could find the domain categories quickly and accurately. With the thematic vocabulary, a classification system of the entire nonferrous metals knowledge map was established, which covered the nonferrous metals academic and scientific literature and the main entity categories of the industry. Of the total number of entities, 17.4% were extracted, and 12.8% of triples were extracted from it. The framework of DNMKG was thus formed, and it was more conducive for knowledge extraction in the subsequent processing stage. We consider that it is an effective approach to construct the domain knowledge map from the thesaurus.

Table 4 Overview of DNMKG

Type	Number of thesaurus	Number of thesaurus + open knowledge graph	Number of thesaurus + open knowledge graph + MHNMCo
Entities	14203	37316	81852
RDF triples	29074	108714	226857

Next, the framework of DNMKG was supplemented with the existing open knowledge bases. At this stage, the number of entities was expanded by 162.7% compared with the previous stage and accounted for 45.6% of the total number of entities. In particular, almost all entity attributes were extracted through the open knowledge bases. A large number of related relationships were extracted, with other types were relatively few. This is because the subordinate relationships and the equivalence relationships were not established in the knowledge bases.

Finally, the entities and relationships for semi-structured and unstructured corpora were extracted to further strengthen the knowledge graph. We sorted large-scale digital corpora, which is a highlight of this work. At this stage, the DNMKG constructed in this work was used as a priori knowledge. The number of related entities extracted increased by 119.3% compared with before, and the proportion of new entities added reached 54.4%, which represents a good extraction result.

A prototype system was constructed to demonstrate the human−computer interaction form of DNMKG (Fig. 8). When the user enters a term, the network connections of this term in DNMKG will appear on the left, and the entities that are semantically related to it can be clearly seen. This interface can help researchers quickly understand the relevant knowledge hierarchy. At the same time, knowledge resources closely related to the term will be retrieved on the right. They are from MHNMCo and can guide users to further reading and learning.

6 Conclusions and future work

DNMKG is a semantic network of knowledge related to nonferrous metals, mineral processing, mining, smelting, and many other related areas. It provides researchers and practitioners with an effective method for the organization, management, and retrieval of large-scale professional content and improves the efficiency of acquiring research intelligence. This form of the knowledge map can be computerized and compared with other fields of knowledge mapping to facilitate interactive sharing and integration, enabling better organization and dissemination of multi-domain knowledge for researchers in the field of nonferrous metals.

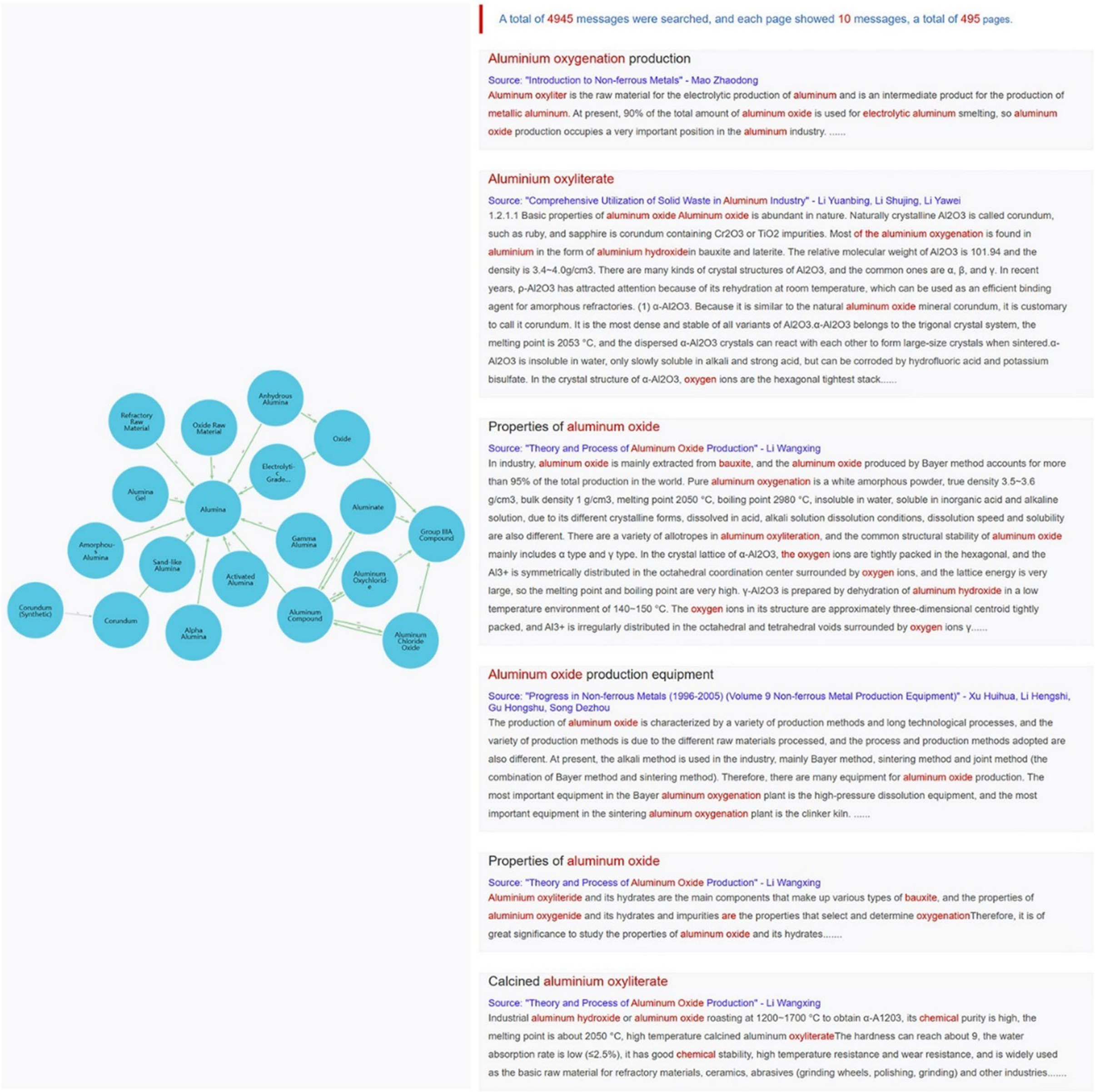
In the future, we plan to improve two aspects: (1) to develop an automated processing mechanism that can discover new entities, their attributes, and relationships on a timely basis and dynamically enrich DNMKG; (2) to further expand the types of entity relationships, such as those of alloy materials, to meet the needs of materials design, analysis, manufacturing, selection, and other scenarios in the field of metals.

