IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 16, 2019, accepted July 30, 2019, date of publication August 5, 2019, date of current version August 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2933370

From Vision to Content: Construction of Domain-Specific Multi-Modal Knowledge Graph

XIAOMING ZHANG, XIAOLING SUN, CHUNJIE XIE, AND BING LUN
School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China

Corresponding author: Xiaoling Sun (0211714022@stu.hebust.edu.cn)

This work was supported in part by the Natural Science Foundation of Hebei Province under Grant F2018208116, in part by the Innovation Foundation for Postgraduate of Hebei Province under Grant CXZZSS2018092, and in part by the National Natural Science Foundation of China under Grant 51271033.

ABSTRACT Knowledge graphs are usually constructed to describe the various concepts that exist in real world as well as the relationships between them. There are many knowledge graphs in specific fields, but they usually pay more attention on text or structured data, ignoring the image vision information, and cannot play an adequate role in the emerging visualization applications. Aiming at this issue, we design a method that integrates image vision information and text information derived from Wikimedia Commons to construct a domain-specific multi-modal knowledge graph, taking the metallic materials domain as an example to illustrate the method. The text description of each image is regarded as its context semantic to acquire the image's context semantic labels based on the DBpedia resource. Furthermore, we adopt deep neural network model instead of simple visual descriptors to acquire the image's visual semantic labels using the concepts from WordNet. In order to fuse the visual semantic labels and context semantic labels, a path-based concept extension and fusion strategy is proposed based on the conceptual hierarchies of WordNet and DBpedia to obtain the effective extension concepts as well as the links between them, increasing the scale of the knowledge graph and enhancing the correlation between images. The experimental results show that the maximum extension level has a significant impact on the quality of the generated domain knowledge graph, and the best extension level number is respectively determined for both DBpedia and WordNet. In addition, the results of this paper are compared with IMGpedia to further show the effectiveness of the proposed method.

INDEX TERMS Knowledge graph, information fusion, multi-modal knowledge, metallic materials knowledge.

I. INTRODUCTION

The research aiming at the semantic representation of domain-specific data generally focuses on text data or structured data, such as transforming structured databases into knowledge graph to provide semantic query services [1] and extracting knowledge from unstructured source data to build new ontologies (e.g., STSM [2], X. Zhang [3]). Even some ontology (e.g., MMOY [4] and Materials ontology [5]) do not focus on multimedia data at the beginning of construction, so the ability of these knowledge graph to process multimedia data (such as images, video, audio, etc.) is limited. In fact, the knowledge contained in vision information is as important as that contained in textual information. Knowledge in text messages can help users understand more of the hidden content, while knowledge in visual information can help users understand more of the visible content. Multi-modal knowledge graph can provide users with better query experience in Engineering applications, e.g. Visual Question Answering [6]–[8], by fusing the knowledge in text information and visual information.

IMGpedia [9] is a large multi-modal knowledge graph which includes two types of relationships: image-to-image and image-to-text. The image-to-image relationships are mined based on the simple visual descriptors of images, and the image-to-text relationships are mined based on the

The associate editor coordinating the review of this manuscript and approving it for publication was Chang Choi.

image-title pairs. And most of the text resources are associated with DBpedia [10]. However, simple visual descriptors are difficult to integrate with text and may differ greatly from the semantic information expressed in the images. In contrast, the vision semantic labels of images obtained by using the deep neural network model can represent the advanced visual characteristics of images, which can provide more semantics from the vision perspective of the image. In addition, the relationships between titles and images are broad and coarse-grained, while the relationships between the entities in the images' association text and the images are fine-grained and contextual, although they may be indirect and implicit. DBpedia is one of the largest domain knowledge graphs in the world, which contains a large amount of knowledge, such as related knowledge in the metallic materials domain [11]. WordNet [12] is a wide range of English lexical semantic network, organized into different synonym networks according to different parts of speech, and different synonym networks are linked by a variety of relationships. Therefore, based on the image information in Wikimedia Commons, DBpedia and WordNet are used to provide the domain background knowledge to connect the visual semantic labels and context semantic labels.

In this paper, a domain-specific multi-modal knowledge fusion method is proposed, which combines the visual semantic with the context semantic of the images and fuses them based on the conceptual hierarchies of WordNet and DBpedia, so as to construct a domain-specific multi-modal knowledge graph. We demonstrate the main contributions of this work as follows:

1) We proposed an approach to gradually extract domain-specific image data from the Wikimedia Commons. In order to obtain domain data, we define the scope of data acquisition by constructing a domain concept dictionary, and then obtain candidate image data (including images and their corresponding text) through these concepts in the concept dictionary. We also give a strategy to filter the candidate image data to make the result data more accurate.
2) We use DBpedia-Spotlight [13] to annotate the image association text to obtain the context semantic labels of the image, which can represent the semantic summarized from the text description of the image using the entities derived from DBpedia resources. Meanwhile, we use the ImageNet dataset to train VGG-Net [14] model to obtain the visual semantic labels of each image, which can represent the visual semantic information of the image using the concepts derived from WordNet resources.
3) A path-based concept extension and fusion strategy is designed based on the hierarchy of WordNet and DBpedia to extend the obtained semantic labels. Among them, visual semantic labels are extended according to the hierarchy of WordNet, and context semantic labels are extended by the hierarchy of DBpedia. Then the common extension concepts during the extension are recognized as the connection points between images. Through the concept extension, more relevant resources can be obtained, not only increasing the scale of the knowledge graph, but also enhancing the correlation between images, thus forming an interconnected domain knowledge graph. In addition, we have also considered the strategy to determine a suitable level of concept extension to control the abstractness of the extension concepts.

The rest of this paper is organized as follows: In Section II, we discuss the related work. The problem description is presented, and the related concepts are defined in Section III. Then in Section IV, we describe the entire process and implementation of the method in detail and perform the experiments in Section V. Finally, Section VI provides the conclusion and the future work.

II. RELATED WORK

This paper attempts to construct a domain-specific multi-modal knowledge graph based on the image visual features and the image's association text. Therefore, in this section, we will discuss three issues related to the acquisition of image visual knowledge, text semantic knowledge extraction and the construction of knowledge graph.

A. ACQUISITION OF IMAGE VISION KNOWLEDGE

Image vision knowledge is reflected by image features. The image feature acquisition can be divided into two categories: hand-crafted features extraction method [15]–[17] and deep learning extraction method [18]–[20]. The method based on hand-crafted features usually need professional knowledge and use the surface properties of the image to extract image features, so the learning ability of the model suffers from great limitation and can not fully reflect the essential attributes of the object. The method based on deep learning mainly uses convolution neural network (CNN) to automatically extract image features. Compared with the hand-crafted features, the image features extracted by deep learning method are not fixed and more comprehensive, and the features obtained by this method can express the deep semantic information of the image. Therefore, we use the ImageNet dataset to train deep learning model VGG-Net to obtain the visual semantic labels of images, and it can establish links to WordNet resource labels.

B. ACQUISITION OF TEXT SEMANTIC KNOWLEDG

The acquisition of text knowledge mainly includes two aspects: entity extraction and relation extraction. Entity extraction is also called Named Entity Recognition (NER). At present, the main methods of entity extraction are rule-based methods [21]–[23], unsupervised methods [24], [25], neural network methods [26]–[28] and so on. The purpose of relation extraction is to judge whether two entities are related from a sentence [29]. The image's description text in Wikimedia Commons knowledge base is a single entity or short text. Therefore, this paper uses entity extraction method to get the entities in text. Unlike the above approach, we use the existing method named DBpedia-Spotlight to obtain entities in text, and the obtained entities have corresponding resource labels in DBpedia knowledge base.

C. CONSTRUCTION OF KNOWLEDGE GRAPH

Knowledge graph has become a hotspot in industry and academia because of its powerful semantic expression and organizational abilities. The knowledge graph can be classified into general knowledge graph, such as Freebase [30], YAGO [31], Knowledge Vault [32], Microsoft Concept Graph [33], [34], and domain knowledge graph, such as medical knowledge graph SIDER [35], music knowledge graph MusicBrainz [36], movie knowledge graph IMDB [37]. The domain knowledge graph has gradually attracted lots of attention because of its strong professionalism and cohesiveness of domain knowledge.

The methods of construction domain knowledge graph can be divided into two categories according to the structure feature of the data source. One is to build knowledge graph based on structured or semi-structured information in Wikipedia or other existing knowledge base (e.g., MMKG [11], BabelNet [38], and WordNetGraph [39]). MMKG [11] extracts data from DBpedia and Wikipedia to construct knowledge graph. BabelNet [38] is a multilingual knowledge graph based on Wikipedia and WordNet. WordNetGraph [39] is based on the conceptual structure defined by WordNet and use the classifier to automatically label the terms defined by natural language to construct the knowledge graph.

The other category of methods [40]–[46] tries to construct knowledge graph by automatically extracting knowledge from unstructured data source such as texts, images, or other media. Among them, [40]–[42] is to extract structured knowledge from domain-related texts to construct domain knowledge graph. However, HDSKG [44] extracts relational triples from Web pages and then uses a pre-trained SVM classifier and domain dictionary to determine the domain relevance of the extracted triples. MeSH-like [45] extracts entity attributes from medical textbooks, medical online websites, and other domain-related texts using a rule-based approach and fuses them with SimHash-TF-ID algorithm. KnowEdu [46] is an educational domain knowledge graph constructed by extracting domain entities from domain-related heterogeneous data sources (such as textbooks) using recursive neural networks.

All the aforementioned methods extract knowledge from text-based data. However, there is also a large amount of visual knowledge residing in images, which can not only be used as auxiliary information to improve the effect of knowledge graph construction (e.g., IKRL model [47]), but also be used to construct multi-modal knowledge graph with the knowledge derived from images. Multi-modal knowledge graph can enhance people's understanding of knowledge from different aspects according to the characteristics of different modal. MKBE [48] uses embedding to encode different modal data and connects entity and multi-modal data with existing relational models, so as to construct domain-specific knowledge graph. Recently, the development of representation learning provides new methods for constructing knowledge graph. For example, Thomas *et al.* [49], [50] used Inception-v3 [51] to get the vector of visual features, Word2vec [52] to get the vector of word embedding, and TransE [53] to get KG-Entity Embedding, and then used vector splicing to achieve multi-modal fusion. In order to further enhance the effect of multi-modal knowledge graph construction, Mousselly *et al.* [54] embed visual and linguistic to complete the multi-modal knowledge graph under the verification of structured knowledge. AMVAE model [55] captures the fine-grained information between data modal by introducing the attention mechanism of bidirectional long short-term memory to realize the fusion of link and multi-modal content.

Unlike the above methods, the method proposed in this paper tries to connect the discrete entities derived from the multi-modal data source including unstructured text and images in a specific domain. Furthermore, these entities have semantic context coming from WordNet and DBpedia respectively, based on which the entities are moderately expanded by concept extension to enrich the result knowledge graph. Therefore, the fusion strategy in our method considers more semantic relations computing rather than just vector computing.

III. PROBLEM DESCRIPTION AND CONCEPT DEFINITION

Both image vision and image association text contain semantic information. Therefore, this paper attempts to construct a domain-specific multi-modal knowledge graph in Wikimedia Commons by combining the image visual semantic content and the image association text. Thus, we should address the following issues:

1) How to determine whether a concept belong to a specific domain, and then obtain images and associated text of the concept.
2) How to obtain the image visual semantic information and the semantic context of the image association text.
3) How to fuse these two types' semantic information.

