IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 20, 2019, accepted August 1, 2019, date of publication August 5, 2019, date of current version August 19, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2933249

TriTag-NFPF: Knowledge Denoising for Chinese Encyclopedia based on Triple Tag-Constructed Potential Function

TING WANG[1], HANZHE GU[1], JIE LI[1], AND JINGYAO XIE[2]
[1]School of Management and Engineering, Capital University of Economics and Business, Beijing 100070, China
[2]State Grid Beijing Electric Power Company, Beijing 100031, China

Corresponding author: Ting Wang (wangting@cueb.edu.cn)

This work was supported in part by the Scientific Research Project of Beijing Municipal Education Commission (General Social Science Project) under Grant SM201910038010, and in part by the Youth Excellent Teachers Grant of Capital University of Economics and Business under Grant 23491854840429.

ABSTRACT In this paper, a novel method is proposed for Chinese large-scale online encyclopedia knowledge denoising. Firstly, the initial similarity of the triples is acquired by the similarity computing method integrating the Edit-Distance and TongYiCiCiLin similarity algorithm. Secondly, a novel nuclear field-like potential function of the Infobox knowledge triples is constructed in virtue of Chinese encyclopedia entry semantic tag. Finally, large-scale knowledge triple clustering and denoising are performed by means of the improved potential function proposed in this paper for the purpose of minimizing the influence of massive repetition and ambiguity in the Chinese open encyclopedia Knowledge Base (KB). The proposed method has solved the problems of semantic duplication, ambiguity and inappropriate classification of knowledge triples arising from constructing Chinese KBs. The experimental results indicate that the open-domain oriented Chinese encyclopedia KBs constructed by the method proposed in this paper is outperformed than the state-of-the-art methods.

INDEX TERMS Knowledge base, online encyclopedia, knowledge denoising, similarity computing, nuclear field-like potential function.

I. INTRODUCTION

The vision of the Semantic Web is to create a ''Web of Data'', so that a machine is able to understand the semantic information on the internet [1]. With the development and application of internet technology, it has gradually become an open platform for information release, communication and sharing. Information inquiry and knowledge acquisition have been gradually transformed from offline to online. People's life has entered into the era of big data and knowledge graph (KG) due to the rapid development of Web 3.0. Construction of knowledge base (KB) which serves as the collection of important knowledge triples for storing, organizing, processing and providing knowledge services, is becoming the foundation for many industries to carry out knowledge management.

With the increasingly diverse means of KBs construction, the tendency of large-scale size can be seen. The flourishing online encyclopedia provides a high-quality data source for knowledge discovery, integration and KG construction. With the maturing and emerging of Chinese online encyclopedia, the academic circle has started focusing and researching for automatic knowledge extraction and KB construction. However, a large number of synonymy, ambiguity and improper classification of knowledge in Chinese entries have resulted in problems of low efficiency and precision of Chinese online encyclopedia KB.

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

II. RELATED WORK

In order to achieve the vision of ''Web of Data'', in the current decade, the research on Semantic Web has undergone vigorous development of KB construction, linked open data (LOD) and KG.

A. KNOWLEDGE BASE

The birth and widespread using of online encyclopedia has provided abundant knowledge sources for constructing the KB. In recent years, researches on large-scale knowledge

acquisition for online encyclopedia systems which are in the ascendant and mainly depend on two aspects: i) knowledge extraction by parsing the structured or semi-structured information from entry's web page, such as Infobox, and ii) knowledge extraction from the plain-text of entry's web page based on semantic annotation and machine learning algorithms.

The large-scale KB (e.g., DBpedia) based on the online encyclopedia system is also deemed as the source of online knowledge. Kylin system developed by Wu and Weld [2], [3] not only focuses on the structured information arising from the web page of encyclopedia items, but also tries to extract the knowledge triples from the unstructured text. Freebase [4] is a practical, scalable triple database used to structure commonsense knowledge. It not only extracts information from Wikipedia, but also makes use of the MusicBrainz and Notable Names Database etc. Some other typical works such as YAGO [5], [6] and DBpedia [7], [8], which are large-scale and multilingual semantic KBs built on the basis of Wikipedia system, extracting knowledge from the structured information of Wikipedia pages through specific wiki system middleware. The hierarchical relation in encyclopedia open classification system is confirmed as the ''is-a'' relationship between concepts. The Infobox on entry page contains a number of knowledge triples. DBpedia as the Hub of LOD [9], a multilingual and giant LOD Web has been established.

At the same time, as the second largest language in the world, Chinese KB plays an important part in the LOD and KG constructions, therefore many researches on Chinese KB are gradually unfolded. By extracting knowledge triples and constructing large-scale KB from the structured information in Chinese online encyclopedia system, Wang *et al.* extract the hierarchical relationship between concepts and obtain the Infobox knowledge triples in the entry's web page and encyclopedia entries based on the classification system of Chinese encyclopedia, and finally establish a Chinese encyclopedia KB [10]. Jingwei+ [11] is; Fu *et al.* [12] crawl the entry's web pages of BaiduBaike and Hudong through semantic crawler, then parse the structured information (e.g., Infobox) and generate the RDF triples; secondly, download the Chinese version of DBpedia KB; finally store the three largest Chinese encyclopedia KBs into Jingwei+ (a large-scale distributed Chinese KB [11]). However, in the case of ambiguity in the data sets, it cannot achieve integration. Wang *et al.* [13], [14] propose a new approach of automatic building for domain ontology based on machine learning algorithm, and by which the large-scale e-Gov ontology is built automatically. In the first stage, rough mapping from thesaurus to ontology is carried out, and domain rough ontology is formed according to the conversion rules proposed. In the second stage, they firstly merge and align concepts between thesaurus and encyclopedia KB based on Edit-Distance and TongYiCi-CiLin algorithm. Secondly, based on the TF-IDF algorithm, the structured Infobox knowledge triples are sorted, refined and automatically merged with thesaurus to build domain ontology with rich semantic information. Li *et al.* [15] use twelve effective features of Chinese Wikipedia tag data which are designed from the perspectives of lexical, grammatical and structural aspects to predict and extract is-a relationship. The Skip-gram model is used for training word embedding, mining and descripting of semantic relationships in the web page. The is-a relation inference method based on language pattern, heuristic rules and association rules mining is proposed, and the upper concept of classification tree is extracted. In the end, a large-scale Chinese classification system construction algorithm is proposed to design and implement the Chinese classification system query system: CTCS2 based on constructing the classification system so as to meet the requirements of semantic query.

The method of automatic information annotation and acquisition by supervised learning in plain-text of Chinese Wikipedia is earlier proposed by Chen *et al.* They extract the large-scale Chinese knowledge triples from the unstructured text of the Wikipedia based on the Infobox and statistical learning algorithm [16]. Liu *et al.* [17] propose a synonym decision method to extend the set of attributes, which is based on the characteristics of attribute phrases and make full use of structural features of Chinese online encyclopedia to discover different expressions of attributes. In order to enhance the precision of KB, Wang *et al.* [18] propose a self-expanded learning method to predict on the semantic relations between subjects and objects while extracting the knowledge triples from the plain-texts of entry's web page of Chinese encyclopedia.

B. LINKED OPEN DATA

In recent years, with a growing number of heterogeneous large-scale KBs published on the web and in order to achieve knowledge sharing and semantic interoperation, there are many researches in the field of LOD mainly focus on the following two aspects: i) interlinking data sets on the schema-level that is ontology mapping, and ii) interlinking data sets on the level of instances.

As a typical scenario of LOD on schema-level, ontology mapping has been widely studied. Melnik *et al.* propose a structural level ontology mapping algorithm called ''similarity flooding'' that uses the concept of ontology to build a similarity propagation map and then spreads and corrects any similarities between the concepts. Cohen *et al.* [19] survey some typical similarity computing algorithms on element-level and evaluate their performance. Giunchiglia and Yatskevich [20] propose that semantic relationships should be discovered based on linguistic method by introducing shared knowledge dictionary (e.g., WordNet [21]). Isaac *et al.* [22] propose an Instance-Level ontology mapping algorithm that would measure the similarities between the concepts according to the number of similar instance of ontology concepts. Nikolov *et al.* [23] make use of the hierarchy between concepts in order to pick up the most suitable mapping parameter and approach. Zhong *et al.* [24] develop the RiMOM system, which is a multi-strategy mapping system based on ontology instance, concept name, ontol-

ogy structure and other characteristics. Jain *et al.* propose the BLOOMS system [25], which can efficiently build the Schema-Level interlinking in LOD environment. Furthermore, Tongsacom [26] is a sequence alignment-based Chinese ontology mapping model, it can deal with the Chinese out-of-vocabulary mapping task.

