Quality Evaluation of Triples in Knowledge Graph by Incorporating Internal With External Consistency

Taiyu Ban, Xiangyu Wang, Lyuzhou Chen, Xingyu Wu, Qiuju Chen, and Huanhuan Chen, *Senior Member, IEEE*

***Abstract*— The evaluation of knowledge quality (KQ) in multisource knowledge graphs (KGs) is an essential step for many applications, such as fragmented knowledge fusion and knowledge base construction. Many existing quality evaluation methods for multisource knowledge are based on validation from high-quality knowledge bases or statistical analysis of knowledge related to a specific fact from multiple sources, named external consistency (EC)-based methods. However, high-quality KGs are difficult to obtain, and there might exist incorrect knowledge in multisource KGs interfering with KQ evaluation. To address the issue, this article refers to the internal structure of a KG to evaluate the degree to which the contained triples conform to the overall semantic pattern of the KG, such as KG embedding and logic inference-based approaches, defined as internal consistency (IC) evaluation. The IC is integrated with the EC to identify possible incorrect triples and reduce their influences on the KQ evaluation, thus alleviating the interference of incorrect knowledge. The proposed method is verified with multiple datasets, and the results demonstrate that the proposed method could significantly reduce wrong evaluations caused by incorrect knowledge and effectively improve the quality evaluation of triples.**

***Index Terms*— Consistency, knowledge graph (KG), knowledge quality (KQ), quality control.**

I. Introduction

THE evaluation of knowledge[1] quality (KQ) plays an important role in quality control[2] of knowledge graphs (KGs) [1], especially for the construction of multisource knowledge bases, fragmented knowledge fusion [2], and KG-based applications, such as dialogue systems [3], [4] and search engine assistant [5]. Due to the wide range of knowledge sources and the diversification of knowledge acquisition methods, there may be incorrect or contradictory knowledge/triples in KGs from different sources [6], [7], which seriously affects the formation of large-scaled, high-quality KGs. Therefore, evaluating the quality of knowledge/triples in multisource KGs is of great significance for selecting high-quality knowledge and filtering out low-quality knowledge [8], as shown in Fig. 1.

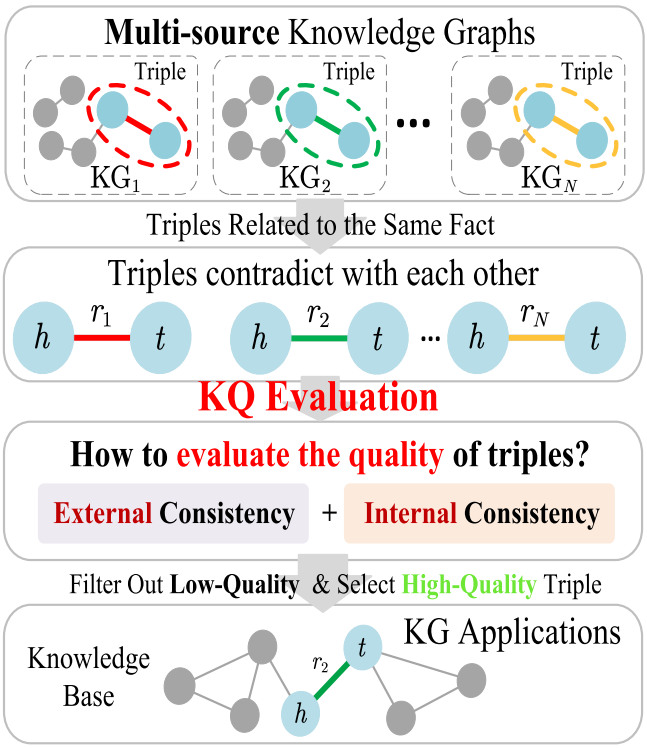


Fig. 1. Description of the process selecting high-quality triples from multisource KGs for KG applications. The quality evaluation of triples in multisource KGs is an essential step for further applications of KGs such as fragmented knowledge fusion and knowledge base construction.