CRediT authorship contribution statement

Hai-liang LI: Conceptualization, Methodology, Software, Data curation, Writing − Original draft, Visualization, Investigation; **Hai-dong WANG:** Writing − Reviewing and editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.



A total of 4945 messages were searched, and each page showed 10 messages, a total of 495 pages.

Aluminium oxygenation production

Source: "Introduction to Non-ferrous Metals" - Mao Zhaodong

Aluminum oxyliter is the raw material for the electrolytic production of aluminum and is an intermediate product for the production of metallic aluminum. At present, 90% of the total amount of aluminum oxide is used for electrolytic aluminum smelting, so aluminum oxide production occupies a very important position in the aluminum industry. ......

Aluminium oxyliterate

Source: "Comprehensive Utilization of Solid Waste in Aluminum Industry" - Li Yuanbing, Li Shujing, Li Yawei

1.2.1.1 Basic properties of aluminum oxide Aluminum oxide is abundant in nature. Naturally crystalline Al2O3 is called corundum, such as ruby, and sapphire is corundum containing Cr2O3 or TiO2 impurities. Most of the aluminium oxygenation is found in aluminium in the form of aluminium hydroxidein bauxite and laterite. The relative molecular weight of Al2O3 is 101.94 and the density is 3.4~4.0g/cm3. There are many kinds of crystal structures of Al2O3, and the common ones are α, β, and γ. In recent years, ρ-Al2O3 has attracted attention because of its rehydration at room temperature, which can be used as an efficient binding agent for amorphous refractories. (1) α-Al2O3. Because it is similar to the natural aluminum oxide mineral corundum, it is customary to call it corundum. It is the most dense and stable of all variants of Al2O3.α-Al2O3 belongs to the trigonal crystal system, the melting point is 2053 °C, and the dispersed α-Al2O3 crystals can react with each other to form large-size crystals when sintered.α-Al2O3 is insoluble in water, only slowly soluble in alkali and strong acid, but can be corroded by hydrofluoric acid and potassium bisulfate. In the crystal structure of α-Al2O3, oxygen ions are the hexagonal tightest stack......

Properties of aluminum oxide

Source: "Theory and Process of Aluminum Oxide Production" - Li Wangxing

In industry, aluminum oxide is mainly extracted from bauxite, and the aluminum oxide produced by Bayer method accounts for more than 95% of the total production in the world. Pure aluminum oxygenation is a white amorphous powder, true density 3.5~3.6 g/cm3, bulk density 1 g/cm3, melting point 2050 °C, boiling point 2980 °C, insoluble in water, soluble in inorganic acid and alkaline solution, due to its different crystalline forms, dissolved in acid, alkali solution dissolution conditions, dissolution speed and solubility are also different. There are a variety of allotropes in aluminum oxyliteration, and the common structural stability of aluminum oxide mainly includes α type and γ type. In the crystal lattice of α-Al2O3, the oxygen ions are tightly packed in the hexagonal, and the Al3+ is symmetrically distributed in the octahedral coordination center surrounded by oxygen ions, and the lattice energy is very large, so the melting point and boiling point are very high. γ-Al2O3 is prepared by dehydration of aluminum hydroxide in a low temperature environment of 140~150 °C. The oxygen ions in its structure are approximately three-dimensional centroid tightly packed, and Al3+ is irregularly distributed in the octahedral and tetrahedral voids surrounded by oxygen ions γ......