There are also some typical systems establishing the interlinking between LOD datasets at the instance level. Silk [27], [28] is a framework for building the interlinking between different datasets by a declarative language. Users can configure the linking strategy, such as the type and conditions, and it also support the remote interlinking. RDF-AI [29] is a matching framework that can match, fuse and interlink RDF datasets based on sequence alignment similarity algorithms. Hassanzadeh *et al.* [30] propose a common and extensible system: LinQL, which has integrated some existed link discovery method. The purpose of this system is to help users to select the most suitable method to interlink their own dataset. Meanwhile, it also support publishing RDF triples from Relational Data Base (RDB) by using D2RQ [31] and Virtuoso. In order to achieve the sharing, reusing, and interoperation of knowledge bases in LOD environment, it is necessary to link ontologies described in different languages. Ngomo and Auer [32] rely on user-defined rules to determine the attributes to be compared between the entities. HolisticEM [33] construct a graph of potential matching pairs firstly and aligns instances in KBs based on Personalized PageRank. Earlier, Niu *et al.* [34] use the rich source data of Chinese encyclopedia to construct the Chinese semantic KB and developed a Chinese linked data application system: Zhishi.me. Specifically, a semi-supervised learning algorithm is proposed to iteratively mine matching rules and find equivalent semantic relationships [35]. This method greatly reduces the cost of manual design matching rules and similarity computing, and still maintains high precision. Wang *et al.* [36] propose extracting the hierarchical relationships between concepts and the concept property contained in Infobox based on the DMOZs (a kind of open classification project) of Chinese wiki: BaiduBaike[1] and Hudong[2]; furthermore, based on a simple keyword-matching method, eventually building Chinese encyclopedia KBs and establishing the co-reference relationships between instances of DBpedia. Wang *et al.* [37] propose a multi-source KB entity alignment algorithm based on semantic tag of entry, which can align Chinese encyclopedia entities by comprehensive usage of attribute tags, category tags and keywords of plain-texts. There are also some typical researches on the cross-linguistic links, e.g., Wang *et al.* [38] [39] propose a method for building cross-linguistic interlinking in LOD. Firstly, with the help of a small amount of cross-language and internal links seed, using the concept annotation method to enrich the internal links; secondly, a regression model is used to predict potential cross-language links between Chinese and English wiki. Wang *et al.* [40] link articles between Wikipedia and BaiduBaike by using a bilingual topic model and translation features based on SVM.

[1]http://baike.baidu.com/

[2]http://www.hudong.com/

C. KNOWLEDGE GRAPH

Nowadays, KBs in the form of KGs have been widely used in many applications. Research on the Semantic Web has entered into the era of KGs. But for similar reasons and same as KBs, many KGs have been created separately for particular purposes. Furthermore, KG embedding models have been widely applied to address KG completion tasks that aim to predict missing entities or relations based on existing triples in a KG. There are some typical translation model-based methods have been proposed to learn entity embeddings. TransE [41] models relationships by interpreting them as translations operating on the low-dimensional embeddings of the entities. TransH [42] models a relation as a hyperplane together with a translation operation on it. TransR [43] builds entity and relation embeddings in separate entity space and relation spaces. TransD [44] not only considers the diversity of relations, but also entities. It can be applied on large scale graphs. Xie *et al.* [45] explore two encoders for knowledge graphs taking advantages of entity descriptions, including continuous bag-of-words and deep convolutional neural models to encode semantics of entity descriptions. Wang and Li [46] take advantage of rich context information in a corpus and expand the semantic structure of the knowledge graph. This method can better handle the 1-to-N, N-to-1 and N-to-N relations. Trisedya *et al.* [47] exploit large numbers of attribute triples in the knowledge graphs and generate attribute character embeddings. They use a transitivity rule to further enrich the number of attributes of an entity to enhance the attribute character embedding.

Besides, some non-translation based approaches have also been proposed to learn entity embeddings. Bordes *et al.* [48] use neural network learning method to embed triples in different KBs into a flexible continuous vector space. Socher *et al.* [49] discover that performance can be improved when entities are represented as an average of their constituting word vectors. HolE [50] also use tensor-based factorization and represent relations with matrices.

To sum up, having a high-quality and high-precision KB is the premise and guarantee to realize the vision of LOD and KGs. Unfortunately, there are a large amount of noise knowledge still in Chinese online encyclopedia KBs brought by previous work, which is mainly caused by the ambiguity and inappropriate classification of Infobox triples due to the open collaborative characteristics and tag settings etc. of online encyclopedia. Therefore, this paper proposes a novel knowledge denoising model to enhance the precision of encyclopedia KBs. The major contributions of this paper are listed as following:

(1) Our method employs the Edit-Distance [51] and TongYiCiCiLin algorithm for computing the initial similarity value of triples in Chinese encyclopedia KBs, and we

separately regard the triples as the data entities in field and the initial similarity value as the mass of triple in data field.

(2) In order to further improve the precision of KB, a novel Infobox knowledge triple-constructed nuclear field-like potential function is proposed by means of initial similarity value and distance between semantic tags of Chinese encyclopedia entry to compute the target similarity value of the triples. Specifically, we propose a novel piecewise function based on semantic distance between triple's tags and embed it into the nuclear field-like potential function. The proposed piecewise function is used for punishing the improper classified triples so as to optimize and decrease its initial similarity value in its triples set.

(3) The processes of re-ranking and deleting of lower-ranked knowledge triples are carried out based on the target similarity value for the purpose of denoising a large number of improper classification and ambiguity of triples in Chinese encyclopedia KBs.

III. PROBLEM DESCRIPTION AND DEFINITION

A. PROBLEM DESCRIPTION

According to the classification system of encyclopedia, each entry can belong to one or more classification concepts. Obviously, these entries are instances of the classification concept. The structured information of Infobox appearing on each entry's web page can be parsed into a set of knowledge triples, that is to say, all triples in this set belong to the instance of entry. The 11 top-classes in BaiduBaike and Hudong refer to the classification tree of both of them, namely: People, Sports, Life, Culture, Science, Economy, History, Society, Geography, Nature, and Art.

The semantic tags discussed in this paper are divided into two parts:

i) The one part is the encyclopedia entries' tags labeled by its open collaborative editors. Especially in the BaiduBaike, there are two kinds of tags for entries: classification tag and property tag. These two kinds of tags are not distinguished on the web page of entries. Therefore, entry's tags of BaiduBaike can be used to describe both the classification and the properties of entries.

ii) The other part is the misclassified-tags labeled by the previous work because of the ambiguity caused by a large number of encyclopedia entries which have the same name but different meanings. These tags are named as encyclopedia ambiguous tag. The ambiguous categorization of entries will eventually result in all the Infobox knowledge triples contained in the entries being labeled the ambiguous tags.

In this paper, tags in the two parts mentioned above are collectively called as semantic tag. They form a set T of semantic tags, in which each tag appearing in the classification tree of Chinese encyclopedias.

There are two possible cases where noise triples would be generated. Below we will take the BaiduBaike's entry ''Verona'' for example to explain:

i) Incorrect categorization may occur when entries are categorized according to entry's tags, that is: if an entry contains a tag describing its properties, then all the triples of this entry will be inappropriately classified under the concept of the classification tree corresponding to the property tag.

For example, the ''Sports'' tag is used to characterize the properties of entry: ''Verona''(a famous football club in Italy) rather than to declare the classification to which it belongs.

ii) The inappropriate categorization of these ambiguity entries will result in the ambiguous categorization of all the Infobox knowledge triples contained in the entry. Besides, because of open collaborative editing mode of online encyclopedia, it should be noted that the encyclopedia tags of the part i) of the entry itself may also contain some improperly categorized tags by the entry's editors, which will also lead to inappropriate categorization of the triples.

For example, the entry ''Verona'' has five homonymous terms, these ambiguity entries belong to five different Encyclopedia sub-classes (tags): Scenic Spots, Geography, Italy, Football and Sports. Just as said in part i) in this section, some of these tags are used for describing the property of entry ''Verona''. Therefore, when the previous work system classifies the entry, it will cause the ambiguity of the entry itself, which will lead to all the ambiguity entries being classified to every sub-class in the set T of semantic tags. That is, causing the following improper classified results:

Verona(Scenic Spots) → Scenic Spots, Geography, Italy, Football, Sports
Verona(Geography) → Scenic Spots, Geography, Italy, Football, Sports
Verona(Italy) → Scenic Spots, Geography, Italy, Football, Sports
Verona(Football) → Scenic Spots, Geography, Italy, Football, Sports
Verona(Sports) → Scenic Spots, Geography, Italy, Football, Sports

The word in brackets indicates the correct classification of each entry with the same name. ''→'' means ''be classified''.

All of the problems mentioned above will bring a large number of noise and error information to the large-scale open-domain KB, and finally degrade KB's precision.

The method proposed in this paper tries to solve the problems mentioned above, namely the noise challenge in Chinese KB. We label all the classification concepts in set T of an entry in the form of semantic tags (including classification, property, and ambiguous tags) onto each triple of this entry. In the following description, a tag in the set T of a triple is collectively referred to as t_i.

B. PROBLEM DEFINITION

In order to clearly define the problem of knowledge denoising in this paper, based on the view of set theory, we declare a set of formalization definitions as following:

Definition 1 (Knowledge triple): Knowledge triple is denoted by $< S, P, O >$, in which subject, predicate and object are denoted as S, as P, and O, respectively. We formulate a triple as $b_i =< b_{is}, b_{ip}, b_{io} >$ and $h_j =< h_{is}, h_{ip}, h_{io} >$ in BaiduBaike and Hudong respectively.

Definition 2 (Entry instance): an instance of entry e can be defined as a set of Infobox triples that appear on its web page. It is formulated as $e = \{b_1, b_2, \ldots, b_i, \ldots, b_n\}$ in BaiduBaike and $e = \{h_1, h_2, \ldots, h_i, \ldots, h_n\}$ in Hudong etc. Specifically, the literals of subject of each triple are equal to the literals of e.

Definition 3 (Set of classification tags of entry): The classification tags are the entry tags(class) appearing on its web page which declare the classification to which it belongs. This kind of tags forms a set of classification tags C of Entry.

Definition 4 (Set of property tags of entry): The property tags are the entry tags appearing on its web page which describe the property of entry, which form a property tag set Y. All tags in the set $C \cup Y$ are the original entry tags which are indiscriminately written together into a HTML label named "Entry Tag" by editors and presented on entry's web page. It should be noted that Hudong does not set property tag to describe entry.