An important KQ evaluation method in multiple KGs is based on external consistency (EC) of knowledge [9], [10]. Some work refers to this method as the external-based method [11], which is referred to as the EC-based method in this article. EC is whether external knowledge agrees with the investigated knowledge under certain criteria. For example, in the scenario of triple validation from external knowledge

Manuscript received 16 January 2022; revised 12 April 2022 and 6 June 2022; accepted 20 June 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2021ZD0111700, in part by the National Nature Science Foundation of China under Grant 62137002 and Grant 62176245, in part by the Key Research and Development Program of Anhui Province under Grant 202104a05020011, in part by the Key Science and Technology Special Project of Anhui Province under Grant 202103a07020002, in part by the Fundamental Research Funds for the Central Universities, and in part by the Special Foundation for Science and Technology Innovation and Entrepreneurship of CCTEG under Grant 2020-2-TD-CXY006. *(Corresponding author: Huanhuan Chen.)*

Taiyu Ban, Xiangyu Wang, Lyuzhou Chen, Xingyu Wu, and Huanhuan Chen are with the School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, China (e-mail: hchen@ustc.edu.cn).

Qiuju Chen is with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3186033.

Digital Object Identifier 10.1109/TNNLS.2022.3186033

[1]The knowledge here refers to the triples contained in KGs.

[2]KG quality control refers to the use of certain methods to ensure that the KG is of high quality or to meet application requirements.

bases, the EC of the investigated triple is true if its entities and relation refer to the same concept as those of a triple in the external base. Likely, using high-quality external KGs to evaluate KQ is a common EC-based method. However, due to the wide range of KG applications, high-quality KGs are often difficult to obtain in many domains [12], [13].

Another EC-based method compares the investigated knowledge with an amount of external knowledge[3] to evaluate its quality [11]. This kind of method usually assumes that most of the external knowledge is correct. However, due to the fact that most of KGs are built with natural language processing (NLP)-based methods [14], [15], error knowledge is inevitable. If most of the triples related to a specific fact are incorrect, it could lead to wrong quality evaluation of the investigated knowledge. To alleviate the interference caused by incorrect knowledge, some methods depend on additional information, such as domain expert knowledge or text validation from trustworthy sources [16], [17]. However, the additional information is usually limited to specialized domains, which may not be available in all domains. Hence, it is difficult to be applied to evaluate KQ in general KGs.

Internal consistency (IC) is defined as the extent to which triples conform to the overall semantic pattern of the KG under particular knowledge representation, such as logic expressions and embedded vectors [18], [19]. For example, under given logic rules, the consistency of entities and relations in each triple could be verified and the conflicted ones are with low consistency since they do not conform to the logic patterns presented by other triples. This article focuses on the embedding way to represent knowledge, which embeds triples by minimizing certain distances among the vectors of entities and relations based on the topology structure of the KG [20]. The IC of a triple is evaluated through the distance among the vectors of its entities and relation. This IC of triples is a kind of self-verification of their KG, where the triples not conforming to the semantic pattern of the whole KG are commonly caused by errors during KG construction and could be regarded as low-quality knowledge [21]. Moreover, the evaluation of IC does not rely on additional information or external knowledge [11]. Therefore, it could be generally utilized to detect low-quality knowledge for KQ evaluation in multisource KGs. However, since it is independent on external knowledge, there lacks validation from the trustworthy knowledge related to the same fact. In this case, what is stated by the consistent knowledge[4] could be incorrect [1]. Therefore, the IC evaluation might not be reliable enough to indicate the correctness of knowledge, which needs to be integrated with EC properly.

To address the above problems, this article proposes a KQ evaluation method that integrates the IC and EC. IC evaluation is employed to provide prior information for the quality of triples, which is further integrated with EC for KQ evaluation. Through introducing IC evaluation into an EC-based process, the possible incorrect knowledge is detected through the value of its IC metric. The influence of this knowledge on the quality estimation of other knowledge is adjusted lower to alleviate the interference incorrect knowledge during the EC-based process. On the other side, the statement of the investigated knowledge is validated through the EC comparison to promise the stability of IC utilization, where the wrong statements of high IC knowledge could also be detected.

To properly evaluate KQ, an iterative method is proposed to update the quality of the triples in all KGs simultaneously. The external triples with low IC are limited to have less influence on the quality estimation of the investigated triple in an iteration, thus reducing the interference caused by incorrect external knowledge. For the IC evaluation, a transformation method is proposed to transform the IC evaluation results in different KGs into a uniform framework to produce a proper IC metric, which is further utilized to adjust the weight of triples in the EC comparison. The framework of the proposed method is presented in Fig. 2.