Aluminum oxide production equipment

Source: "Progress in Non-ferrous Metals (1996-2005) (Volume 9 Non-ferrous Metal Production Equipment)" - Xu Huihua, Li Hengshi, Gu Hongshu, Song Dezhou

The production of aluminum oxide is characterized by a variety of production methods and long technological processes, and the variety of production methods is due to the different raw materials processed, and the process and production methods adopted are also different. At present, the alkali method is used in the industry, mainly Bayer method, sintering method and joint method (the combination of Bayer method and sintering method). Therefore, there are many equipment for aluminum oxide production. The most important equipment in the Bayer aluminum oxygenation plant is the high-pressure dissolution equipment, and the most important equipment in the sintering aluminum oxygenation plant is the clinker kiln. ......

Properties of aluminum oxide

Source: "Theory and Process of Aluminum Oxide Production" - Li Wangxing

Aluminium oxyliteride and its hydrates are the main components that make up various types of bauxite, and the properties of aluminium oxygenide and its hydrates and impurities are the properties that select and determine oxygenationTherefore, it is of great significance to study the properties of aluminum oxide and its hydrates.......

Calcined aluminium oxyliterate

Source: "Theory and Process of Aluminum Oxide Production" - Li Wangxing

Industrial aluminum hydroxide or aluminum oxide roasting at 1200~1700 °C to obtain α-A1203, its chemical purity is high, the melting point is about 2050 °C, high temperature calcined aluminum oxyliterateThe hardness can reach about 9, the water absorption rate is low (≤2.5%), it has good chemical stability, high temperature resistance and wear resistance, and is widely used as the basic raw material for refractory materials, ceramics, abrasives (grinding wheels, polishing, grinding) and other industries.......

Fig. 8 Prototype of human−computer interaction form of DNMKG

References

[1] LI X, SHAN G C, ZHAO H B, SHEK C H. Domain knowledge aided machine learning method for properties prediction of soft magnetic metallic glasses [J]. Transactions of Nonferrous Metals Society of China, 2023, 33(1): 209−219.

[2] XIAO Zhu, DING Yan-jun, WANG Ze-jun, JIA Yan-lin, JIANG Yan-bin, GONG Shen, LI Zhou. Research and development of advanced copper matrix composites [J]. Transactions of Nonferrous Metals Society of China, 2024, 34(12): 3789−3821.

[3] LEHMANN J, ISELE R, JAKOB M, JENTZSCH A, KONTOKOSTAS D, MENDES P N, HELLMANN S, MORSEY M, VAN KLEEF P, AUER S, BIZER C. Dbpedia — A large-scale, multilingual knowledge base extracted from Wikipedia [J]. Semantic Web, 2015, 6(2): 167−15.

[4] HOFFART J, SUCHANEK F M, BERBERICH K, WEIKUM G. YAGO2: A Spatially and temporally enhanced knowledge base from Wikipedia [J]. Artificial Intelligence, 2013, 194: 28−61.

[5] BOLLACKER K, EVANS C, PARITOSH P, STURGE T, TAYLOR J. Freebase: A collaboratively created graph database for structuring human knowledge [C]//SIGMOD '08: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. Vancouver: ACM, 2008: 1247−1250.

[6] BOLLACKER K, COOK R, TUFTS P. Freebase: A shared database of structured general human knowledge [C]// Proceedings of the 22nd AAAI Conference on Artificial

Intelligence. Vancouver: AAAI Press, 2007: 1962−1963.

[7] CYGANIAK R, WOOD D, LANTHALER M. RDF 1.1 Concepts and Abstract Syntax [EB/OL]. [2014−2−25]. https://www. w3.org/TR/rdf11-concepts/.

[8] CHEN H J, YU T, CHEN J Y. Semantic web meets integrative biology: A survey [J]. Brief Bioinform, 2013, 14(1): 109−125.

[9] ERNST P, SIU A, WEIKUM G. KnowLife: A versatile approach for constructing a large knowledge graph for biomedical sciences [J]. BMC Bioinformatics, 2015, 16: 157−170.