Fig. 1 illustrates the problem we are trying to solve.

In order to solve the above problems, we propose a path-based concept extension and fusion strategy. For the convenience of expression, it is necessary to explain the following definitions. And an example is presented at the end of this section to illustrate the meaning of each definition.

Definition 1: Association text. The image association text refers to the image text descriptions, article title and image categories in Wikimedia Commons pages where the images are located. These association texts serve as the basis to obtain context semantic labels of images.

Definition 2: Association concept. Assuming G_i is an image object, the association concept of an image refers to a set of visual semantic labels (V_{vision}^{Gi}) obtained from the image feature and a set of context semantic labels (V_{text}^{Gi}) extracted from association text. The union of them as association

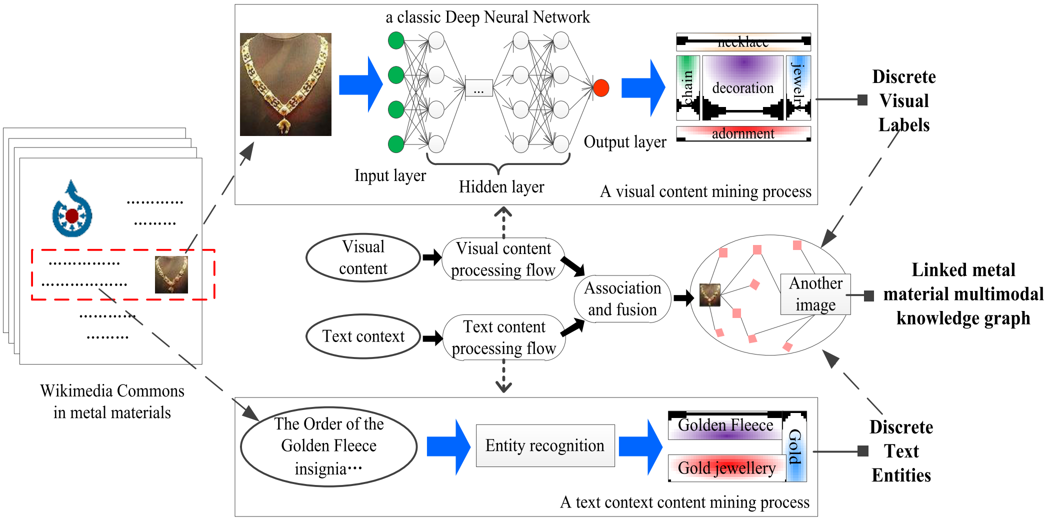
a classic Deep Neural Network
Input layer
Hidden layer
Output layer
necklace
chain
decoration
jewelry
adornment
A visual content mining process
Discrete Visual Labels
Wikimedia Commons in metal materials
Visual content
Visual content processing flow
Text context
Text content processing flow
Association and fusion
Another image
Linked metal material multimodal knowledge graph
The Order of the Golden Fleece insignia···
Entity recognition
Golden Fleece
Gold jewellery
Gold
A text context content mining process
Discrete Text Entities

FIGURE 1. The main problems solved in this paper.

concept set (V^{Gi}) is defined as Equation (1).

$$V^{Gi} = V^{Gi}_{vision} \cup V^{Gi}_{text} \tag{1}$$

Definition 3: Extensional association concept. The upper concepts of the elements in V^{Gi} are called the extensional association concepts (EC_{Gi}), which are derived from the conceptual hierarchy by WordNet or DBpedia. The collection of extensional association concepts is defined as Equation (2).

$$EC_{\mathrm{Gi}} = \left\{ \begin{array}{l} ec\,|(ec\ \ is\ a\ hypernym\ of\ \ s) \wedge \begin{pmatrix} ec \in \ WordNet \\ \vee ec \in \ DBpedia \end{pmatrix}, \\ s \in V^{\mathrm{Gi}} \end{array} \right\} \tag{2}$$

Definition 4: Maximum extension level. The maximum extension level L_{max} is the maximum number of levels that association concept can be extended upwards (generalize) in the hierarchy of the specific background knowledge base. In this paper, it refers to the maximum number of layers that image G_i's V^{Gi}_{vision} in WordNet and V^{Gi}_{text} in DBpedia.

Definition 5: Effective extension concept. The effective extension concept refers to the useful extensional association concepts in the construction of knowledge graph. If G_i and G_j are two different images objects, $EP_{Gi\to Gj}$ is defined as the set of concepts contained in the shortest path from any concept C_t in $V^{Gi} \cup EC_{Gi}$ to the same concept in $V^{Gj} \cup EC_{Gj}$. Then, the effective extension concept set $EE_{Gi,Gj}$ (equal to $EE_{Gj,Gi}$) of the image G_i and G_j is shown in Equation (3), and the total effective extension concept of G_i is the sum of all its relatively effective extension concepts. As shown in Equation (4), n is the total number of images.

$$EE_{Gi,Gj} = \mathrm{EP}_{Gi\to Gj} \cup EP_{Gj\to Gi}, (i \neq j) \tag{3}$$

$$EE_{Gi} = \bigcup\nolimits_{j=1,j\neq i}^{j=n} \mathrm{EP}_{Gi\to Gj} \tag{4}$$

Definition 6: Independent concept. Independent concepts are defined as concepts that do not belong to the collection of effective extension concepts. When the value of L_{max} is fixed, for the image G_i, the independent concepts is marked as $x(l)_{Gi}$, and l is the value of L_{max}. The independent concept set can be defined as Equation (5).

$$x(l)_{Gi} = V^{Gi} + EC_{Gi} - EE_{Gi} \tag{5}$$

Fig. 2 shows an example of effective extension concept for L_{max}=2. The concept path $C1 \to EC1 \to EC12$ of the image G_1 is taken as an example to explain how to obtain the effective extension concepts on this path. First, for the concept *C1*, since the same concept does not exist in the $V^{G2} \cup EC_{G2}$ set of G_2, it is then necessary to judge the concept *EC1*. *EC1* has the same concept in the $V^{G2} \cup EC_{G2}$ set of G_2, so the concepts below *EC1* are all effective extension concepts of the image G_1. As for the concept *EC12*, although the same concept exists in the concept set of G_2, *EC12* is not an effective extended concept because it is not the shortest concept path. Therefore, for this path, the effective extension concept set for G_1 contains concepts *C1* and *EC1*.

IV. CONSTRUCTION METHOD OF DOMAIN-SPECIFIC MULTI-MODAL KNOWLEDGE GRAPH

A. APPROACH OVERVIEW

The main purpose of this subsection is to illustrate how to use the path-based concept extension and fusion strategy proposed in this paper to gradually construct a domain-specific multi-modal knowledge graph. As shown in Fig. 3, the main steps of our approach can be summarized as follows:

1) Obtaining domain-specific multi-modal data from Wikimedia Commons. In order to limit the domain of the source data, we use a supervised approach to build an artificial concept dictionary to determine the domain scope of the data and obtain the dataset by recursively accessing Wikimedia Commons resources.

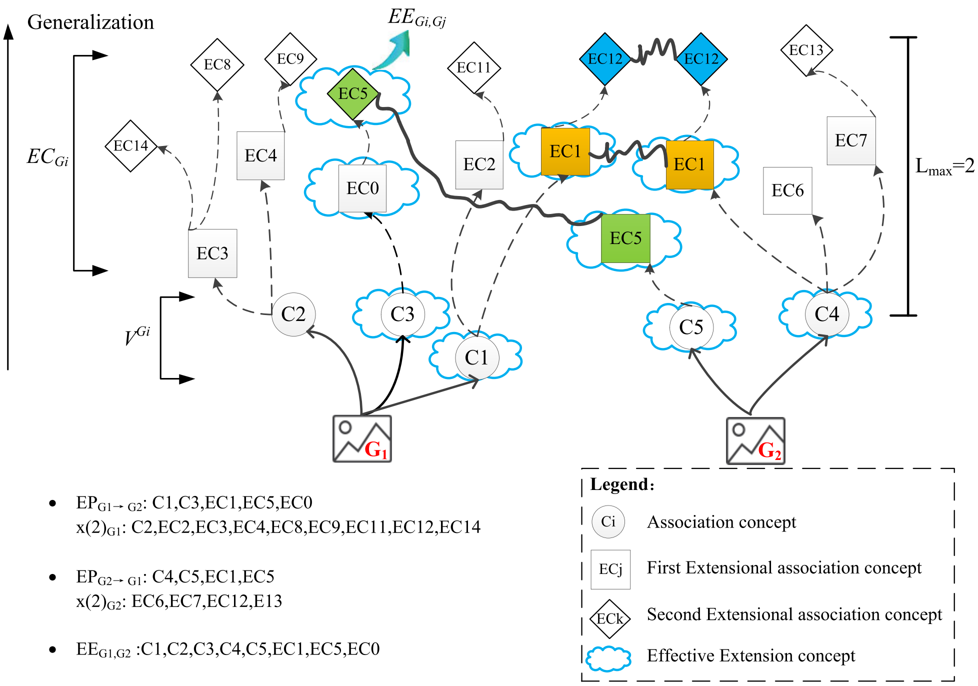
Generalization
EE_{Gi,Gj}
EC8
EC9
EC5
EC11
EC12
EC12
EC13
EC14
EC4
EC0
EC2
EC1
EC1
EC7
EC6
L_max=2
EC_Gi
EC3
EC5
C2
C3
C1
C5
C4
V^Gi
G1
G2
EP_{G1→G2}: C1,C3,EC1,EC5,EC0
x(2)_{G1}: C2,EC2,EC3,EC4,EC8,EC9,EC11,EC12,EC14
EP_{G2→G1}: C4,C5,EC1,EC5
x(2)_{G2}: EC6,EC7,EC12,E13
EE_{G1,G2} :C1,C2,C3,C4,C5,EC1,EC5,EC0
Legend:
Ci Association concept
ECj First Extensional association concept
ECk Second Extensional association concept
Effective Extension concept

FIGURE 2. Obtain the effective extension concepts of any two images. Note: (1) Nodes of the same color represent the same concept (2) Arbitrary shapes with thick solid lines link the same nodes (3) Concepts of the same shape are at the same level of extension.

2) The generation of the context semantic labels from the association text. The V_{text}^{Gi} set of image G_i is generated from the image association text in the Wikimedia Commons by using DBpedia-Spotlight, and then we can get a separate set of entity labels which are resources in DBpedia.
3) The acquisition of visual semantic labels from image vision. Using ImageNet data set to train VGG-Net model to get V_{vision}^{Gi} of each image is acquired from the image in the Wikimedia Commons, and then we can get a separate set of visual semantic labels which are resources in WordNet.
4) Constructing domain-specific multi-modal knowledge graph based on concept extension. Using the results of steps 2) and 3), we can extend the concepts according to concept hierarchy of DBpedia and WordNet to get EC_{Gi} set of each image. We obtain $EE_{Gi,Gj}$ sets of any two images G_i and G_j by using the Equation (3). Finally, using $EE_{Gi,Gj}$ to construct the domain-specific multi-modal knowledge graph.

We take the metallic materials domain as an example to illustrate the method of constructing domain-specific knowledge graph in this paper. The details of each step are illustrated in the following subsections.