Definition 5 (Ambiguous tags set of entry): If an entry e_i $(1 \leq i \leq n)$ has n other homonymous terms: $e_1, e_2, \ldots, e_i, \ldots, e_n$, then all these other homonymous terms will make up the ambiguous tags set $A_i = (C_1 \cup Y_1) \cup (C_2 \cup Y_2) \cup (C_i \cup Y_i) \cup \ldots \cup (C_n \cup Y_n)$. Therefore, for $\forall e_i \in \{e_1, \ldots, e_n\}$, $e_i := C_i \cup Y_i \cup A_i$, where ":=" means "labeled with all tags in set on the right side of this symbol". Then e_i has the "*is-a*" relation with each tag(class) in set $C_i \cup Y_i \cup A_i$.

Definition 6 (Triples set of a tag): An entry set of a tag (class) t consist of all the entries having the "*is-a*" relation with t, then the triples set of t consist of all of the triples belonging to this entry set. It is denoted as B_t and H_t in BaiduBaike and Hudong KBs respectively.

Definition 7 (Set of semantic tags of a triple): All the semantic tags labeled on a triple b_i is denoted as a set $T_{b_i} = C \cup Y \cup A = \{t_1, t_2, \ldots, t_k, \ldots, t_n\}$.

Definition 8 (Noise triple): Suppose an entry $e := C \cup Y \cup A$, all the triples belong to e will be labeled with all the tags of e, that is $\forall b_i \in e, b_i := C \cup Y \cup A$. If $Y \cap A \neq \Phi$, then e is identified as the noise entry instance in Y (not exist in Hudong) and/or A, and $\forall b_i \in e$, b_i is identified as the noise triple in Y and/or A.

It should be noted that all semantic tags(classes) appearing on the web page of entry do not necessarily all exist in the encyclopedia classification tree. In this paper, we do not consider the tags which do not exist in the encyclopedia classification tree.

IV. SYSTEM FRAMEWORK

Taking BaiduBaike as the target KB for denoising, and Hudong as the referenced KB, this paper plans to conduct initial similarity measurement of Chinese knowledge triple by using multi-strategy integrated similarity algorithm combining with cross-language universal Edit-Distance and Chinese TongYiCiCiLin similarity algorithm. Because of Chinese knowledge triples with the literal similarity between words and semantic similarity in the semantic web environment, it is one-sided and inaccurate to only adopt one method.

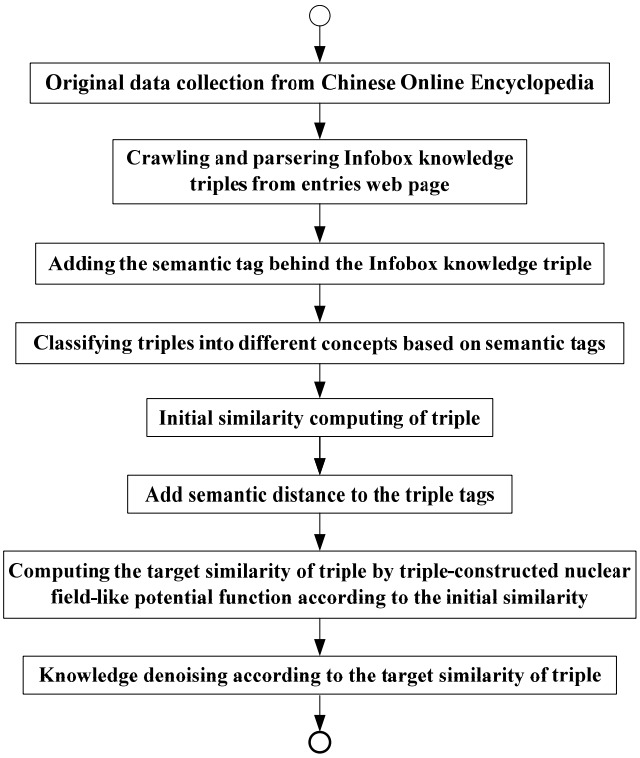


FIGURE 1. Overall architecture of system process.

Thus, after comparing two algorithm results, the larger value is selected for accumulation to the initial similarity result of the BaiduBaike triple. The initial similarity is planned to be defined as the triple proton quality in the potential function of encyclopedia knowledge triples. Based on this, an improved nuclear field-like potential function based on semantic tag is proposed for proton clustering computing for the purpose of maximally denoising and improving system's precision.

In general, the whole triple denoising process is systematically divided into the following two stages:

First Stage: Initial similarity computing of triple by using multi-strategy integrated similarity algorithm combining with cross-language Edit-Distance similarity algorithm and Chinese TongYiCiCiLin similarity algorithm.

Second Stage: Target similarity computing based on initial similarity value of triple by using the improved nuclear field-like potential function for the purpose of optimizing, revising the initial similarity value and finally denoising the Chinese encyclopedia KB.

The process of denoising method: "TriTag-NFPF" proposed in this paper is shown in Fig.1.

A. FIRST STAGE - INITIAL SIMILARITY COMPUTING FOR INFOBOX TRIPLES

The initial similarity computing in this paper is conducted between two triple sets corresponding to sub-classes of BaiduBaike and 11 top-classes of Hudong. For example, for initial similarity computing of BaiduBaike sub-class triple set: "Bank", the triple initial similarity computing together

TABLE 1. Denoising tasks of BaiduBaike.

Denoising tasks of sub-classes of BaiduBaike	Number of triples in sub-class of BaiduBaike	11 Top-classes of Hudong	Number of triples in top-classes of Hudong	Initial similarity computing tasks
Landform	143	Geography	455182	Landform→Geography
River	810			River→Geography
Lake	490			Lake→Geography
Bank	848	Economy	129312	Bank→Economy
Teacher	4768	People	512014	Teacher→People
Politics and law	5978	Society	734776	Politics and law→Society
Scenic spot	3498	Life	567438	Scenic spot→Life
Food	3591			Food→Life
Competition	93	Sports	44371	Competition→Sports
Calligraphy and paintings	1522	Culture	277476	Calligraphy and paintings→Culture
Prose	2130			Prose→Culture
Language	857			Language→Culture
Band	663	Art	164656	Band→Art
Architecture	3222			Architecture→Art
Collectible	426			Collectible→Art
Earthquake	975	Nature	1642294	Earthquake→Nature
Constellation	628			Constellation→Nature

with Hudong top-class ''Economy'' triple set is required because ''Bank'' is sub-class of ''Economy'', which is also the top-class in BaiduBaike. Therefore, mapping of all the triples in the ''Economy'' top-class in Hudong is required for initial similarity computing and accumulation. We randomly select triple sets of several sub-classes from BaiduBaike KB and the triple sets of the 11 top-class of Hudong encyclopedia as the denoising tasks of BaiduBaike. The precision will be measured for the denoising method proposed in this paper. As shown in Table 1.

Specifically, 1) Computing the Edit-Distance similarity of triple based on the Edit-Distance algorithm; 2) Computing the TongYiCiCiLin similarity of triple based on TongYiCiCiLin algorithm; and 3) Conducting complementary integration of the Edit-Distance and the TongYiCiCiLin similarity value to obtain the initial similarity value of triple.

1) SIM_E: EDIT-DISTANCE SIMILARITY COMPUTING OF TRIPLE

High efficient computing method with less resource demand needs to be considered in initial similarity computing. Thus, the Edit-Distance based initial similarity computing method is adopted. The literal similarity between the triples can be obtained through the Edit-Distance algorithm SIM_E, and the semantic correlation will be ignored.

Suppose any triple in the triples set B_t of a BaiduBaike sub-class t as $b_i =< b_{is}, b_{ip}, b_{io} >$, any triple in set H of Hudong top-class corresponding to set H_t as $h_j =< h_{js}, h_{jp}, h_{jo} >$, the Edit-Distance similarity between the subjects in the triples b_i and h_j can be given by Eq. (1). Here, the Edit-Distance similarity between predicate and objects can be obtained in a similar way.

$$\mathrm{SIM_E}\left(b_{is}, h_{js}\right) = \frac{1}{1 + \frac{|\mathrm{STEP}(b_{is}, h_{js})|}{\max(\mathrm{len}(b_{is}), \mathrm{len}(h_{js}))}} \tag{1}$$

where, $|\mathrm{STEP}(b_{is},h_{js})|$ is the required edit-operation times to make b_{is} and h_{js} equal. The length of characters b_{is} and h_{js} denoted as $\mathrm{len}(b_{is})$ and $\mathrm{len}(h_{js})$.

The Edit-Distance similarity of BaiduBaike knowledge triple can be obtained after computing.

2) SIM_T: TONGYICICILIN SIMILARITY COMPUTING OF TRIPLE

''TongYiCiCiLin'' edited by Mei *et al.* in 1983 is a Chinese synonym dictionary with the original aim at providing more synonymous expressions and assisting creation and translation work [52] . This dictionary contains not only the synonym of an entry but also a certain number of entries of the same kind, namely, related entries in a broad sense. In the Chinese synonym dictionary TongYiCiCiLin, each vocabulary after coding is organized in a tree structure in a hierarchical relation with five layers from top to bottom. There are code identifiers in corresponding levels respectively, which are arranged from left to right to constitute CiLin code of lexical unit. Each node in the tree represents a concept, and the connotative semantic correlation between entries will increase with the increase of layer. The Chinese

TABLE 2. Example of a TongYiCiCiLin code of "China".

Code Bit	1	2	3	4	5	6	7	8
Sub-Code	D	i	0	2	A	0	3	= or # or @
Meaning	Broad heading	Middle heading	Small heading		Word group	Atomic word group		Synonymous /unequal /isolated
Layer	1	2	3		4	5		

words co-reference relationship identification can actually be abstracted as the issue of identifying Chinese synonyms. We actually adopted the extended version in this invention, namely: HIT TongYiCiCiLin (extended)[3], as the dictionary lexicon for the TongYiCiCiLin similarity computing [26]. It is the largest dictionary of Chinese synonyms at present. It contains the largest number of Chinese synonyms.