[10] VARDE A S, BEGLEY E E, FAHRENHOLZ-MANN S. MatML: XML for information exchange with materials property data [C]//DMSSP'06: Proceedings of the 4th International Workshop on Data Mining Standards, Services and Platforms. Philadelphia: ACM, 2006: 47−54.

[11] OJALA T, OVER H H. Approaches in using MatML as a common language for materials data exchange [J]. Data Science Journal, 2008, 7: 179−195.

[12] GRUBER T R. A translation approach to portable ontology specifications [J]. Knowledge Acquisition, 1993, 5: 199−220.

[13] CHEUNG K, HUNTER J, DRENNAN J. MatSeek: An ontology-based federated search interface for materials scientists [J]. IEEE Intelligent Systems, 2009, 24: 47−56.

[14] PREMKUMAR V, KRISHNAMURTY S, WILEDEN J C, GROSSE I R. A semantic knowledge management system for laminated composites [J]. Advanced Engineering Informatics, 2014, 28: 91−101.

[15] QIAO Bo, FANG Kui, CHEN Yi-ming, ZHU Xing-hui. Building thesaurus-based knowledge graph based on schema layer [J]. Cluster Computing, 2017, 20(1): 81−91.

[16] ZHANG Xiao-ming, LIU Xin, LI Xin, PAN Dong-yu. MMKG: An approach to generate metallic materials knowledge graph based on DBpedia and Wikipedia [J]. Computer Physics Communications, 2017, 211: 98−112.

[17] WANG Ting, GU Han-zhe, WU Zhuang, GAO Jing. Multi-source knowledge integration based on machine learning algorithms for domain ontology [J]. Neural Computing and Applications, 2020, 32: 235−245.

[18] YUAN Man, CHU Bing, XIAO Yao. Knowledge graph of reservoir structure based on thesaurus [J]. Journal of Jilin University (Information Science Edition), 2020, 38(1): 72−78. (in Chinese)

[19] GU Pei-qin, CHEN Hua-jun, YU Tong. Ontology-oriented diagnostic system for traditional Chinese medicine based on relation refinement [J]. Computational and Mathematical Methods in Medicine, 2013(2): 317803.

[20] ZHANG Wei-ping, LUO Hai-ji, CUI Dong-yue. China thesaurus for nonferrous metals industry [M]. Beijing: China Machine Press, 1993. (in Chinese)

[21] WANG Zhi-chun. Knowledge extraction from Chinese wiki encyclopedias [J]. Frontiers of Information Technology & Electronic Engineering, 2012, 13(4): 268−280.

[22] WANG Zhi-gang, LI Juan-zi, WANG Zhi-chun, LI Shuang-jie, LI Ming-yang, ZHANG Dong-sheng, SHI Yao, LIU Yong-bin, ZHANG Peng, TANG Jie. XLore: A large-scale English−Chinese bilingual knowledge graph [C]// Proceedings of the 12th International Semantic Web Conference (1035). Sydney: CEUR-WS.org, 2013: 121−124.

[23] MANNING C, SURDEANU M, BAUER J, FINKEL J, BETHARD S, MCCLOSKY D. The Stanford CoreNLP natural language processing toolkit [C]//Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Baltimore: Association for Computational Linguistics, 2014: 55−60.

[24] MIKOLOV T, CHEN K, CORRADO G, DEAN J. Efficient estimation of word representations in vector space [C]// Proceedings of the 1st International Conference on Learning Representations. Scottsdate: Microtome Publishing, 2013.

[25] REHUREK R, SOJKA P. Software framework for topic modelling with large corpora [C]//Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Paris: European Language Resources Association, 2010: 45−50.

DNMKG：一种基于多语料库构建有色金属领域知识图谱的方法

李海亮[1]，王海东[2]

1. 有科期刊出版(北京)有限公司，北京 100088；
2. 中南大学 资源加工与生物工程学院，长沙 410083

摘　要：为解决有色金属领域中文研究资料的利用不足问题，提出了一种构建有色金属领域知识图谱(DNMKG)的方法。该方法以领域词表为基础，将实体和关系映射为 RDF 三元组，构建知识图谱的框架结构。通过从开放知识库中提取属性和相关实体，进一步丰富了知识图谱的内容。此外，还从近期文献中构建了一个包含超过 10 亿字的多源异构大规模语料库，以进一步扩充 DNMKG。以知识图谱为先验知识，结合自然语言处理技术对语料库进行分析，生成词向量。采用一种新的实体评估算法识别并提取真实的领域实体，并将其纳入 DNMKG。同时，开发了一个原型系统用于知识图谱的可视化展示以及支持人机交互。研究结果表明，DNMKG 能够有效提升有色金属领域的知识发现能力并显著提高研究效率。

关键词：知识图谱；有色金属；词表；词向量模型；多源异构语料库

(Edited by Wei-ping CHEN)