B. OBTAINING MULTI-MODAL DATA FROM WIKIMEDIA COMMONS IN METALLIC MATERIALS DOMAIN

In this section, we accomplish the acquisition and filtering of domain-specific data in two steps. Firstly, we define a concept dictionary consisting of prepared candidate concepts, which are used to identify and retrieve Wikimedia Commons resources that only contain a fraction of the images and texts we need. Secondly, all the above images and texts are filtered to form the domain-specific source data. In Fig. 4, a candidate concept C_1 is taken as an example to illustrate how to gradually acquire images and text in metallic materials domain from Wikimedia Commons.

1) CONSTRUCTING AN EFFECTIVE CONCEPT DICTIONARY

There are many sources such as existing domain knowledge bases or domain ontologies can be used to choose candidate concepts, which can be selected by domain users according to their requirement. To ensure that the candidate concepts could be found in the Wikimedia Commons, we utilize the titles of Wikimedia Commons document to check the selected candidate concepts, and it can be described as Equation (6).

$$(\forall c)(\forall t)(c \in C \wedge t \in T \wedge isSim(c,t) \rightarrow put(c,D)) \quad (6)$$

As shown in Equation (6), C is the set of all candidate concepts manually selected and T is the set of all Wikimedia commons documents' title or categories. D is the concept dictionary that we are going to build. Furthermore, the meaning of *isSim(c, t)* is that c and t are identical by string comparison when case is ignored, and the meaning of *put(c, D)* is that c can be put into D. This rule means that if a Wikimedia Commons document title could match with the candidate concept c, c could be added to the concept dictionary D. For example, the candidate concept *Quasicrystal*, which does not have a page directly titled *Quasicrystal* in Wikimedia

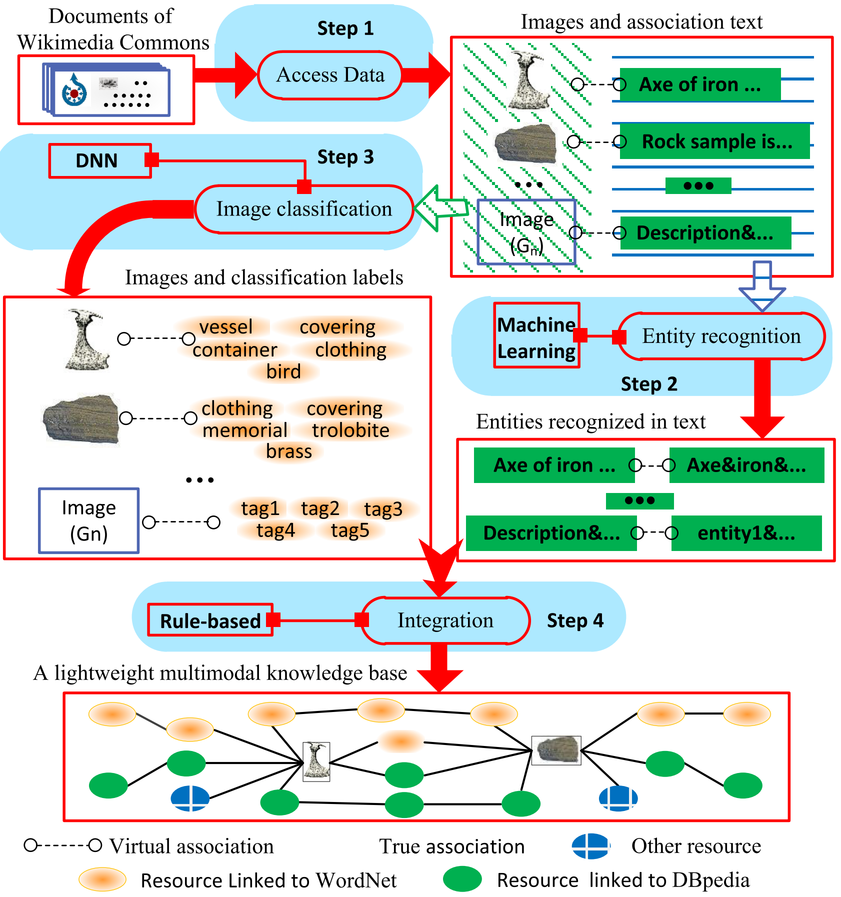

FIGURE 3. Approach overview.

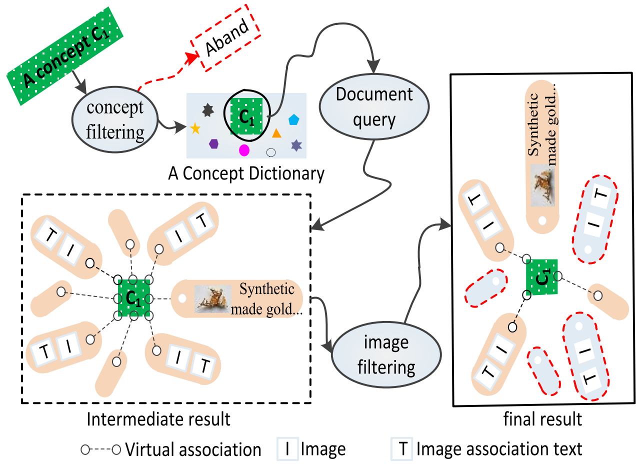

FIGURE 4. The sample process of getting source data based on a candidate concept C_1.

Commons query results, will be discarded and cannot be added to the concept dictionary.

The concept dictionary does not need to have a structure or any superfluous description. It is just a simple list of domain concepts in metallic materials science. After the concept dictionary has been built, we can start the document query.

2) OBTAINING SOURCE DATA

Wikimedia Commons has two types of documents, i.e. Category page and Gallery page. The former aims to display the image with a manually defined type structure, which does not contain details. The latter is designed to show the details of the image, including the image name, image text descriptions, Wikimedia article URL, and so on. For each concept in the concept dictionary, the image information is obtained from the Category pages and Gallery pages respectively. Since the Wikimedia Commons article may contain multiple languages, and this article only deals with English articles, these images need to be filtered according to the rules in Equation (7).

$$(\forall i)\,(\forall w)\left(\begin{array}{l} i\in I\wedge w\in W\wedge isAppear\,(i,w)\wedge isEnglish\,(w)\\ \rightarrow store\,(i)\end{array}\right) \tag{7}$$

In Equation (7), i is the set of all images and w is the set of all articles acquired from Wikimedia Commons by the concept dictionary. The meaning of *is Appear(i,w)* is that i appears in w, and *is English(W)* means that w is an English article, and *store(i)* means that i could be stored as source data. This rule means that if an image i appear in an English Wikipedia document w, it can be saved as source data. In the meanwhile, each image and its corresponding associated text establish a virtual connection with the same ID. The concept ID and image ID are connected through an index table.

3) EVALUATING THE ACCURACY OF SOURCE DATA

The semantics of images may not match the conceptual semantics of the retrieval, so a new evaluation criterion is proposed to evaluate the accuracy of the whole data, as shown in Equation (8), which combines the vision semantics and the corresponding relationship between the current query concepts.

$$Accuracy(D, I) = \frac{\sum_{d \in D} \sum_{i \in I_d} check(d, i)}{\sum_{d \in D} |I_d|} \tag{8}$$

D is a set of concepts in the concept dictionary. *I* is a set of all images obtained by each concept in the concept set *D*. *d* is a value in concept set *D*, and I_d is a set of images obtained by concept *d*. *check(d,i)* is the judgment result of the corresponding relation between *d* and *i*, which is calculated according to Equation (9). *Accuracy (D,I)* means calculating the overall accuracy of the image set *I* obtained from each concept in concept set *D*.

$$(\forall d)\,(\forall i)\left(\begin{array}{l} d \in D \wedge i \in I \wedge isQueried(d, i) \wedge \ isMaterial\ (d, i) \\ \wedge\, isDirect\ (d, i) \rightarrow\ check\ (d, i) \end{array}\right) \tag{9}$$

As shown in Equation (9), if *i* is obtained from Wikimedia Commons by *d*, *isQueried(d,i)* is true. *isMaterial(d,i)* indicates that the true semantics of *d* in the image *i* belong to the concept of the metallic materials domain, and *isDirect(d,i)* means that *i* has direct semantic relation with *d*. *isCheck(d,i)* means that the value of *check(d,i)* is *1*. The value of *check(d,i)* can only be *1* or *0*. This rule implies that if *d* is a concept about metallic materials and *d* has direct semantic relation with *i*, the value *check(d,i)* is *1*, otherwise the value of *check(d,i)* is *0*.

Fig. 5 shows some concrete examples of judging the value of *check(d,i)*. Since the *Gold* in the images (a) and (f) does not represent the meaning of the *Gold* in the field of metallic materials, the values of *isMaterial(Gold,a)* and *isMaterial(Gold,f)* are all false. Likewise, the value of *isMaterial(Silver,c)* is false. Meanwhile image (d) not has direct semantic relation with *Silver*, so the value of *isDirect(Silver,d)* is false. In a similar way, the values of *isDirect(Titanium,k)* and *isDirect(Titanium,l)* are all false. Then the values of *check(Gold,a)*, *check(Gold,f)*, *check(Silver,c)*, *check(Silver,d)*, *check(Titanium,k)* and *check(Titanium, l)* are *0*. Hence the value of *Accuracy(D,I)* for all data in Fig. 5 is $Accuracy(D, I) = \frac{0+1+1+0+0+0+1+1+1+1+0+0}{12} = 0.5$. Through this approach, it can be seen that the image vision semantic content and the corresponding concepts do not always correspond to each other. The value of the *Accuracy(D,I)* can be appropriately adjusted by data filtering according to the user's requirement for the data. This article considers that it can be used when the *Accuracy(D,I)* value exceeds *0.8*.

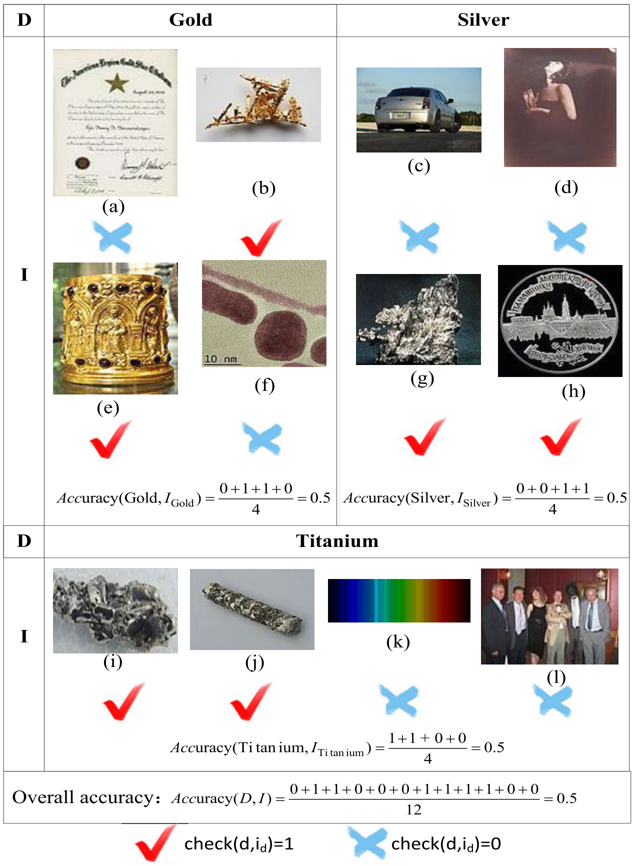
D
Gold
Silver
I
(a)
(b)
(c)
(d)
(e)
(f)
10 nm
(g)
(h)
$Accuracy(\text{Gold}, I_{\text{Gold}}) = \frac{0+1+1+0}{4} = 0.5$
$Accuracy(\text{Silver}, I_{\text{Silver}}) = \frac{0+0+1+1}{4} = 0.5$
D
Titanium
I
(i)
(j)
(k)
(l)
$Accuracy(\text{Ti tan ium}, I_{\text{Ti tan ium}}) = \frac{1+1+0+0}{4} = 0.5$
Overall accuracy: $Accuracy(D, I) = \frac{0+1+1+0+0+0+1+1+1+1+0+0}{12} = 0.5$
✓ check(d,i_d)=1
✗ check(d,i_d)=0

FIGURE 5. Some examples of calculating ***Accuracy(D,I)***.