This article is the first work to provide a triple quality evaluation framework incorporating internal with EC of knowledge. It has close relevance with neural network (NN) and learning system (LS) communities. KG quality management plays an important role in the lifecycle of KG applications [21], and this article could assist KG quality management by assessing the KG construction methods in practical scenarios, especially for NN-based methods or LS. For most NN-based methods or LS, model evaluation metrics are usually available during the training or testing process. However, in the practical application, such as knowledge extraction and knowledge fusion, the quality of their results is difficult to be guaranteed or accurately estimated due to the lack of labeled data. In this case, the method proposed in this article could provide a promising evaluation method for the practical performance of NN-based methods or LS in applications due to its generality. The contributions of this article are summarized as follows.

1) A new perspective for evaluating the quality of triples is proposed, which integrates the IC of KGs and EC with external knowledge to evaluate the quality of the triples.
2) An iterative method based on EC comparison is proposed, which eliminates the interference of incorrect external knowledge by utilizing the IC to adjust the influence of knowledge on the KQ evaluation.
3) An IC transformation method is proposed, which transforms the evaluation of different KGs to a uniform framework for proper utilization of IC in the KQ evaluation.

The rest of this article is organized as follows. After introducing the related work of this article in Section II, the proposed method is illustrated in Section III. Experimental results and discussions are presented in Sections IV and V, respectively. Finally, Section VI concludes this article.

II. Related Work

To evaluate KQ in multiple sources, EC-based methods are usually employed. These methods usually utilize high-quality external knowledge or a large amount of external knowledge

[3]The external knowledge here corresponds to triples related to the same fact as that of the investigated knowledge in other KGs. Note that many methods do not specify the investigated or external knowledge, which usually treats all the knowledge equally and evaluates KQ through the consistency comparison.

[4]Here, the consistent knowledge refers to the knowledge evaluated as high IC.

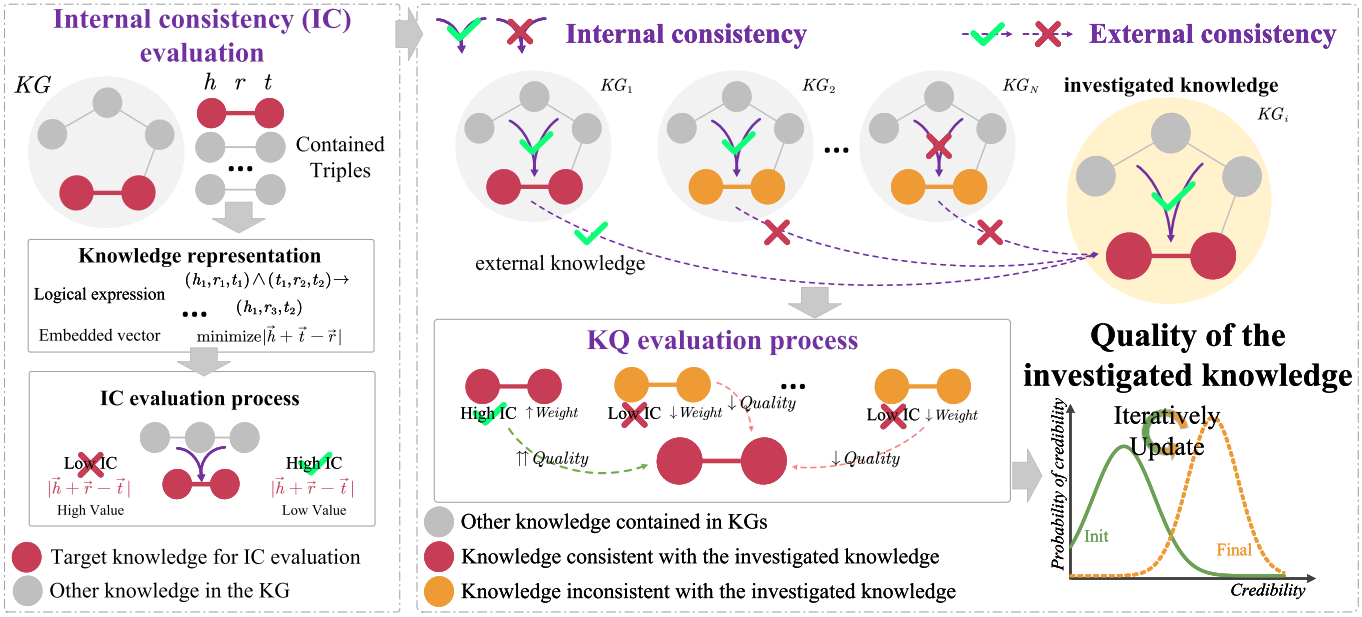


Fig. 2. Framework of the proposed method. The box on the left is IC evaluation to give each triple an IC metric, which is based on embeddings in this article. The IC metric of triples is subsequently integrated with an iterative KQ update method based on EC, shown as the box on the right. The subbox "KQ evaluation process" shows the update procedure where the effect of external triples on quality estimation of the investigated triple is adjusted according to their IC metrics.