Taking the lexical unit " 中国(China)" as an example (TongYiCiCiLin code: "Di02A03="), its TongYiCiCiLin code format as indicated in Table 2.

Firstly, TongYiCiCiLin code parsing of subject, predicate and object in the triple according to the structural characteristics of TongYiCiCiLin for sub-code extraction from the 1-th to the 5-th layer and comparison from the sub-code in the first layer. In case of different sub-codes, give the corresponding similarity weight of the mapping according to the appeared layer. The deeper layer of the sub-code appears, the higher the similarity weight. At the same time, the number of branch nodes at each layer also affects the similarity.

The semantic similarity between the triples can be obtained by the TongYiCiCiLin similarity algorithm: $\mathrm{SIM_T}$. In this case, $\mathrm{SIM_T}$ is adopted to explain with the similarity computing between the subjects in two triples as an example. The similarity computing between two predicates and objects can be obtained in a similar way:

$$\mathrm{SIM_T}\left(b_{is}, h_{js}\right) = \lambda \times \frac{L_n}{|L|} \times cos\left(N_T \times \frac{\pi}{180}\right) \times \left(\frac{N_T - D_c + 1}{N_T}\right) \quad (2)$$

To adjust the parameter semantic correlation factors, thus controlling the possible similarity degree of lexical units in branches at different layers, $\lambda \in (0,1)$. When λ is set to 0.9, our method can achieve the best overall performance. Because TongYiCiCiLin tree consists of 5 layers altogether, so $L = \{1, 2, 3, 4, 5\}$. $|L|$ is the number of elements in the set L and equals to 5 in this system. Setting $\forall L_n \in L$, L_n is the n-th layer's number where different sub-codes appearing between b_{is} and h_{js}. N_T is the total number of nodes of b_{is} and h_{js} on the branch of the n-th layer. D_c is the code-distance at b_{is} and h_{js}'s branch.

3) THE INITIAL SIMILARITY COMPUTING OF TRIPLE

Considering the semantic complementarity of $\mathrm{SIM_E}$ and $\mathrm{SIM_T}$ algorithms, complementary integration of the similarity results of these two algorithms in this paper is proposed: the maximum value is selected.

The initial similarity S of a triple b_i in triples set B_t of BaiduBaike proposed in this paper is shown in Eq. (3):

$$S_{b_i}(B_t) = \sum_{j=1}^{n} \begin{pmatrix} 0.3 \times \max\left(\mathrm{SIM_E}\left(b_{is}, h_{js}\right), \mathrm{SIM_T}\left(b_{is}, h_{js}\right)\right) \\ +0.5 \times \max\left(\mathrm{SIM_E}\left(b_{ip}, h_{jp}\right), \mathrm{SIM_T}\left(b_{ip}, h_{jp}\right)\right) \\ +0.2 \times \max\left(\mathrm{SIM_E}\left(b_{io}, h_{jo}\right), \mathrm{SIM_T}\left(b_{io}, h_{jo}\right)\right) \end{pmatrix} \quad (3)$$

where 0.3, 0.5 and 0.2 represent the weight coefficients of the subject similarity, predicate similarity and object similarity during the initial similarity computing of the whole triple b_i, which can be adjusted according to the target effect. Specifically, because we consider that the predicate of a triple can best reflect the semantic characteristics of the classification to which it should belong, followed by the subject and the object. Therefore, by setting the weight coefficients of the predicate to the highest among three of them, triples that have been correctly classified can get higher similarity values compared with others of improper classified in its triples set. Furthermore, after tuning parameters of the system according to the output, we find that when the weight coefficients of subject, predicate and object are 0.3, 0.5 and 0.2 respectively, the system can obtain the best P-value.

The initial similarity of knowledge triples in a BaiduBaike sub-class's triple set B can be obtained after computing.

B. ADD SEMANTIC DISTANCE TO ENTRY TAGS OF INFOBOX KNOWLEDGE TRIPLE

If the name of a tag of triple b_i is the same as the name of the current triples set (sub-class) to which the triple b_i belongs, then it is the center tag in the T_{b_i} and is denoted as t_i^c. For one triple, the semantic distance between tags is defined as the shortest path length between the center tag and other tags of it.

As mentioned above, all the entry tags involved in this paper belong to the open classification system of encyclopedia. We stipulate that if the current tag belongs to the current top-class set, then the current tag must be the parent-class or sub-class of the center tag. If the current tag is the direct parent-class or sub-class of the center tag, then the semantic distance is 1. Because the maximum depth of classification tree of BaiduBaike and Hudong are both equal to 2, so the maximum semantic distance between tags is 6. The semantic distance between the center tag t_i^c itself is 0 in its triples set.

[3]http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm

For example, a triple < Verona, Sports, Football > has been improper classified to the class: Scenic spots as it has an ambiguous tag: Scenic spots, meanwhile it also has another ambiguous tag: Geography, because both of them are used for describing the property of another ambiguous one of the entry: "Verona". In the classification tree of BaiduBaike, Scenic spots is sub-class of Tourism and Tourism is sub-class of Life. While both Geography and Life are the top-classes, so the semantic path between Scenic spots and Geography can be expressed as:

Scenic spots → Tourism → Life →*ROOT*→ Geography.

So the semantic distance between Scenic spots and Geography is 4.

C. SECOND STAGE - A NOVEL SEMANTIC TAG-BASED NUCLEAR FIELD-LIKE POTENTIAL FUNCTION FOR TARGET SIMILARITY COMPUTING OF INFOBOX KNOWLEDGE TRIPLE

A semantic tag-based "potential function" is proposed in this paper while computing the triple target similarity value. In order to further improve the precision of KB, the purpose of this stage is to optimize the similarity value obtained in the first stage.

It should be noted that the field which refers to the mapping of one vector to another or number in mathematics and a space region where each point is under the effect of force in physics, originally referred to magnetic field, electric field, gravitational field and other physical fields. In the above-mentioned physical fields, it is common to describe the mutual effect between particles with vector field majority function and the scalar potential function, which can also be defined in data fields just like in physics fields. The theory of data field [53] is proposed based on the field theory in physics. It is a description method of finalizing into field theory by abstracting the mutual relation between data in the space of field as the issue of mutual effect between material particles. The theory which expresses the interaction relationship between different data through the potential function can manifest the distribution characteristics of data, conduct clustering partition of data set according to the equipotential line structure in the data field.

The shortest path length (semantic distance) between center and other tags is $d = |t_i^c\text{-}t_j|$. We regard the triple as the data entity in our new nuclear field-like potential function, and the initial similarity value as the mass of triple in the field. Thus, the nuclear field-like potential function of the interaction between center tag t_i^c and the other tag t_j of the triple b_i is expressed as Eq. (4):

$$f_{t_j}(b_i) = S_{b_i}\left(B_{t_j}\right) \times e^{-\left(\frac{|d|}{\sigma}\right)^k} = S_{b_i}\left(B_{t_j}\right) \times e^{-\left(\frac{\left|t_i^c - t_j\right|}{10}\right)^2} \quad (4)$$

where $S_{b_i}(B_{t_j})$ represents the initial similarity of the triple b_i in its other triples set B of tag(sub-class): t_j, which represents the influence intensity of t_j on b_i. $\sigma \geq 0$ is called the influence factor, which determines the scope of influence of the elements. The potential function value increases with the increase of σ. In this paper, we set $\sigma = 10, k = 2$ in order to make the semantic distance have a greater impact on the target similarity value and which will make our method achieve the best overall performance.

Tags not belonging to the current top-class are punished for the purpose of weakening the initial similarity of triples containing tags with far semantic distance for secondary ranking based on the target similarity. The triples of lower-ranking behind will be removed from the KB. Therefore, the field majority computing based on the triple's tags is required to obtain the target similarity, which is the correction result of the initial similarity. The improved and optimized piecewise function (5) for punishing the improper classified triples is shown as follows:

$$\varphi_{t_j}(b_i) = \begin{cases} +f_{t_j}(b_i), & d \in (0, 3] \\ -f_{t_j}(b_i), & d \in (3, 6] \end{cases} \quad (5)$$

"+" Namely, tag t_j belongs to the top-class of the center tag t_i^c.

"−" Namely, tag t_j does not belong to the top-class of the center tag t_i^c.

In which, $\varphi_{t_j}(b_i)$ is the piecewise function we built. The "+" represents whether the current tag t_j and center tag t_i^c. belong to the same top-class. If yes, it indicates that the tag has a positive acting force on the triple b_i; if no, it indicates that the tag has an opposite acting force on the triple b_i.

The BaiduBaike triple final target similarity computing algorithm is shown as Eq. (6), where F_i represents the final target similarity value of triple b_i.

$$F_i = S_{b_i}\left(t_i^c\right) + \sum_{j=1}^{n} \varphi_{t_j}(b_i) \quad (6)$$

D. KNOWLEDGE DENOISING BASED ON THE TARGET SIMILARITY OF INFOBOX KNOWLEDGE TRIPLE

After the improved nuclear field-like potential function processing of the initial similarity, descending sorting according to the target similarity value is performed. After sorting the initial and target similarity sets of triples, according to the descending order of similarity value, it can be seen that this method has good effect on rank descending of incorrect triples and will have more conductive to KB denoising. For the triples' target similarity set of a sub-class, the latter 41% of triples will be deleted according to the idea of golden section point to obtain the final denoised and refined KB.