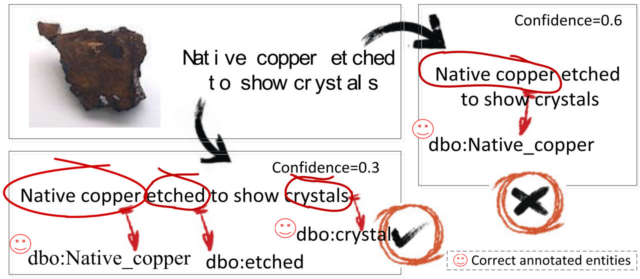
Nat i ve copper et ched t o show cr yst al s
Confidence=0.6
Native copper etched to show crystals
dbo:Native_copper
Confidence=0.3
Native copper etched to show crystals
dbo:Native_copper
dbo:etched
dbo:crystal
Correct annotated entities

FIGURE 6. Combine visual content of images to evaluate the accuracy of entity annotations.

C. GENERATION OF CONTEXT SEMANTIC LABELS FROM ASSOCIATION TEXT

The purpose of this section is to obtain context semantic labels for each image association text in the source data set. DBpedia-Spotlight is a tool for automatically annotation mentions of DBpedia resources in text, as shown in Fig. 6, enter a piece of text and output a bunch of discrete annotation mentions. In order to acquire the context semantic labels of image association text, we use DBpedia-spotlight to annotate the image association text to obtain the entities and use virtual links (same ID) to connect annotated results

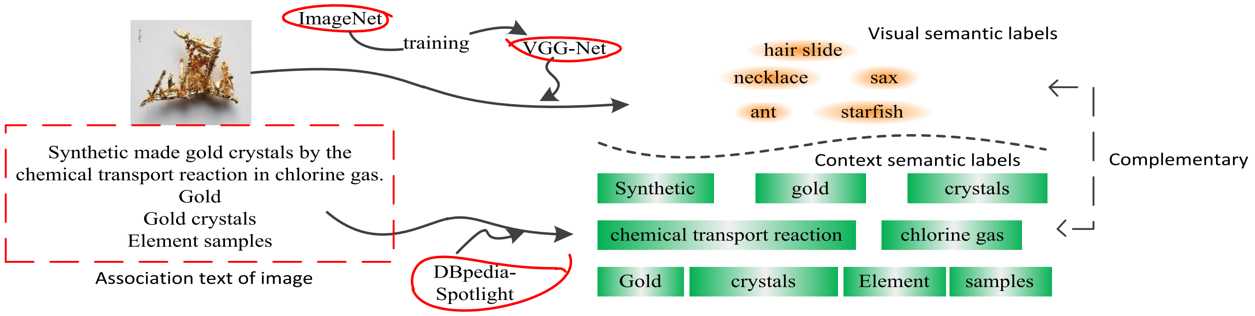

FIGURE 7. There is a complementary relationship between the processing results of the vision context and the processing results of the image association text.

to association text. By using DBpedia-Spotlight to annotate the association text of the image G_i, the annotation entity set V_{text}^{Gi} of the image can be obtained and used as the semantic context of the image association text.

However, annotation entities may not accurate, so we design a criterion which combines the semantics of annotation entities with the semantics of text description and image vision to judge the accuracy of context semantic labels of each image. As shown in Equation (10).

$$(\forall e)(\forall G_i)\begin{pmatrix} G_i \in I \wedge e \in V_{text}^{G_i} \wedge isMatchText(e, G_i) \\ \wedge isMatchVisual(e, G_i) \rightarrow isRightMatch(e) \end{pmatrix} \tag{10}$$

I is the set of images. *isMatchText(e,i)* indicates that entity *e* has the same meaning as the image association text, *isMatchVisual(e,*G_i*)* means that entity *e* can match the vision semantics of G_i, and the meaning of *isRightMatch(e)* is that the annotation of *e* is accurate. The overall meaning of this formula is that if *e* is an annotation entity for the association text of G_i and matches the visual semantics of G_i, *e* is marked as an exact annotation.

DBpedia-Spotlight has a parameter that can be manually adjusted, called *confidence*, to control how strict the system is with the results of the comments. Its range of values is [0, 1]. The smaller the value is, the higher the recall rate of annotation result will be, meaning that the more annotation mentions are obtained. In our experiment, the real recall rate of annotation mentions is not statistically significant, so based on the *confidence* of DBpedia-Spotlight, we design a new function *F(cnf)*, which is an approximation of *F1*, it can approach the true value infinitely as the sample size increase. When judging the accuracy of the annotation mentions result, we evaluate the data according to Equation (11). Experiments (Fig. 12) show that for the data in this paper, when the *confidence* is *0.3*, the overall entity annotation effect is better.

$$F(cnf) = 2 \cdot \frac{\frac{M(cnf)}{N(cnf)} \cdot \frac{M(cnf)}{M(0)}}{\frac{M(cnf)}{N(cnf)} + \frac{M(cnf)}{M(0)}} \tag{11}$$

The *cnf* is the value of *confidence*. *M(cnf)* is the number of annotation entities that the tool correctly generates when the *confidence* is *cnf*. *N(cnf)* is the total number of annotation entities that the tool generates when the *confidence* is *cnf*. *M(0)* is the number of annotation entities that the tool generates when *confidence* is *0*, and it is used as the basis for recall rate calculation.

Fig. 6 shows a concrete example to evaluate the entity annotation results for an image association text. In this example, the annotation mentions result with *confidence* of *0* happens to be the same as the result with a *confidence* of *0.3*. When the *confidence* is *0* or *0.3*, the annotation mentions results are both *dbo:Native_copper, dbo:etched and dbo:crystals*. When *confidence* is *0.6*, the DBpedia-spotlight annotation mention includes just *dbo:Native_copper*. Combined with Equation (10) and Equation (11), the value of *F(0.3)* is *0.33* and the value of *F(0.6)* is *0.25*. In this case, we think the annotation result when *confidence* is *0.3* is better.

D. THE ACQUISITION OF VISUAL SEMANTIC LABELS FROM IMAGE VISION

The purpose of this section is to obtain a set of visual semantic labels for each image of the source data results. ImageNet is a manually annotated image dataset containing 1000 classifications, and its annotation vocabulary is derived from WordNet, thus providing a way to link images with WordNet entities. In this paper, the result model for train VGG-Net model with ImageNet datasets is called VGG-ImgNet-Result and is used to obtain top-5 visual semantic labels for each image, and each visual semantic label can establish links to WordNet resources. And each image is associated with its visual semantic label set via a virtual link (same ID).

A more detailed illustration of why we use ImageNet is shown in Fig. 7. Five labels we have got from VGG-ImgNet-Result are *hair slide*, *necklace*, *sax*, *ant* and *starfish*. From the meaning of the visual semantic label itself, these five visual semantic labels can be roughly divided into two categories. *hair slide*, *necklace* and *sax* are products which are made of metal and may have hidden semantic association with the concept of *gold*, while the other category (i.e. *ant* and *starfish*) of visual semantic labels has no association with *gold*. However, if we look at the image and its visual semantic labels from a different perspective, the image in Fig. 7 does look like a swarm of ants or starfish. We believe that the visual semantic labels obtained from the vision perspective may be also meaningful. The visual semantic labels of images express the visual features of images and, in some cases, can be complementary to the association text of the image, so we retain these labels.

As shown in Fig. 7, the context semantic labels of the association text and the visual semantic labels of the image are acquired for the descriptions of the image content. Although they describe the image from the different aspects, the results show that they are both reasonable, so it is further verified that the image vision and the association text of image are complementary to each other. We did not adjust or optimize the VGG-Net model, but we design a new standard to evaluate the results of the visual semantic annotation results, which called *Satisfaction*.

$$Satisfaction(i, L) = \frac{\sum_{l \in L} relevance(i, l)}{|L|} \tag{12}$$

As shown in Equation (12), *i* is an image and *L* is the set of visual semantic labels related to *i*. The *relevance(i, l)* is a value which is only *0* or *1*. If *i* is similar to *l* from the vision perspective, the value of *relevance(i, l)* is *1*, otherwise, the value of *relevance(i, l)* is *0*.

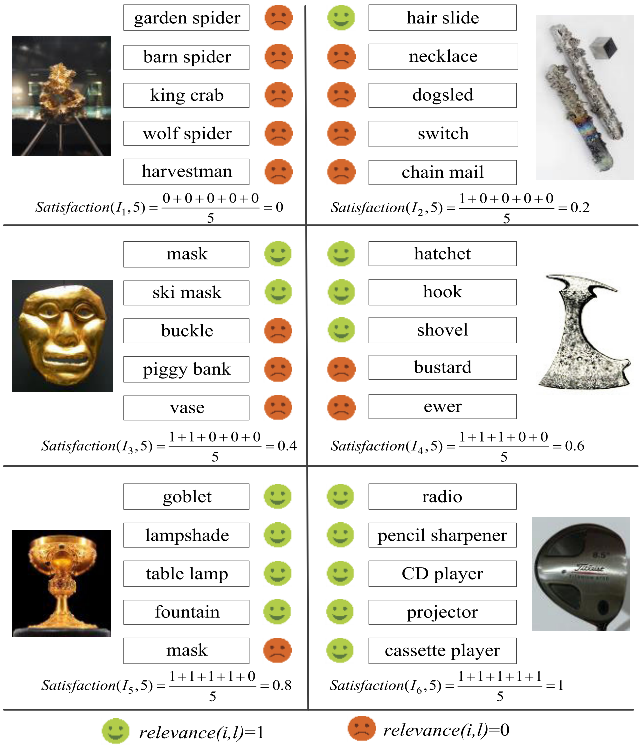

FIGURE 8. **An example of calculating image *Satisfaction(i,L)*. Assume that the value of | *L* | is *5*.**

Some examples of visual semantic labels *Satisfaction* are listed in Fig. 8. Assuming that the value of | *L* | is *5*, after obtaining the visual semantic labels of each image, the value of *relevance(i, l)* requires artificial judgments according to the vision semantics and the visual semantic labels. The average *Satisfaction* of the images reflects the reasonableness of the visual semantic annotation results. When the *Satisfaction* of each image is low, we can improve the *Satisfaction* by adjusting the number of vision labels or retraining a vision semantic model according to dataset.

E. CONSTRUCTION OF THE MULTI-MODAL METALLIC MATERIALS KNOWLEDGE GRAPH BASED ON CONCEPT EXTENSION

After subsections IV.C and IV.B, the V_{text}^{Gi} set and V_{vision}^{Gi} set of each image are acquired. However, these labels describing the image are discrete entities rather than knowledge graph with relations between the entities. Therefore, this paper proposes a method to extend the concept of semantic labels based on the conceptual hierarchies of DBpedia and WordNet, to establish the relationships of different images.

1) MAXIMUM EXTENSION LEVEL SELECTION

WordNet and DBpedia are two large knowledge bases, and they provide rich semantic associations that can determine hierarchical relationships between entities. However, if we extend the concept only according to the hierarchy of the knowledge graph, all the images will be related, and the relationships between them may be very abstract and most of them may be meaningless. This result is blind and undesirable, so we must reasonably limit the maximum extension level (L_{max}). We have designed two strategies working together to select L_{max}.

The first method is to calculate the growth rate (*Grow(l)*) of the independent concepts when L_{max} becomes larger. We consider that a suitable L_{max} makes the *Grow(l)* approach zero. As extension level *l* increases, the number of independent concepts will gradually decrease. Once *l* reaches a certain value, the *Grow(l)* will becomes a constant or even negative, indicating that the hierarchy of the knowledge graph begins to converge. If we continue to extend, there will be large-scale associations between concepts. Therefore, the extension level is limited by the size of *Grow(l)*. As shown in Equation (13), the $x(l)$ is the number of independent concepts when the number of extension level is *l*.

$$Grow(l) = \frac{x(l) - x(l-1)}{x(l-1)}, \quad (l \in [1, L_{\max}]) \tag{13}$$

The second method is to calculate the proportion of effective extension concepts. This paper argues that the higher the extension level, the less important the effective extension concept is. For the extension results of different *l*, the proportion of the effective extension concept is calculated by Equation (14). The larger the *select(l)* value, the greater the proportion of effective extension concepts in the current *l*, and the better the extension effect.

$$Select(l) = \frac{\sum_{i=1}^{i<=l} \sin(\frac{1}{i} * \frac{\pi}{2}) * |EEi|}{|ECl|} \tag{14}$$

As shown in Equation (14), the *l* is the number of extension level, and EC_l is the set of extensional association concepts

Algorithm 1 Get the Effective Extension Concepts set for any two Images

```
Input: ArrayList imageID
    int $L_{max}$ // Maximum extension level
Output: HashSet EE_Gi_Gj // the effective extension concepts set of $G_i$ and $G_j$
FOR EACH id in imageID
    ArrayList visual_semantic_labels_id ← get all $V^{Gi}_{vision}V^{Gi}_{vision}$ for each id
    ArrayList context_semantic_labels_id ← get all $V^{Gi}_{text}V^{Gi}_{text}$ for each id
    FOR EACH vl in visual_semantic_labels_id
        HashSet evl ← obtain the $L_{max}$ layer extended concept set for the vl
    END FOR
    FOR EACH cl in context_semantic_labels_id
        HashSet ecl ← obtain the $L_{max}$ layer extended concept set for the cl
    END FOR
END FOR
FOR EACH id_i in imageID
    HashSet el_i ← get the extension concepts set of id_i (evl+ ecl)
    FOR EACH id_j in imageID
        HashSet el_j ← get the extension concepts set of id_j (evl+ ecl)
        FOR EACH hs_el_i in el_i
            IF hs_el_i in el_j
                HashSet EP_Gi_Gj ← the shortest path from hs_el_i to el_j
            END IF
        END FOR
        FOR EACH hs_el_j in el_j
            IF hs_el_j in el_i
                HashSet EP_Gj_Gi ← get the shortest path from hs_el_j to el_i
            END IF
        END FOR
    END FOR
    EE_Gi_Gj = EP_Gi_Gj +EP_Gj_Gi ← get the common EE of $G_i$ and $G_j$
END FOR
RETURN EE_Gi_Gj
```

when the extension level is l. i is the distance between the current extension concept and the association concept. EE_i is the set of effective extension concepts when $l = i$. And $sin((1/i)^*(\pi/2))$ is a computational weight function to ensure that higher level effective extension concept has smaller weight.

2) ACQUISITION OF EFFECTIVE EXTENSION CONCEPT BASED ON WORDNET AND DBPEDIA

The concept extension based on WordNet is similar to the concept extension method based on DBpedia. Therefore, this section takes the extension concept of DBpedia knowledge graph as an example to explain how to extend the image association concepts based on the hierarchy of knowledge graph, thus obtaining effective extension concept of any two images. This method is mainly divided into three steps. Firstly, for each image, V^{Gi}_{text} is obtained by the method in Section IV.C, and the concepts in the set are associated with the corresponding concepts in DBpedia knowledge graph. Secondly, through the hierarchy of DBpedia and the value of L_{max}, the concepts in V^{Gi}_{text} are respectively extended to obtain the EC_{Gi}. Finally, we can get the shortest path set $EP_{Gi\rightarrow Gj}$ from G_i to G_j, and the $EP_{Gj\rightarrow Gi}$. The $EE_{Gi,Gj}$ set of the two images can reflect the correlation of the two images from the side, obtained by Equation (3). The two images are linked using paths in the $EE_{Gi,Gj}$ set. The procedure is shown in Algorithm 1.

Fig. 9 is an example (assuming that L_{max}=2), to illustrate how to obtain V^{Gi}_{text} and EC_{Gi} for each image G_i and $EE_{Gi,Gj}$ between two images G_i and G_j through the above three steps. In Fig. 9, *ImageA* is about a silver coin and *ImageB* is about a gold coin. For the above three types of sets, $i = A$ represents the corresponding set of *ImageA* and $i = B$ represents the corresponding set of *ImageB*. Through step one, you can get that both the size of V^{GA}_{text} and V^{GB}_{text} are *3* (Orange Oval). After extension, the size of EC_{GA} is *20* (Green ellipse attached to G_A), and the size of EC_{GB} is *13* (Green ellipse attached to G_B). Finally, an effective extension concept set of the two images is obtained through step *3*, where $EE_{GA,GB}$ is *{Transition metals, 1996 coins, Coins by year}* and $EE_{GB,GA}$ is *{Transition metals, Coins by year}*. Eventually, *ImageA* and *ImageB* can connect via *Transition metals* or *Coins by year*.

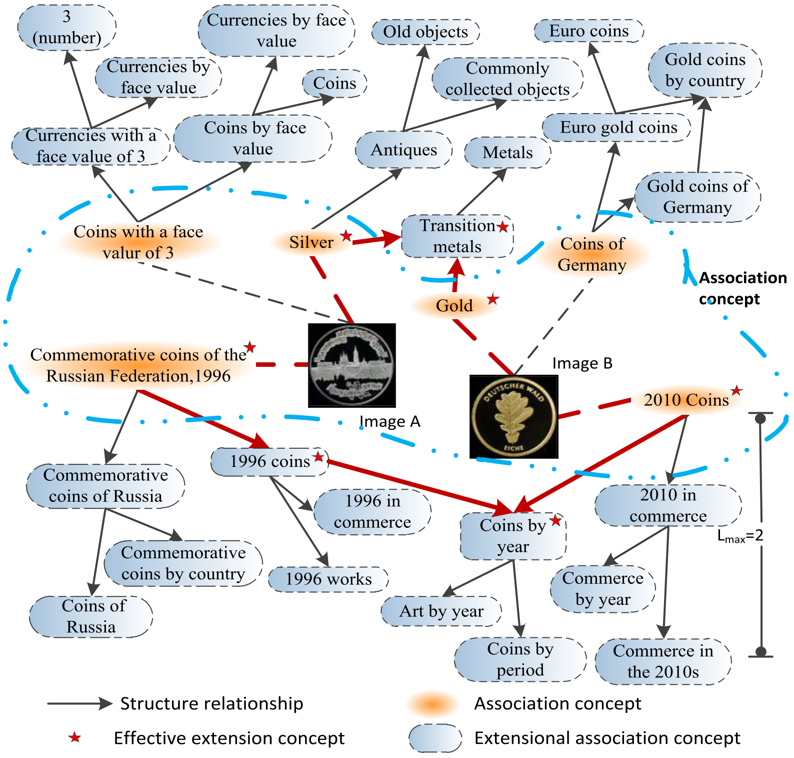

FIGURE 9. Discovery of hidden relationships between images by extending the concept.

TABLE 1. The relationships defined to integrate the images and text.

Relationship	Domain	Range	Type	Extension
MTKG:links2D	Image	DBpedia: resourceName	Image-to-concept	directly
MTKG:links2W	Image	WordNet : resourceName	Image-to-concept	directly
MTKG:name_is	Image	String	Image-to-attValue	-
MTKG:url_is	Image	URL	Image-to-attValue	-
MTKG:Format_is	Image	String	Image-to-attValue	-
MTKG:hyper2D	DBpedia resource	DBpedia resource	Concept-to-concept	indirectly
MTKG:hyper2W	WordNet resource	WordNet resource	Concept-to-concept	indirectly

3) CONSTRUCTION OF THE DOMAIN KNOWLEDGE GRAPH

After the above chapters, we can get four sets of semantic labels of each image G_i, which are V_{text}^{Gi}, V_{vision}^{Gi}, $EE_{Gi,Gj}$, and EC_{Gi}. In addition, the image itself also has its own attributes, such as image format, image size and so on. However, the semantic labels in these sets are discrete. Therefore, this article links them through the relationships which we have defined. There are three types of relationships: *Concept-to-concept*, *Image-to-concept* and *Image-to-attValue*. The types of extension concept are divided into direct extension concept and indirect extension concept. Table 1 shows the types we defined. Among them, *MTKG-**** is the prefix of the relational label. *DBpedia:resourceName* indicates that the semantic label comes from the DBpedia knowledge graph, *WordNet:resourceName* indicates that the semantic label comes from the WordNet, and *resourceName* represents the resource name.

The types defined above show the attributes contained in the knowledge graph we constructed. We use an image as an example to illustrate how to use it, as shown in Fig. 10.

V. EXPERIMENTS

A. EVALUATION ON SOURCE DATA QUALITY

To obtain higher quality source data, a concept dictionary containing 10 common metal concepts was used to obtain source data from Wikimedia Commons, which are *{Copper, Gold, Iron, Manganese, Mercury, Nickel, Silver, Titanium, Tungsten, Zinc}*. Based on these 10 concepts, for each concept C_t, the corresponding image data from Gallery Pages and Category Pages is obtained, respectively. And stores the results obtained from the above two types of pages into R_{G}^{Ci} and R_{C}^{Ci} sets, respectively. The *10* metal concepts obtained a total of 1275 images (each containing relevant text) from Gallery Pages (246 images) and Category Pages.