For clarity, the "TriTag-NFPF" algorithm consisting of two stages can also be presented with the manner of*"end-to-end"* as Algorithm 1:

V. ANALYSIS AND COMPARISON OF EXPERIMENT RESULTS

To verify the effectiveness of the new method of constructing Chinese KB proposed in this paper, we have extracted data sets from Chinese open encyclopedias such as BaiduBaike and Hudong, and designed two experiments as follows:

Algorithm 1 TriTag-NFPF(B, H, T)

Input: a triples set B_x of the tag x of BaiduBaike for denosing,
tag t_k's corresponding triples set H_{t_k} of Hudong Top-class,
tags set T_{b_i} of each triple b_i in B_x
Output: triple set *Map_B'* after denoised

1. for each triple b_i in B_x
2. for each tag t_k in T_{b_i}
3. for each triple h_j in H_{t_k}
4. $S_{b_i}(B_{t_k}) \leftarrow S_{b_i}(B_{t_k}) + 0.3 \times \max(SIM_E(b_{is}, h_{js}), SIM_T(b_{is}, h_{js})) + 0.5 \times \max(SIM_E(b_{ip}, h_{jp}), SIM_T(b_{ip}, h_{jp})) + 0.2 \times \max(SIM_E(b_{io}, h_{jo}), SIM_T(b_{io}, h_{jo}))$
5. end for
6. end for

// confirm the center tag *tc i* of triple b_i in its tag set

7. for each t_k in T_{b_i}

// "*literalof*" means get the literal value of input strings

8. if *literalof*(t_k) == *literalof* (B_x) then
9. $t_i^c \leftarrow t_k$ // tag t_k is confirmed as the center tag t_i^c of b_i
10. end if
11. end for

// get the target similarity value of each triple in set B_x

12. $F_i \leftarrow S_{b_i}(t_i^c)$
13. for each $t_k \in T_{b_i}$
14. $d \leftarrow |t_i^c - t_k|$
15. if $0 < d \leq 3$ then
16. $F_i \leftarrow F_i + S_{b_i}(B_{t_k}) \times e^{-\left(\frac{d}{10}\right)^2}$
17. end if
18. else then
19. $F_i \leftarrow F_i - S_{b_i}(B_{t_k}) \times e^{-\left(\frac{d}{10}\right)^2}$
20. end else
21. *Map_B'.put*(b_i, F_i)
22. end for
23. end for

// the latter 41% of triples will be deleted

24. descendsortingbytargetsimilarity(*Map_B'*, 0.41)
25. return *Map_B'*

i) Compared the changes in the number of incorrectly classified triples in the latter 41% of each sub-class triples set of BaiduBaike after the first stage and the second stage processing.

The purpose of the experiment is mainly to observe the changes in number of incorrectly classified triples in the latter 41% of BaiduBaike sub-class triples set after the clustering sorting by the improved nuclear field-like potential function, and to compare the number of incorrect triples in the latter 41% with descending sorting of initial similarity set without this processing, thus reasonably analyzing and comparing the results.

TABLE 3. Evaluation statistical table.

11 Top-class	Sub-classes	Evaluation
Geography	landform	*P*
	river	*P*
	lake	*P*
Economy	bank	*P*
People	teacher	*P*
Society	politics and law	*P*
Life	scenic spot	*P*
	food	*P*
Sports	competition	*P*
Culture	calligraphy and paintings	*P*
	prose	*P*
	language	*P*
Art	band	*P*
	architecture	*P*
	collectible	*P*
Nature	earthquake	*P*
	constellation	*P*

ii) Precision can be obtained after comparing with the processing at two stages and state-of-the-art methods

With Chinese online encyclopedia KB as the experimental data source in this paper, the crawler toolkit HTMLParser is used for crawling and parsing of the Infobox structured information contained in the open classification page and entry page of BaiduBaike and Hudong, respectively.

The strategy for building KBs of BaiduBaike and Hudong is based on the methods reported in Ref. [11], [36]. Then, the information is organized in the form of Chinese triples to establish the initial large-scale Chinese open domain KB. Then, the crawled data can be divided into 11 top-class concept sets, each of which contains subclass concepts and each sub-class concept includes several corresponding triples.

The number of entry triples in the 11 top-classes of Hudong, sub-classes of BaiduBaike, and corresponding relationships can be seen in Table 1. We invite four Chinese senior grade students from Capital University of Economics and Business to use the manual identification and manual annotation-based method for semantic annotating. This is the manually annotated result with the principle of whether the triple matches the related sub-class attributes. If it matches, *Y* is marked, otherwise *N* is marked.

Evaluation of experimental data: the precision of the selected triples:

P = number of outputted Y / total number of outputs $*100\%$

The experiment in this paper is aimed at denoising, namely, the precision of triple sets(sub-classes) is required and triples marked as *Y* need to be extracted. This system effect can be completely satisfied by observing the precision.

After selecting some triple sets under the 11 top-class in the classification tree of Chinese online encyclopedia, it can be found that precision (*P*) can better reflect the efficiency brought by the algorithm.

TABLE 4. Information of a part of BaiduBaike KB preparing to be denoised.

11 Top-classes	Sub-class	The number of triples	The number of correctly classified triples	The number of inappropriately classified triples
Geography	landform	143	120	23
	river	810	587	223
	lake	490	399	91
Economy	bank	848	622	226
People	teacher	4769	1773	2996
Society	politics and law	5977	5537	440
Life	scenic spot	3498	3137	361
	food	3591	3112	479
Sports	competition	93	81	12
Culture	calligraphy and paintings	2130	1568	562
	prose	1522	383	1139
	language	857	136	721
Art	band	663	396	267
	architecture	3222	2385	837
	collectible	426	405	21
Nature	earthquake	475	316	159
	constellation	755	638	117

Thus, the evaluation is shown in Table 3. The detailed a part of BaiduBaike KB selected randomly and preparing to be denoised in this paper is shown in Table 4.

A. EXPERIMENT 1 – RESULT ANALYSIS FOR THE SEMANTIC TAGS-BASED NUCLEAR FIELD-LIKE POTENTIAL FUNCTION

The following is a result analysis of a triple belongs to the sub-class of "scenic spots": < Verona, sports, football >. Since the term Verona is both a historic city and a football club in northern Italy, therefore, while categorizing, all the Infobox triples belong to the entry will be inappropriately classified because of ambiguity. Therefore, we can see that triples belong to the term "Verona" related to the "football club" will be inappropriately classified under the sub-class "scenic spots".

In this paper, the schematic process of modifying the initial similarity value in order to denoise the Infobox triples is presented as shown in the Fig.2.

Where, the number above the arrow represents the semantic distance d between every two tags. The "Scenic spot" is the classification of the triple currently discussed, then the center tag of the triple "< Verona, Sports, Football >" is "Scenic spot", the d of center tag is 0, and the initial similarity value of this triple is +3175.2356 in the triple set 'Scenic spot".

According to the algorithm 1 proposed in this paper, the initial similarity value will be optimized based on the other semantic tags it has. "Italy" is a sub-class of the top-class "Geography", and the semantic path between "Scenic spot" and "Italy" can be expressed as:

"Scenic spots"→"Tourism"→"Life"→*ROOT*→"Geography"→"Country"→"Italy".

So the semantic distance between "Scenic spot" and "Italy" is 6, while the semantic distance between "Scenic spot" and the top-class "Geography" is 4.

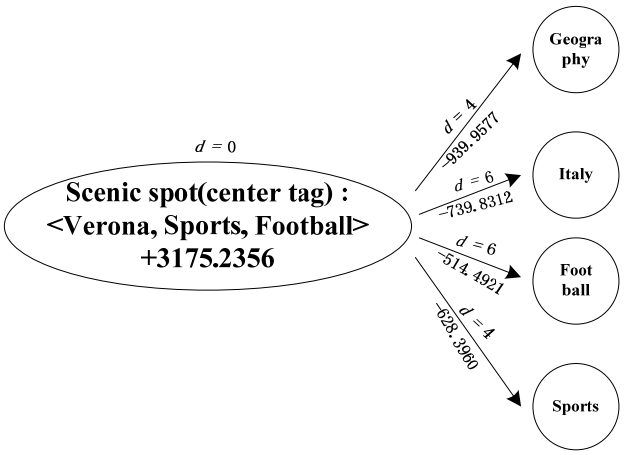


FIGURE 2. The schematic process of modifying the initial similarity value.

"Football" is a sub-class of top-class "Sports", and "Football" is also a sub-class of "Ball Games". Therefore, the semantic path between "Scenic spots" and "Football" can be expressed as:

"Scenic spots"→"Tourism"→"Life"→*ROOT*→"Sports"→"Ball games"→"Football".

So the semantic distance between "Scenic spot" and "Football" is 6, and the distance between "Scenic spots" and "Sports" is 4.

"+/−" represents whether the current tag and the center tag are in the same top-class. If it is, it means that the current tag exerts positive force on the current triple, then the calculation result of piecewise function (5) is positive; if not, it means that the current tag exerts negative force on the current triple, then the calculation result of piecewise function (5) is negative.

The detailed computing process and intermediate results are as follows:

TABLE 5. Changes in the number of improper classified triples in the latter 41% of sub-classes triple sets of BaiduBaike in every stage.