To evaluate the *Accuracy* of the images and correspondence concepts, 1275 images are used as dataset. For each concept sets R_{G}^{Ci} and R_{C}^{Ci}, each image is evaluated according

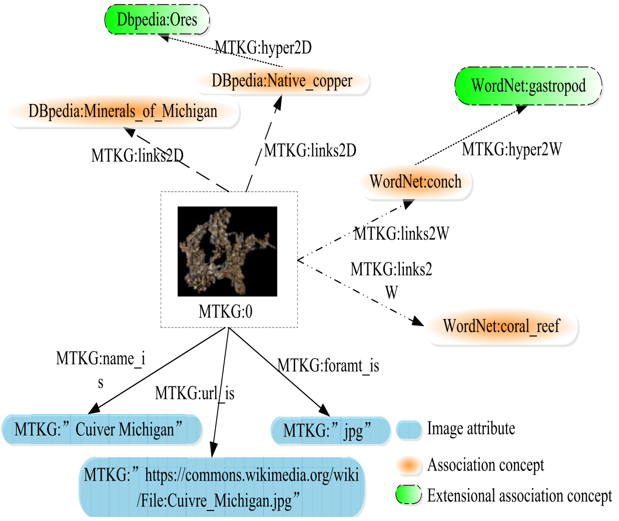

FIGURE 10. Relationship diagram. ***MTKG:LINKS2****: Used to connect direct extension concepts. ***MTKG:HYPER2****: Used to connect indirect extension concepts. and * has a value of ***W*** or ***D***.

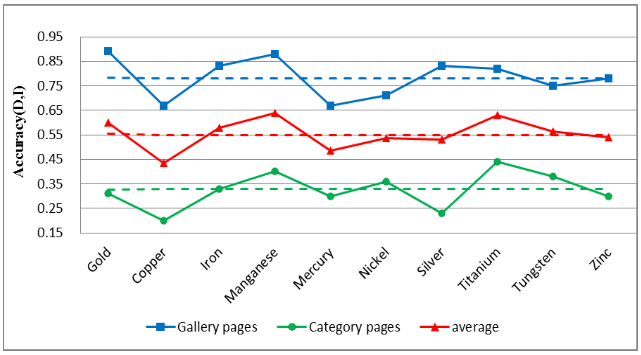

FIGURE 11. The ***Accuracy(D,I)*** of 10 concepts on different pages.

to the rules shown in Equation (8), and finally the *Accuracy* value of each set of images corresponding to the 10 concepts is evaluated, as shown in Fig. 11. For example, for the concept *Gold*, the *Accuracy* in the R_{G}^{Gold} set is *0.89*, and the *Accuracy* in the R_{C}^{Gold} set is *0.31*.

The average *Accuracy (D,I)* value for the 10 concepts in the Gallery page is *0.78* and in the Category pages is *0.33*, and their average *Accuracy (D,I)* value is *0.56*. The results show that the images with 10 concepts obtained from Gallery pages have high *Accuracy (D,I)*, but the number of images is relatively small. This means that if all the image data obtained from the Gallery page and Category page are used as source data, the size of the data will be increased at the expense of *Accuracy (D,I)*. In order to ensure the quality of the experimental data, this paper uses the image data set obtained from Gallery page as the data source of the subsequent experiments, including 246 images, named *246MetImg*.

B. EVALUATION ON THE SEMANTIC CONTEXT RESULT OF THE IAMGE'S ASSOCIATION TEXT

In order to select the appropriate *confidence* of the whole data, we evaluate the results obtained by setting different *confidence* according to Equation (11) and *precision*. The *precision* value of each V_{text}^{Gi} set is the ratio of the number of labeled entities in the semantics of the association text matching the image to the total number of labeled entities under the current confidence level. The *F(cnf)* and *precision* of the overall data are the mean values of *F(cnf)* and *precision* for all images at the current confidence.

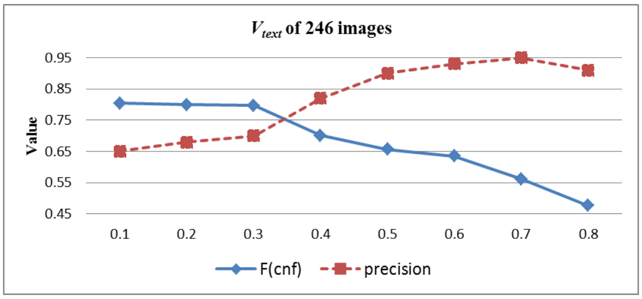

FIGURE 12. ***F(cnf)*** and ***precision*** values for ***Confidence*** between ***0.1*** and ***0.8***.

The calculation result of *F(cnf)* and *precision* is shown in Fig. 12. As the value of *confidence* increases, the value of *F(cnf)* shows a downward trend, while the value of *precision* shows an upward trend. And we also could find that higher *confidence* can improve the *precision* of the DBpedia-Spotlight system, but the cost is the lower recall rate. When the *confidence* is less than *0.3*, the value of *F(cnf)* changes little, but the *precision* increases gradually. In order to obtain the high *precision* result, the *confidence* level is set as *0.3*, where the value of *F(cnf)* is *0.79*, and the *precision* value is *0.70*, which indicates that the context semantic labels of image association text can express the text semantic information of the image very well.

C. EVALUATION ON THE SATISFACTION OF VISUAL SEMANTIC LABELS WITH VGG-IMGNET-RESULT

It has been confirmed that the top-5 classification error of 16-layer-VGG-Net in the ILSVRC-2014-test set is *7.5%*. Since the experimental data and experimental purposes of this paper are different from ILSVRC-2014-test. To this end, the *Satisfaction* designed in Equation (12) is used to evaluate the visual semantic labels of each image obtained by *VGG-ImgNet-Result.*

For the *246MetImg*, we obtain top-5 visual semantic labels for each image. According to the theorem of large numbers, the frequency of random events approximates its probability when repeated many times. Therefore, in terms of statistical satisfaction, three times (i.e. t1, t2 and t3 in Fig. 13) of experiments were carried out randomly. Five groups of samples were selected in each experiment, and the number of each group was (30, 50, 70, 90, 110).

As shown in Fig. 13, by comparing the calculation results, it can be found that as the number of samples increases, the overall *Satisfaction(i,L)* value shows an upward trend. When the number of samples reaches a certain number,

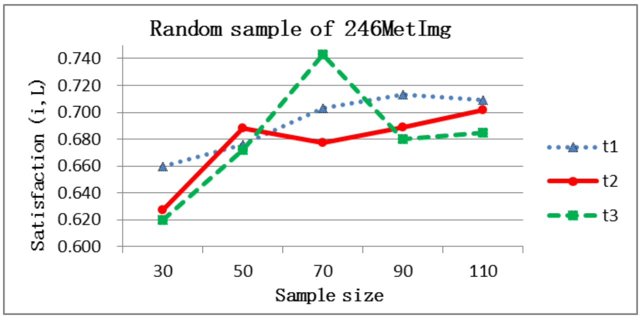

FIGURE 13. Evaluation of the results of VGG-ImgNet-Result processing of image data collected in this article by different sample sizes.

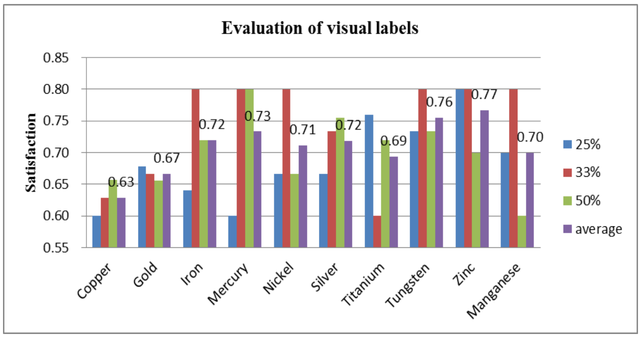

FIGURE 14. The *Satisfactio(i,L)* evaluation of image visual labes in different classes. *N*% represents the proportion of evaluation data in the overall data under the current category. The average represents the average of the three evaluatios dat of each category.

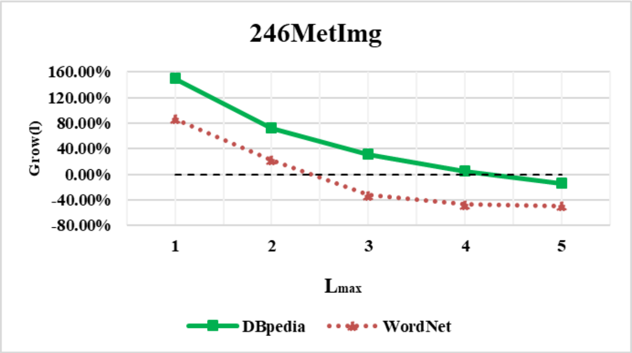

FIGURE 15. The results of the *Grow (l)* values for *246MetImg*.

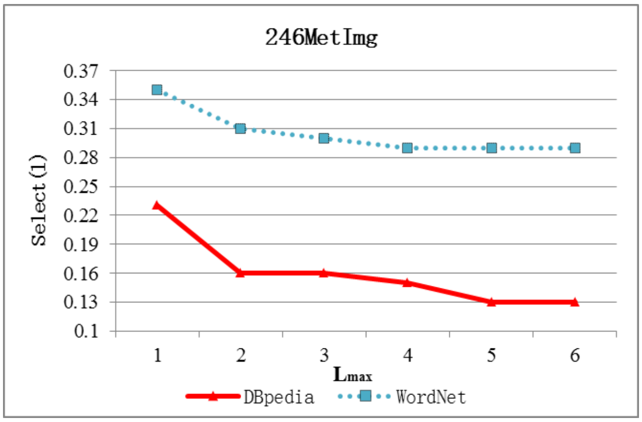

FIGURE 16. The *Select(l)* evaluation of 246MetImg under different L_{max} values.

the value of *Satisfaction(i,L)* begins to stabilize and finally converges between *0.69* and *0.71*. The results show that the overall *Satisfaction(i,L)* with the visual semantic labels of images obtained by the proposed strategy is acceptable.

In order to have an overall understanding of the *Satisfaction(i,L)* of image visual labels, we evaluate the image visual labels in different categories. For each category of images, *25%*, *33%* and *50%* of images were selected to evaluate their *Satisfaction* according to Equation (12). As shown in Fig. 14, the statistics show that the average *Satisfaction(i,L)* of visual labels in each category is about *0.7*. We think the result is acceptable.

D. LMAX VALUE SETTING

In order to set a reasonable maximum extension level (L_{max}) value, this paper uses the extensional association concept, the effective extension concepts and the independent concepts of *246MetImg* to evaluate the effects of different L_{max} from two aspects. We set the L_{max} range from *1* to *6* to select a reasonable L_{max} based on Equations (13) and (14).

We calculate the *Grow(l)* values which is defined in Equation (13) for *246MetImg* with different L_{max}, as shown in Fig. 15. *Grow(l)>0* indicates that there are new independent concepts appears in the extension concepts when extending from the L_{max}*-1* layer to the L_{max} layer; *Grow(l)<=0* indicates that there is no new independent concepts in the extension concepts when extending from the L_{max}*-1* layer to the L_{max} layer. As the value of L_{max} increases, the value of *Grow(l)* keeps decreasing. When *Grow(l)* starts to be less than *0*, it indicates that there are no new independent concepts in the further upward extension, and the hierarchy begins to converge. Therefore, when the concept is extended upward continuously, the large-scale association will appear, and the extension is meaningless at this time. According to Fig. 15, when *Grow(l)>0*, the value of L_{max} should less than *4* for the DBpedia resource and the value of L_{max} should be less than *2* for the WordNet resource.

Like the L_{max} selection of *Grow(l)*, in order to test the relationships between L_{max} and effective extension concepts, we calculate the *Select(l)* values for *246MetImg* with different L_{max}, as shown in Fig. 16. For extension concepts, as the value of L_{max} increases, the value of *Select(l)* keeps decreasing. This indicates that a higher level of extension results in many concepts that are meaningless to data extensions. It also shows that the L_{max} value is not the larger the better. Combining the results of Fig. 15 with Fig. 16, considering the influence of L_{max} on the redundancy of the data, the L_{max} value is set as *3* for DBpedia, and *2* for WordNet.

E. THE PERFORMANCE COMPARISON BEFORE AND AFTER EXTENSION

The purpose of this section is to compare the effects and impacts of path-based concept extension strategies before

TABLE 2. Effective extension results when L_{max} equals to 3 for DBpedia and 2 for WordNet.

Image Objects	Effective extension concepts		Relations
	WordNet	DBpedia	
246	967	11265	120028

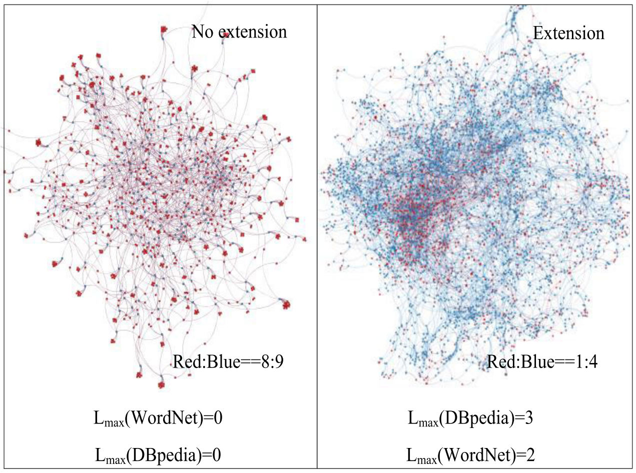
No extension
Extension
Red:Blue==8:9
Red:Blue==1:4
L_{max}(WordNet)=0
L_{max}(DBpedia)=0
L_{max}(DBpedia)=3
L_{max}(WordNet)=2
● (Red) Independent concept
● (Blue) Effective concept

FIGURE 17. Ratio of independent concepts to effective concepts in extension and non-extension data.

and after concept extension. According to the results of subsection V.D, the association concepts of *246MetImg* in the experiment is extended; the association concepts of *246MetImg* were obtained about 10,000 nodes and 100,000 edges in DBpedia and WordNet, as shown in Table 2. These nodes and edges are derived by using the effective extension concepts, enabling images and text to be closely related.

The effectiveness of the new method can be verified by evaluating the characteristics of the constructed knowledge graph from two perspectives, i.e. a macro perspective and a micro perspective.

From the point of macro perspective, we try to analyze the changes of these multi-modal entities as well as the relations between them, and to illustrate the effectiveness of the fusion strategy based on the effective extension concepts. Therefore, the data changes before and after the extension by DBpedia and WordNet are compared. According to the results of subsection V.D, the optimal L_{max} for DBpedia is *3* and for WordNet are *2*. The comparison of the results before and after expansion is shown in Fig. 17.

In Fig. 17, the ratio of the independent concepts to the effective concepts means the strength of the association concept. When the proportion of independent concepts is large, the correlation of the concepts is weak; otherwise it means that the concepts are highly correlated. When a fusion strategy based on the effective extension concepts is not used, the ratio of the independent concepts to effective concepts inside the Knowledge graph is *8:9*, indicating that there are many independent concepts and the data correlation is weak. After the concept extension through the effective extension concepts fusion strategy, the ratio of independent concepts to effective concepts within the data is *1:4*, indicating that there are fewer independent concepts and the data is highly correlated. In addition, Fig. 17 also illustrates that based on the effective extension concepts fusion strategy significantly increases the scale of the effective extension concepts.

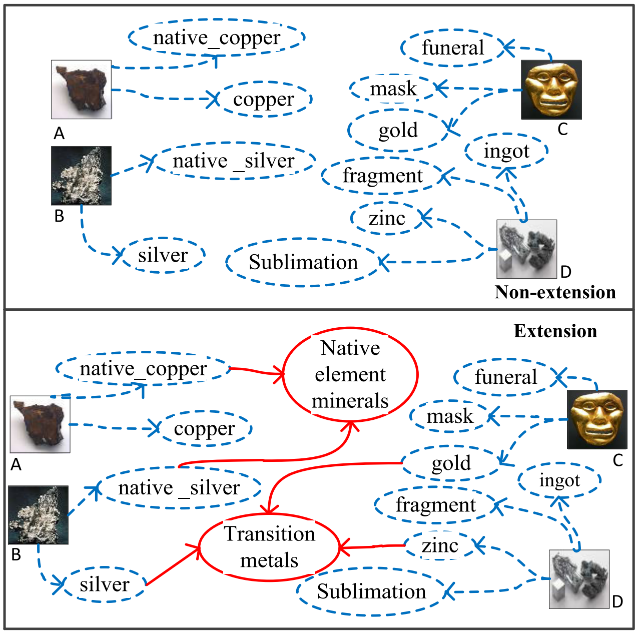

FIGURE 18. After concept extension, some images are associated with each other by effective extension concepts.

From the point of micro perspective, we randomly extract a set of image samples and observe the changes of data associated with these images before and after using extension. Fig. 18 shows four randomly image samples, named A, B, C, and D, respectively. The four images were not significantly related to each other. After using the effective extension concepts fusion strategy, Image A and Image B are related by the effective extension concept *Native element minerals*, while Image B, Image C and Image D are related by the effective extension concept *Sublimation*, which show that our strategy brings significant positive changes to the data.

In order to further illustrate that the characteristics of the knowledge graph, we try to compare the constructed multi-modal knowledge graph with IMGpedia. As shown in Fig. 19, a set of randomly sampled images and their association entities in the datasets are acquired and compared. It is not difficult to see that the sampled images are generally associated with few resources in IMGpedia, but they are associated with richer resources in the data set generated in this paper, and many resources have applied value because they can describe the semantic content of the image or visual characteristics. In addition, IMGpedia does not include the upper concept of the entity. For the entities associated with the same image, the datasets in this paper can provide hierarchical relationships for these entities. By the comparison with IMGpedia data, it is further verified that the effective extension concepts fusion

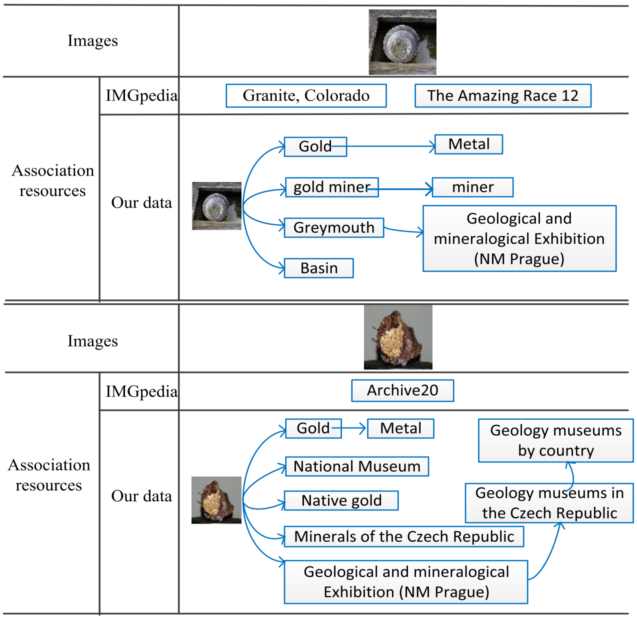

FIGURE 19. The comparison of the association resources of images in our data and IMGpedia.

strategy designed in this paper can not only improve the relevance of data, but also improve the richness of data.

VI. CONCLUSION AND FUTURE WORK

We proposed a method to construct domain-specific multi-modal knowledge graph. Based on the resources of Wikimedia Commons, the scope of the field is defined by a manually constructed concept dictionary to obtain the domain-specific images and association texts. By acquiring the context semantic labels of association text and the visual semantic labels of image, those semantic labels are extended by using the concept hierarchy of DBpedia and WordNet. And a method of limiting the level of concept extension is designed. The effective extension concepts of images are obtained by the proposed method, and the domain-specific knowledge graph is constructed by taking the metallic materials domain as an example. Compared with IMGpedia, this method can not only obtain many semantic labels related to image, but also make semantic relations between discrete images.

Some new issues have gradually emerged during the implementation of the method. First, we only use the common concepts in two sets of effective extension concepts to connect two images, and the relationships generated between the images may be incomplete. We think the method based on similarity computation or representation learning [56] may improve our method and discover more new links in the domain knowledge graph. Second, the relationships defined in the constructed knowledge graph are fixed, thus lacking practical semantics. Furthermore, since the label of ImageNet data set is not adjusted according to the label of metallic materials domain when training the VGG-Net model, it may lead to the problem that the predicted label does not conform to the metallic materials domain.

The domain-specific multi-modal domain knowledge graph constructed in this paper can be used in the applications such as Question Answering System in the field of metallic materials. It can provide the support to implement the cross-modal retrieval with this image-text knowledge. Moreover, it is also helpful to calculate the semantic similarity between images in metallic materials domain.

In the future, we will explore the following research directions: (1) we will try to discover new relationships between entities in the generated domain knowledge graph to implement the knowledge graph completion. (2) We will explore the details of how to combine multi-modal knowledge base with Visual Question Answering system, so as to show more attractive practical significance of the multi-modal knowledge graph. (3) We will continue to extend the method to make it suitable to construct general knowledge graphs. (4) We will use domain-specific labels to fine-tune the labels of the image data to improve the satisfaction of the visual detection model for predicting labels.

REFERENCES

[1] S. Zhao and Q. Qian, "Ontology based heterogeneous materials database integration and semantic query," *AIP Adv.*, vol. 7, no. 10, p. 105325, Oct. 2017. doi: 10.1063/1.4999209.

[2] X. Zhang, P. Lv, and J. Wang, "STSM: An infrastructure for unifying steel knowledge and discovering new knowledge," *Int. J. Database Theory Appl.*, vol. 7, no. 6, pp. 175–190, Dec. 2014. doi: 10.14257/ijdta.2014.7.6.16.

[3] X. Zhang, Z. Zhang, H. Wang, M. Meng, and D. Pan, "Metallic materials ontology population from LOD based on conditional random field," *Comput. Ind.*, vol. 99, pp. 140–155, Aug. 2018. doi: 10.1016/j.compind.2018.03.032.

[4] X. Zhang, D. Pan, C. Zhao, and K. Li, "MMOY: Towards deriving a metallic materials ontology from Yago," *Adv. Eng. Informat.*, vol. 30, no. 4, pp. 687–702, Oct. 2016. doi: 10.1016/j.aei.2016.09.002.

[5] T. Ashino, "Materials ontology: An infrastructure for exchanging materials information and knowledge," *Data Sci. J.*, vol. 9, pp. 54–61, Jul. 2010. doi: 10.2481/dsj.008-041.

[6] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6077–6086.

[7] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, vol. 211, pp. 457–468, Nov. 2016. [Online]. Available: http://aclweb.org/anthology/D16-1044

[8] H. Ben-younes, R. Cadene, M. Cord, and N. Thome, "MUTAN: Multimodal Tucker fusion for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2631–2639.

[9] S. Ferrada, B. Bustos, and A. Hogan, "IMGpedia: A linked dataset with content-based analysis of wikimedia images," (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Cham, Switzerland: Springer, 2017, pp. 84–93.