11 Top-class	BaiduBaike Sub-classes	First Stage (Before optimized)	Second Stage (After optimized)	Difference value
Geography	landform	15	17	+2
	river	112	128	+16
	lake	45	54	+9
Economy	bank	123	115	-8
People	teacher	1288	1300	+12
Society	politics and law	6	353	+347
Life	scenic spot	159	214	+55
	food	163	271	+108
Sports	competition	2	12	+10
Culture	calligraphy and paintings	435	517	+82
	prose	256	312	+56
	language	300	338	+38
Art	band	126	143	+17
	architecture	306	301	-5
	collectible	7	9	+2
Nature	earthquake	90	98	+8
	constellation	72	84	+12
Total		3505	4266	+761

$f_{Geography}(<\text{Verona, Sports, Football}>)$
$= S_{<\text{Verona,Sports,Football}>}(Geography) \times e^{-0.16}$
$= 1103.0554448085834 \times e^{-0.16}$
$=939.9577$
$f_{Italy}(<\text{Verona, Sports, Football}>)$
$= S_{<\text{Verona,Sports,Football}>}(Italy) \times e^{-0.36}$
$= 1060.4162781416999 \times e^{-0.36}$
$=739.8312$
$f_{Footbal}(<\text{Verona, Sports, Football}>)$
$= S_{<\text{Verona,Sports,Football}>}(Footbal) \times e^{-0.36}$
$= 737.4328285519572 \times e^{-0.36}$
$=514.4921$
$f_{Sports}(<\text{Verona, Sports, Football}>)$
$= S_{<\text{Verona,Sports,Football}>}(Sports) \times e^{-0.16}$
$= 737.4328285519572 \times e^{-0.16} = 628.3960$
From Eq. (5) and (6), we can get that:
$F_{<\text{Verona,Sports,Football}>} =$
$S_{<\text{Verona,Sports,Football}>}(\text{Scenic spot})$
$+\varphi_{Geography}(<\text{Verona, Sports, Football}>)$
$+\varphi_{Italy}(<\text{Verona, Sports, Football}>)$
$+\varphi_{Footbal}(<\text{Verona, Sports, Football}>)$
$+\varphi_{Sports}(<\text{Verona, Sports, Football}>)$
$= +3175.2356\text{-}939.9577\text{-}739.8312\text{-}514.4921\text{-}628.3960$
$= +352.5586$

In summary, we get the target similarity value of +352.5586, which shows that the similarity value of the triple decreases a lot after optimized by the new potential function. In order to delete the noise triple from KB, we finally sort the triple set of the sub-class ''Scenic spot'' in descending order according to the target similarity value. Compared with the descending order according to the initial similarity value, it can be seen that the ranking of this improper classification triple has dropped from 2258 to 3374. Results show that the method proposed in this paper has a good effect on dropping the ranking of improper classified triples, and will effectively improve the accuracy of KB.

B. EXPERIMENT 2 – CHANGES IN THE NUMBER OF IMPROPER CLASSIFIED TRIPLES IN THE LATTER 41% OF SUB-CLASSES TRIPLE SETS OF BAIDUBAIKE IN EVERY STAGE

After the optimized processing by potential function based on distance between triple's tags, according to the similarity value of the two stages, each set of triples is sorted in descending order for every stage respectively. And then the changes in the number of improper classified triples in the latter 41% of the respective sub-class triple sets are compared respectively.

As shown in Table 5, the number of improper classified triples in the latter 41% of most BaiduBaike sub-class' triple set has increased after the modified processing by the semantic tags-based potential function proposed in this paper. And we also demonstrate the experimental results in the schematic diagrams indicated in Fig.3 and Fig.4. The most significant is the sub-class ''politics and law'', with the number of improper classified triples increased by 347. Meanwhile, the number of improper classified triples in the sub-class' triple sets ''bank'' and ''architecture'' decreased and the reduced degree is relatively small. It could be negligible considering that the total number of improper classified triples has increased by 761. The Result indicates that the denoising method proposed in this paper will improve the precision of Chinese online encyclopedia KB construction.

C. EXPERIMENT 3 – RESULT ANALYSIS FOR PRECISION OF ALL DENOISING TASK

Precision is an important criterion to evaluate information retrieval effect. This experiment is divided into two stages.

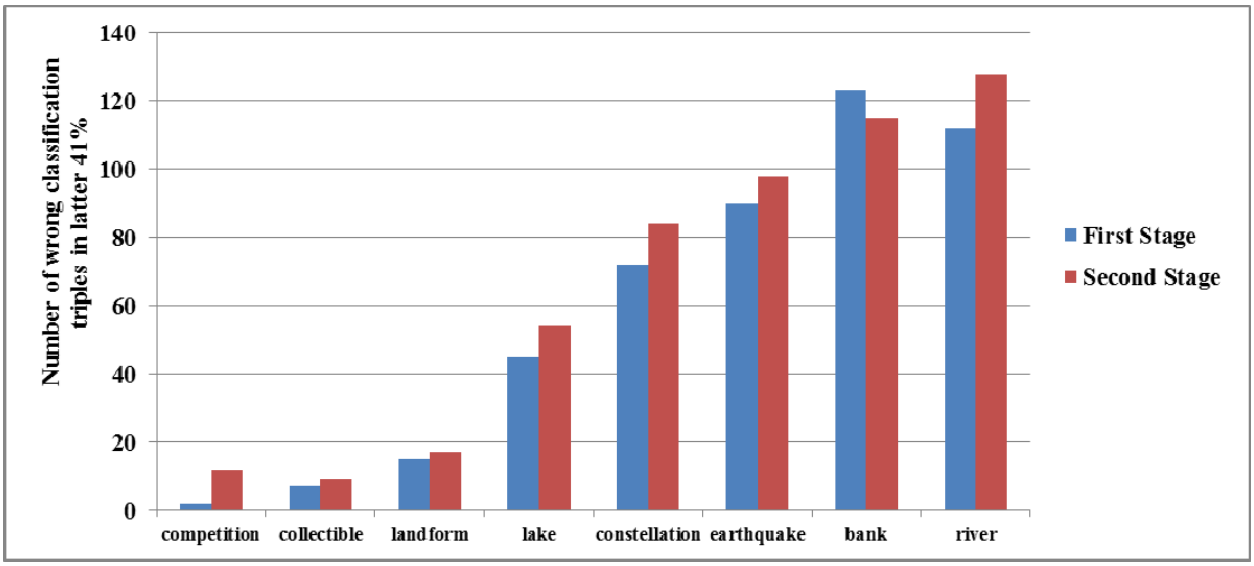


FIGURE 3. The number of improper classification triples in each sub-class triple set at every stage (histogram) – part I.

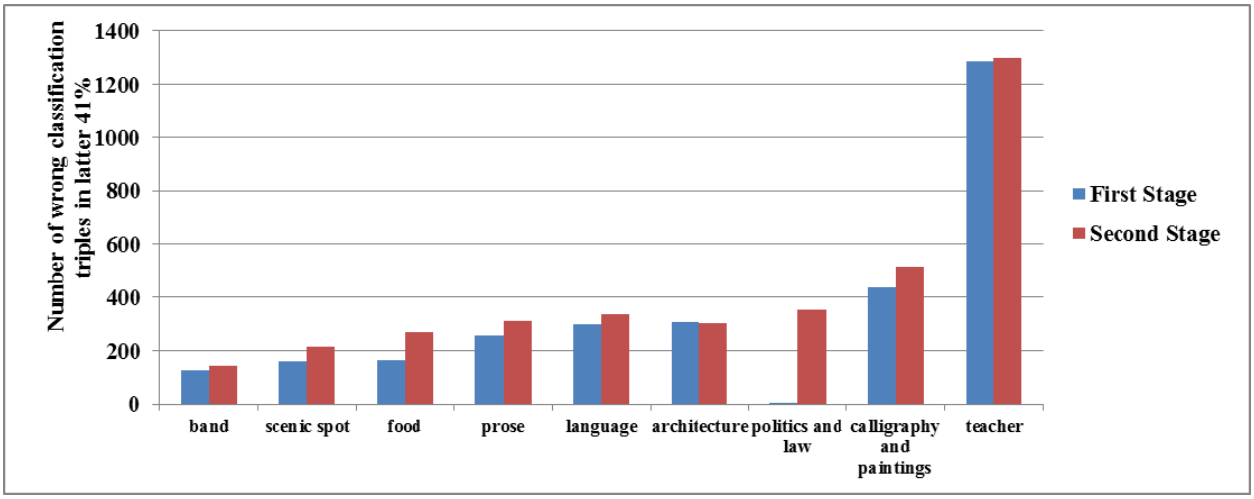


FIGURE 4. The number of improper classification triples in each sub-class triple set at every stage (histogram) – part II.

Stage one: sorting in descending order in triples sets for denosing according to the initial similarity value of triples in BaiduBaike. Stage two: sorting in descending order in triples sets for denosing according to target similarity value after the modified processing by potential function. The latter 41% of the triples of sub-classes are deleted by our method firstly, and then the *P*-values of the remaining triples can be calculated for following comparison with other state-of-the-art methods.

Table 6 is made for listing the *P*-values of each sub-class triple set at two stages proposed in this paper. The *P*-values at the first and second stage are compared with state-of-the-art Chinese encyclopedia KBs construction methods: Ref. [11], [14] and [36]. There is comparability between our method and Ref. [14] because knowledge triples are both refined in the second stage of each system but different algorithms are used. Specifically, the latter 41% of triples in each sub-classes triple set are also deleted according to the descending order of triples' TF-IDF values.

It can be seen that The *P*-values obtained by TF-IDF algorithm [14] is better than that by method [36]. This is because most triples which can reflect the Infobox characteristics of their concept achieve a high ranking, so as to easily eliminate the noise triples which rank relatively lower. So it can be proved that as long as we use appropriate intelligent algorithms to refine knowledge, we will get KBs with high quality and precision. But the *P*-values of our method is higher than TF-IDF algorithm [14] , it is mainly because that TF-IDF algorithm can only compute the frequency of appearing of

TABLE 6. Comparisons of *P*-values with state-of-the-art methods.

Hudong 11 Top-classes	BaiduBaike Sub-classes	*P*-values				
		Ref. [36]	Ref.[11]	Ref. [14]	Our First Stage	Our Second Stage
Geography	landform	0.8392	0.8210	0.9010	0.9052	0.9289
	river	0.6580	0.6370	0.6478	0.6547	0.6882
	lake	0.8143	0.8350	0.8088	0.8409	0.8720
Economy	bank	0.7335	0.7067	0.7563	0.7941	0.7781
People	teacher	0.3882	0.3481	0.4110	0.4209	0.4252
Society	politics and law	0.9264	0.9200	0.9380	0.8769	0.9753
Life	scenic spot	0.8968	0.9100	0.9245	0.9021	0.9288
	food	0.8666	0.8666	0.8874	0.8509	0.9018
Sports	competition	0.8710	0.8709	0.8496	0.8178	1.0000
Culture	calligraphy and paintings	0.2516	0.2532	0.2342	0.2160	0.3073
	prose	0.7362	0.6314	0.7444	0.7565	0.8011
	language	0.1587	0.1274	0.2074	0.1674	0.2425
Art	band	0.5973	0.6237	0.6459	0.6395	0.6830
	architecture	0.7402	0.7120	0.7581	0.7207	0.7180
	collectible	0.9507	0.8599	0.9513	0.9443	0.9523
Nature	earthquake	0.6653	0.7022	0.7023	0.7538	0.7823
	constellation	0.8450	0.8045	0.8355	0.8990	0.9259
Average value		0.7023	0.6883	0.7179	0.7153	0.7595

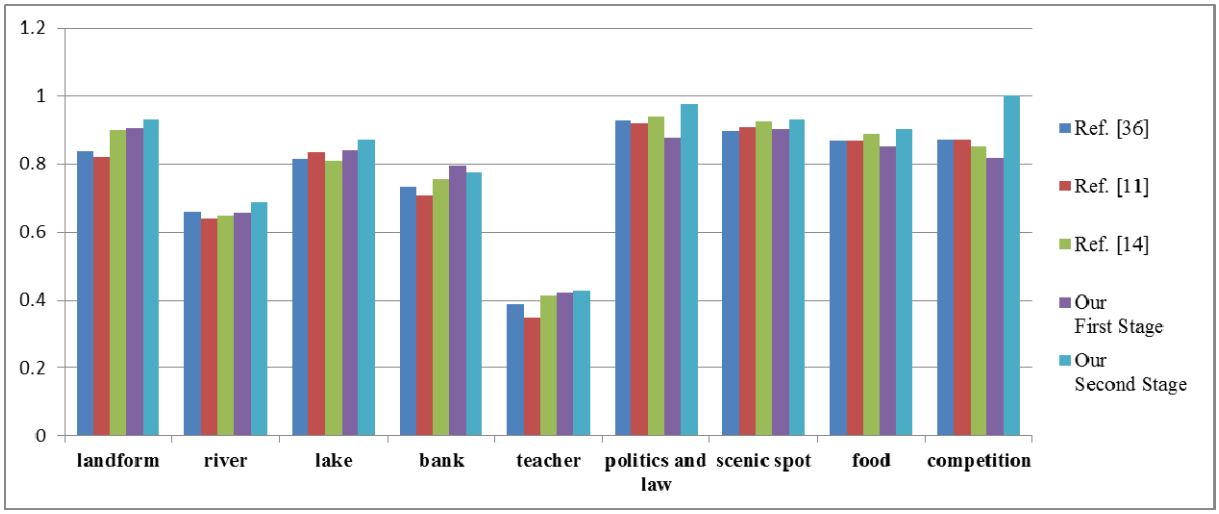


FIGURE 5. The *P*-values of each sub-class triple set at every stage (histogram) – part I.

Infobox mechanically and ignore its semantic features completely. Therefore, some randomness will be involved into the results of TF-IDF algorithm. By comparison, our method not only introduces the TongYiCiCiLin (extended version) as the semantic dictionary but also introduces the tags of Infobox triples as their semantic feature, so our enhanced nuclear field-like potential function will have more rationality and stability in the re-ranking and re-clustering process of triples.

It is notable from the data that *P*-values after denoising processing is significant increased than that when not processed at all. Only the sub-class ''architecture'' has declined slightly, which is insignificant for the increase of overall *P*-values improvement. Because in open-domain scenarios, it is difficult to guarantee that the same algorithm can cover and take into account the characteristics of triples under any different sub-classes. Therefore, negative may occur in few individual set of triples.

Overall, the average of *P*-values of our method is higher than other state-of-the-art methods at the second stage. Thus, it can be concluded that the new method proposed in this paper can improve the precision of encyclopedia KB.

Next, we also demonstrate the experimental results in the schematic diagrams indicated in Fig.5 and Fig.6 that show the comparing *P*-values obtained at the end of two

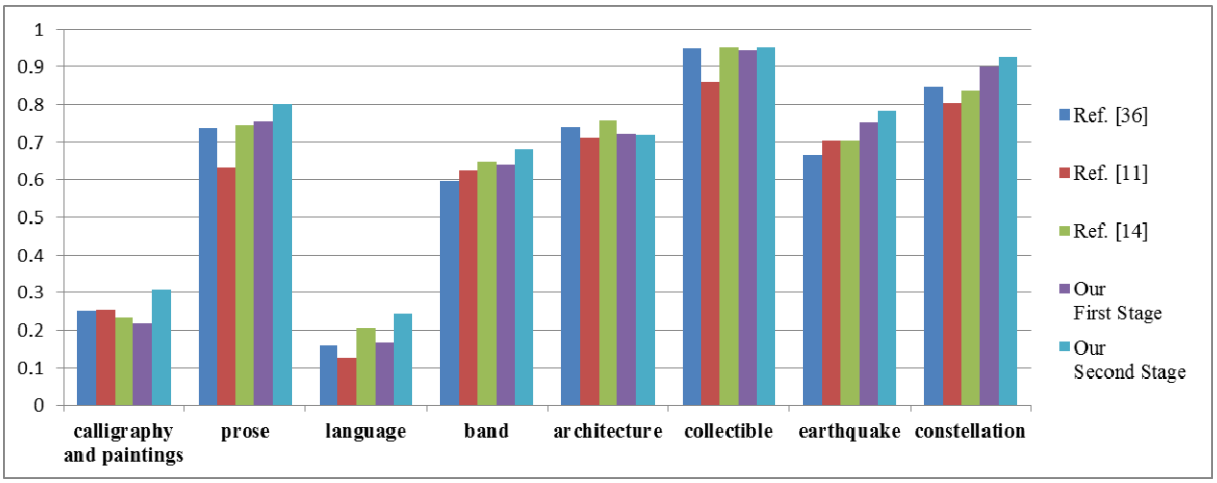


FIGURE 6. The *P*-values of each sub-class triple set at every stage (histogram) – part II.

denoising stages by using the new method proposed in this paper.

VI. CONCLUSION

This paper firstly to measure the initial similarity of knowledge triple by means of the algorithm integrating Edit-Distance with TongYiCiCiLin, then design a new nuclear field-like potential function for obtaining the target similarity to further re-cluster and re-rank the triples in sets. In the end, after descending sorting of triples set based on the target similarity, and the lower-ranked triples are removed from the KB, the goal of knowledge refinement is achieved.

The experimental results reveal that the "TriTag-NFPF" algorithm proposed in this paper not only can optimize and corrects the initial similarity value of triples, but also effectively avoids the issues of ambiguity problem in the Chinese online encyclopedia and improper classification of knowledge triples, thus eventually improving the precision of the Chinese encyclopedia KB.

In the future, with the explosive growth of knowledge on the Internet, we will explore the use of Big Data and parallel technology for more efficient knowledge denoising, so that our system can be provided with scalability.

REFERENCES

[1] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic Web," *Sci. Amer.*, vol. 284, no. 5, pp. 28–37, 2001.

[2] F. Wu and D. S. Weld, "Autonomously semantifying wikipedia," in *Proc. 16th ACM Conf. Inf. Knowl. Manage. ACM*, Nov. 2007, pp. 41–50.

[3] F. Wu and D. S. Weld, "Automatically refining the wikipedia infobox ontology," in *Proc. 17th Int. Conf. World Wide Web*, Apr. 2008, pp. 635–644.

[4] 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*, Jun. 2008, pp. 1247–1250.

[5] 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.

[6] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum, "YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia," *Artif. Intell.*, vol. 194, pp. 28–61, Jan. 2013.

[7] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "DBpedia—A crystallization point for the Web of Data," *J. Web Semantics*, vol. 7, no. 3, pp. 154–165, Sep. 2009.

[8] 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.

[9] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data: The story so far," *Semantic Services, Interoperability Web Appl., Emerg. Concepts*, pp. 205–227, 2011.

[10] Z. Wang, Z. Wang, J. Li, and J. Z. Pan, "Building a large scale knowledge base from Chinese Wiki encyclopedia," in *Proc. Joint Int. Semantic Technol. Conf.* Springer, 2011, pp. 80–95.

[11] X. Wang, L. Jiang, H. Shi, Z. Feng, and P. Du, "Jingwei+: A distributed large-scale RDF data server," in *Proc. Asia–Pacific Web Conf.* Springer, 2012, pp. 779–783.

[12] Y. Fu, X. Wang, Z. Feng, and X. Lv, "Organization and integration of chinese encyclopedia knowledge based on semantic web," (in Chinese) *Comput. Eng. Appl.*, vol. 51, no. 14, pp. 120–126, 2015.

[13] T. Wang, J. Song, R. Di, and Y. Liang, "A thesaurus and online encyclopedia merging method for large scale domain-ontology automatic construction," in *Proc. Int. Conf. Knowl. Sci., Eng. Manage.* Springer, 2013, pp. 132–146.