[10] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer, "DBpedia—A large-scale, multilingual knowledge base extracted from wikipedia," *Semantic Web*, vol. 6, no. 2, pp. 167–195, 2015. doi: 10.3233/SW-140134.

[11] X. Zhang, X. Liu, X. Li, and D. Pan, "MMKG: An approach to generate metallic materials knowledge graph based on DBpedia and Wikipedia," *Comput. Phys. Commun.*, vol. 211, pp. 98–112, Feb. 2017. doi: 10.1016/j.cpc.2016.07.005.

[12] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[13] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer, "DBpedia spotlight: Shedding light on the Web of documents," in *Proc. 7th Int. Conf. Semantic Syst. I-Semantics*, 2011, pp. 1–8.

[14] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' in *Proc. ICLR*, San Diego, CA, USA, 2015.

[15] M. Lenz, R. Krug, C. Dillmann, R. Stroop, and N. C. Gerhardt, ''Automated differentiation between meningioma and healthy brain tissue based on optical coherence tomography *ex vivo* images using texture features,'' *J. Biomed. Opt.*, vol. 23, no. 7, p. 1, Feb. 2018. doi: 10.1117/1.JBO.23.7.071205.

[16] M. Parchami, S. Bashbaghi, and E. Granger, ''Video-based face recognition using ensemble of haar-like deep convolutional neural networks,'' in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2017, pp. 4625–4632.

[17] K. W. Chee and S. S. Teoh, ''Pedestrian detection in visual images using combination of HOG and HOM features,'' in *Proc. 10th Int. Conf. Robot., Vis., Signal Process. Power Appl.*, in Lecture Notes in Electrical Engineering, 2019, pp. 591–597.

[18] R. Yu, A. Li, V. I. Morariu, and L. S. Davis, ''Visual relationship detection with internal and external linguistic knowledge distillation,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1068–1076.

[19] B. A. Plummer, A. Mallya, C. M. Cervantes, J. Hockenmaier, and S. Lazebnik, ''Phrase localization and visual relationship detection with comprehensive image-language cues,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1946–1955.

[20] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, ''Visual relationship detection with language priors,'' in *Computer Vision—ECCV* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Cham, Switzerland: Springer, 2016, pp. 852–869.

[21] L. F. Rau, ''Extracting company names from text,'' in *Proc. 7th IEEE Conf. Artif. Intell. Appl.*, vol. 1, Feb. 2002, pp. 29–32.

[22] R. Gabbard, J. DeYoung, C. Lignos, M. Freedman, and R. Weischedel, ''Combining rule-based and statistical mechanisms for low-resource named entity recognition,'' *Mach. Translation*, vol. 32, nos. 1–2, pp. 31–43, Jun. 2018. doi: 10.1007/s10590-017-9208-0.

[23] P. J. Gorinski, H. Wu, C. Grover, R. Tobin, C. Talbot, H. Whalley, C. Sudlow, W. Whiteley, and B. Alex, ''Named entity recognition for electronic health records: A comparison of rule-based and machine learning approaches,'' 2019, *arXiv:1903.03985*. [Online]. Available: https://arxiv.org/abs/1903.03985

[24] N. Greenberg, T. Bansal, P. Verga, and A. Mccallum, ''Marginal likelihood training of BiLSTM-CRF for biomedical named entity recognition from disjoint label sets,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 2824–2829.

[25] A. Liu, J. Du, and V. Stoyanov, ''Knowledge-augmented language model and its application to unsupervised named-entity recognition,'' Jun. 2019, *arXiv:1904.04458v2*. [Online]. Available: https://arxiv.org/abs/1904.04458

[26] A. Katiyar and C. Cardie, ''Nested named entity recognition revisited,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2018, pp. 861–871.

[27] P. Corbett and J. Boyle, ''Chemlistem: Chemical named entity recognition using recurrent neural networks,'' *J. Cheminf.*, vol. 10, p. 59, Dec. 2018. doi: 10.1186/s13321-018-0313-8.

[28] Y. Xin, J. D. Ruvini, and E. J. Hart, ''Deep hybrid neural network for named entity recognition,'' U.S. Patent 692 392, Feb. 28, 2019.

[29] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, ''Distant supervision for relation extraction without labeled data,'' in *Proc. Joint Conf. 47th Annu. Meeting ACL 4th Int. Joint Conf. Natural Lang. Process. (AFNLP)*, vol. 2, Aug. 2009, pp. 1003–1011.

[30] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, ''Freebase?: A collaboratively created graph database for structuring human knowledge,'' in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2008, pp. 1247–1250.

[31] F. M. Suchanek, G. Kasneci, and G. Weikum, ''YAGO: A large ontology from wikipedia and WordNet,'' *J. Web Semantics*, vol. 6, no. 3, pp. 203–217, 2008.

[32] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang, ''Knowledge vault: A Web-scale approach to probabilistic knowledge fusion,'' in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2014, pp. 601–610.

[33] W. Wu, H. Li, H. Wang, and K. Q. Zhu, ''Probase: A probabilistic taxonomy for text understanding,'' in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2012, pp. 481–492.

[34] Z. Wang, H. Wang, J.-R. Wen, and Y. Xiao, ''An inference approach to basic level of categorization,'' in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2015, pp. 653–662.

[35] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, ''The SIDER database of drugs and side effects,'' *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1075–D1079, Jan. 2016. doi: 10.1093/nar/gkv1075.

[36] A. Swartz, ''MusicBrainz: A semantic Web service,'' *IEEE Intell. Syst.*, vol. 17, no. 1, pp. 76–77, Jan. 2002. doi: 10.1109/5254.988466.

[37] L. E. Blakesley, ''The Internet movie database (IMDb),'' *Electron. Resour. Rev.*, vol. 3, no. 5, pp. 56–57, 1999.

[38] R. Navigli and S. P. Ponzetto, ''BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network,'' *Artif. Intell.*, vol. 193, pp. 217–250, Dec. 2012.

[39] V. S. Silva, A. Freitas, and S. Handschuh, ''Building a knowledge graph from natural language definitions for interpretable text entailment recognition,'' Jun. 2018, *arXiv:1806.07731*. [Online]. Available: https://arxiv.org/abs/1806.07731

[40] C. B. Wang, X. Ma, J. Chen, and J. Chen, ''Information extraction and knowledge graph construction from geoscience literature,'' *Comput. Geosci.*, vol. 112, pp. 112–120, Mar. 2018. doi: 10.1016/j.cageo.2017.12.007.

[41] T. Yu, J. Li, Q. Yu, Y. Tian, X. Shun, L. Xu, L. Zhu, and H. Gao, ''Knowledge graph for TCM health preservation: Design, construction, and applications,'' *Artif. Intell. Med.*, vol. 77, pp. 48–52, Mar. 2017. doi: 10.1016/j.artmed.2017.04.001.

[42] A. Mehta, A. Singhal, and K. Karlapalem, ''Scalable knowledge graph construction over text using deep learning based predicate mapping,'' in *Proc. Companion World Wide Web Conf. (WWW)*, May 2019, pp. 705–713.

[43] R. A. Al-Zaidy and C. L. Giles, ''Extracting semantic relations for scholarly knowledge base construction,'' in *Proc. IEEE 12th Int. Conf. Semantic Comput. (ICSC)*, Apr. 2018, pp. 56–63.

[44] X. Zhao, Z. Xing, M. A. Kabir, N. Sawada, J. Li, and S.-W. Lin, ''HDSKG: Harvesting domain specific knowledge graph from content of Webpages,'' in *Proc. IEEE 24th Int. Conf. Softw. Anal., Evolution Reeng. (SANER)*, Feb. 2017, pp. 56–67.

[45] K. Zhang, K. Li, H. Ma, D. Yue, and L. Zhuang, ''Construction of MeSH-like obstetric knowledge graph,'' in *Proc. Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discovery (CyberC)*, Feb. 2018, pp. 160–1608.

[46] P. Chen, Y. Lu, V. W. Zheng, X. Chen, and B. Yang, ''KnowEdu: A system to construct knowledge graph for education,'' *IEEE Access*, vol. 6, pp. 31553–31563, 2018. doi: 10.1109/ACCESS.2018.2839607.

[47] R. Xie, Z. Liu, H. Luan, and M. Sun, ''Image-embodied knowledge representation learning,'' in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2017, pp. 3140–3146.

[48] P. Pezeshkpour, L. Chen, and S. Singh, ''Embedding multimodal relational data for knowledge base completion,'' in *Proc. EMNLP*, Brussels, Belgium, 2018, pp. 3208–3218.

[49] S. Thoma, A. Rettinger, and F. Both, ''Towards holistic concept representations: Embedding relational knowledge, visual attributes, and distributional word semantics,'' in *The Semantic Web—ISWC* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2017, pp. 694–710.

[50] S. Thoma, A. Rettinger, and F. Both, ''Knowledge fusion via embeddings from text, knowledge graphs, and images,'' Apr. 2017, *arXiv:1704.06084*. [Online]. Available: https://arxiv.org/abs/1704.06084

[51] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, ''Rethinking the inception architecture for computer vision,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[52] T. Mikolov, K. Chen, G. Corrado, and J. Dean, ''Efficient estimation of word representations in vector space,'' in *Proc. ICLR*, Scottsdale, Arizona, 2013.

[53] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, ''Translating embeddings for modeling multi-relational data,'' in *Advances in Neural Information Processing*, vol. 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Q. Weinberger, Eds. New York, NY, USA: Curran Associates, 2013, pp. 2787–2795.

[54] H. Mousselly-Sergieh, T. Botschen, I. Gurevych, and S. Roth, ''A multimodal translation-based approach for knowledge graph representation learning,'' in *Proc. 7th Joint Conf. Lexical Comput. Semantics*, 2018, pp. 225–234.

[55] F. Huang, X. Zhang, J. Xu, C. Li, and Z. Li, ''Network embedding by fusing multimodal contents and links,'' *Knowl.-Based Syst.*, vol. 171, pp. 44–55, May 2019.

[56] Y. Bengio, A. Courville, and P. Vincent, ''Representation learning: A review and new perspectives,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013. doi: 10.1109/TPAMI.2013.50.

XIAOMING ZHANG received the master's degree from Hebei University, China, in 2002, and the Ph.D. degree from the University of Science and Technology Beijing, China, in 2009, both in computer application. He is currently a Professor with the School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, China. His main research interests include knowledge graph, semantic computing, domain-specific information integration, and multi-modal knowledge fusion.

XIAOLING SUN was born in Shijiazhuang, Hebei, China, in 1994. She is currently pursuing the M.S. degree in computer science and technology with the Hebei University of Science and Technology, Shijiazhuang, Hebei, China. Her current research interests include knowledge graph, domain-specific information integration, multi-modal knowledge fusion, and semantic computing.

CHUNJIE XIE received the M.S. degree in computer science and technology from the Hebei University of Science and Technology, Shijiazhuang, Hebei, China, in 2019. His main research interests include knowledge graph, domain-specific information integration, and multi-modal knowledge fusion.

BING LUN was born in Xingtai, Hebei, China, in 1994. He is currently pursuing the M.S. degree in computer science and technology with the Hebei University of Science and Technology, Shijiazhuang, Hebei, China. His research interests include knowledge graph, multi-modal knowledge fusion, knowledge representation learning, and natural language processing.