[14] T. Wang, H. Gu, Z. Wu, and J. Gao, "Multi-source knowledge integration based on machine learning algorithms for domain ontology," in *Neural Computing and Applications*. Springer, 2018, pp. 1–11. doi: 10.1007/s00521-018-3806-5.

[15] J. Li, C. Wang, X. He, R. Zhang, and M. Gao, "User generated content oriented chinese taxonomy construction," in *Proc. Asia–Pacific Web Conf.* Springer, 2015, pp. 623–634.

[16] Y. Chen, L. Chen, and K. Xu, "Learning chinese entity attributes from online encyclopedia," in *Proc. Asia–Pacific Web Conf.* Springer, 2012, pp. 179–186.

[17] Q. Liu, B. Liu, M. He, D. Wu, Y. Liu, and X. Cheng, "Synonymous expansion based entity attribute extraction via online encyclopedia," (in Chinese), *J. Chin. Inf. Process.*, vol. 30, no. 1, pp. 16–24, 2016.

[18] T. Wang, F. Ji, and T. Xu, "A novel knowledge extraction approach orientedon unstructured information of Chinese online encyclopedia," (in Chinese), *Library Inf. Service*, vol. 60, no. 13, pp. 126–133, 2016.

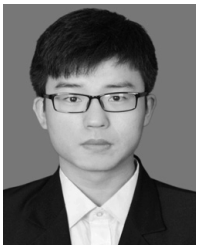
[19] W. W. Cohen, P. Ravikumar, and S. E. Fienberg, "A comparison of string metrics for matching names and records," in *Proc. KDD Workshop Data Cleaning Object Consolidation*, vol. 3, Aug. 2003, pp. 73–78.

[20] F. Giunchiglia and M. Yatskevich, "Element level semantic matching," Univ. Trento, Trento, Italy, Tech. Rep., 2004.

[21] M. M. Stark and R. F. Riesenfeld, "Wordnet: An electronic lexical database," in *Proc. 11th Eurograph. Workshop Rendering*, vol. 37, 1998.

[22] A. Isaac, L. van der Meij, S. Schlobach, and S. Wang, "An empirical study of instance-based ontology matching," in *The Semantic Web*. Springer, 2007, pp. 253–266.
[23] A. Nikolov, V. Uren, E. Motta, and A. Roeck, "Integration of semantically annotated data by the knofuss architecture," in *Proc. 16th Int. Conf. Knowl. Eng., Pract. Patterns*. Springer, 2008, pp. 265–274.
[24] Q. Zhong, H. Li, J. Li, G. Xie, J. Tang, L. Zhou, and Y. Pan, "A gauss function based approach for unbalanced ontology matching," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Jul. 2009, pp. 669–680.
[25] P. Jain, P. Hitzler, A. P. Sheth, K. Verma, and P. Z. Yeh, "Ontology alignment for linked open data," in *Proc. Int. Semantic Web Conf.*. Springer, 2010, pp. 402–417.
[26] T. Wang, T. Xu, Z. Tang, and Y. Todo, "TongSACOM: A TongYiCiCiLin and sequence alignment-based ontology mapping model for Chinese linked open data," *IEICE Trans. Inf. Syst.*, vol. 100, no. 6, pp. 1251–1261, Jun. 2017.
[27] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov, "Silk—A link discovery framework for the Web of data," in *Proc. LDOW*, vol. 538, Apr. 2009, p. 53.
[28] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov, "Discovering and maintaining links on the Web of data," in *Proc. Int. Semantic Web Conf.* Springer, 2009, pp. 650–665.
[29] F. Scharffe, Y. Liu, and C. Zhou, "Rdf-ai: An architecture for rdf datasets matching, fusion and interlink," in *Proc. IJCAI Workshop Identity, Reference, Knowl. Represent.*, Pasadena, CA, USA, Jul. 2009.
[30] O. Hassanzadeh, L. Lim, A. Kementsietsidis, and M. Wang, "A declarative framework for semantic link discovery over relational data," in *Proc. 18th Int. Conf. World Wide Web*, Apr. 2009, pp. 1101–1102.
[31] C. Bizer and A. Seaborne, "D2RQ–treating non-RDF databases as virtual RDF graphs," in *Proc. 3rd Int. Semantic Web Conf.*, 2004, pp. 1–2.
[32] A.-C. N. Ngomo and S. Auer, "LIMES—A time-efficient approach for large-scale link discovery on the Web of data," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 2312–2317.
[33] M. Pershina, M. Yakout, and K. Chakrabarti, "Holistic entity matching across knowledge graphs," in *Proc. IEEE Int. Conf. Big Data*, Oct./Nov. 2015, pp. 1585–1590.
[34] X. Niu, X. Sun, H. Wang, S. Rong, G. Qi, and Y. Yu, "Zhishi.me - Weaving Chinese Linking Open Data," in *Proc. Int. Semantic Web Conf.* Springer, 2011, pp. 205–220.
[35] X. Niu, S. Rong, H. Wang, and Y. Yu, "An effective rule miner for instance matching in a Web of data," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage.*, Nov. 2012, pp. 1085–1094.
[36] Z.-C. Wang, Z.-G. Wang, J.-Z. Li, and J. Z. Pan, "Knowledge extraction from Chinese Wiki encyclopedias," *J. Zhejiang Univ. Sci. C*, vol. 13, no. 4, pp. 268–280, 2012.
[37] X.-P. Wang, K. Liu, S.-Z. He, S.-L. Liu, Y.-Z. Zhang, and J. Zhao, "Multi-source knowledge bases entity alignment by leveraging semantic tags," (in Chinese), *Chin. J. Comput.*, vol. 40, no. 3, pp. 701–711, 2017.
[38] Z. Wang, J. Li, Z. Wang, and J. Tang, "Cross-lingual knowledge linking across Wiki knowledge bases," in *Proc. 21st Int. Conf. World Wide Web*, Apr. 2012, pp. 459–468.
[39] Z. Wang, J. Li, and J. Tang, "Boosting cross-lingual knowledge linking via concept annotation," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, Aug. 2013, pp. 2733–2739.
[40] Y.-C. Wang, C.-K. Wu, and R. T.-H. Tsai, "Cross-language article linking with different knowledge bases using bilingual topic model and translation features," *Knowl. Based Syst.*, vol. 111, pp. 228–236, Nov. 2016.
[41] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proc. 26th Int. Conf. Adv. Neural Inf. Process. Syst.*, Dec. 2013, pp. 2787–2795.
[42] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *Proc. 28th AAAI Conf. Artif. Intell.*, Jun. 2014, pp. 1112–1119.
[43] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Proc. 29th AAAI Conf. Artif. Intell.*, Feb. 2015, pp. 2181–2187.
[44] G. Ji, S. He, L. Xu, K. Liu, and J. Zhao, "Knowledge graph embedding via dynamic mapping matrix," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, 7th Int. Joint Conf. Natural Lang. Process.*, vol. 1, 2015, pp. 687–696.
[45] R. Xie, Z. Liu, J. Jia, H. Luan, and M. Sun, "Representation learning of knowledge graphs with entity descriptions," in *Proc. 30th AAAI Conf. Artif. Intell.*, Mar. 2016, pp. 2659–2665.
[46] Z. Wang and J.-Z. Li, "Text-enhanced representation learning for knowledge graph," in *Proc. IJCAI*, Jul. 2016, pp. 1293–1299.
[47] B. D. Trisedya, J. Qi, and R. Zhang, "Entity alignment between knowledge graphs using attribute embeddings," in *Proc. 33rd AAAI Conf. Artif. Intell.*, Jul. 2019, pp. 297–304.
[48] A. Bordes, J. Weston, R. Collobert, and Y. Bengio, "Learning structured embeddings of knowledge bases," in *Proc. 35th AAAI Conf. Artif. Intell.*, Aug. 2011, pp. 301–306.
[49] R. Socher, D. Chen, C. D. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledge base completion," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 926–934.
[50] M. Nickel, L. Rosasco, and T. Poggio, "Holographic embeddings of knowledge graphs," in *Proc. 30th AAAI Conf. Artif. Intell.*, Feb. 2016, pp. 1955–1961.
[51] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Sov. Phys. Doklady*, vol. 10, no. 8, pp. 707–710, Feb. 1966.
[52] J.-J. Mei, *Tongyici Cilin*. Shanghai, China: Shanghai Lexicographical Publishing House, 1983.
[53] D. Li and Y. Du, *Artificial Intelligence with Uncertainty*. Beijing, China: National Defence Industry Press, 2005.

TING WANG received the B.S. and Ph.D. degrees in computer science and technology from the Beijing University of Technology, in 2008 and 2014, respectively. In 2012, he was a Research Student with the Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is currently an Associate Professor with the School of Management and Engineering, Capital University of Economics and Business. His research interests include semantic web, knowledge discovery, and big data. He is a member of the Chinese Information Processing Society of China, CIPSC.



HANZHE GU received the B.S. degree in e-commerce from the Shandong University of Science and Technology, in 2017. He is currently pursuing the master's degree in management science and engineering with the School of Management and Engineering, Capital University of Economics and Business. His research interests include knowledge discovery and e-commerce.

JIE LI is currently pursuing the bachelor's degree in computer science and technology from the Capital University of Economics and Business. His research interests include big data and semantic web. In 2017, he was awarded the Excellent Cadre Award and the Social Work Scholarship. He also received the National Encouragement Scholarship, the Second Prize for Excellence, the Three Good Students Award for Excellence, and the Research Scholarship, in 2018.

JINGYAO XIE received the B.S. and M.D. degrees in computer science and technology from the Beijing University of Aeronautics and Astronautics, in 2011 and 2014, respectively. She is currently an Engineer with State Grid Beijing Electric Power Company. Her research interests include data mining and virtualization technology.