describing the same facts to validate the knowledge in KGs or in other forms. For example, Gerber *et al.* [22] presented a framework named DeFacto, which finds trustworthy sources on the web for knowledge in KG. This framework combines the evidence from web pages and calculates the score of quality for input triples. Yin *et al.* [23] invented TruthFinder, an algorithm evaluating KQ using the relationships between web sites. The algorithm iteratively updates the quality of knowledge and knowledge sources based on the comparison between information. Similar ideas are implemented in [24]–[26]. In addition, for the verification of multisource knowledge, majorty vote (MV) could be used as an effective method to evaluate KQ, which takes the most occurring knowledge as the correct one. Many MV-based methods are studied to improve the prediction effect by assigning certain weights for knowledge from each source [27]–[29]. In addition, statistical analysis methods are also used in the evaluation of KQ. This type of method usually estimates the credibility of knowledge by using statistical analysis on different descriptions of the same knowledge. Expectation–maximization (EM) algorithm is a representative method. It employs EM to iteratively update and estimate the probability that knowledge is correct based on the statistical distribution of different descriptions of a fact [30].

The IC of knowledge is the degree to which it is consistent with the whole KG [1], [7], [31], which can also be used as evidence for evaluating KQ. It is widely utilized to detect incorrect knowledge in a KG (also known as error detection), only based on the provided KG itself with no additional information or external knowledge. The methods of evaluating IC are concluded as internal methods in [11], which includes methods based on logic rules, knowledge embedding, statistical distribution, graph features, and so on.

For logic-based methods, lots of reasoners based on logic rules have been applied for large ontologies to check their knowledge, such as consequence-based (CB) reasoner and fast classification of terminologies (FaCT++) [32]. For knowledge embedding methods, such models, such as TransE and RotateE, have been employed for error detection tasks of KGs [33], [34]. Additional methods based on statistical distribution or graph features have also been researched to evaluate KQ and improve KG quality accordingly. Paulheim and Bizer [35] used a statistical method to study consistency. This method uses the statistical distribution of the entity attributes and types in knowledge to identify possible wrong relationships. Melo and Paulheim [36] considered type features of entities in knowledge and path features of graphs. Such features are combined into a relation classifier to find inconsistencies. Wienand and Paulheim [37] studied the application of numerical consistency of knowledge in the KG and employed the outlier detection method to find low-quality knowledge.

III. Methods

This section elaborates the proposed method that integrates EC with IC to evaluate triple quality in multisource KGs. Section III-A illustrates the problem definition, including the used symbols and the goal of the task. Section III-B presents the form of KQ and the overall iterative method to update KQ. Section III-C introduces the detailed implementation of the proposed iterative method integrating EC with IC. Section III-D proposes an IC evaluation method providing uniform IC metrics of triples in different KGs. The scheme of the proposed method is shown in Fig. 3, and the logic diagram of containing main formulas is presented in Fig. 4.

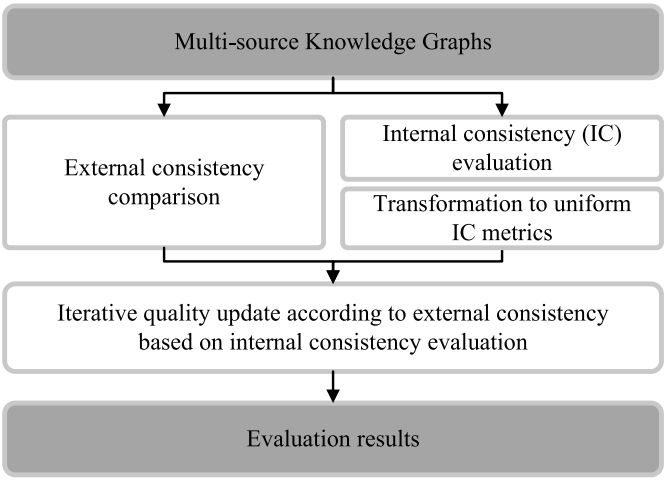


Fig. 3. Scheme of the proposed method.

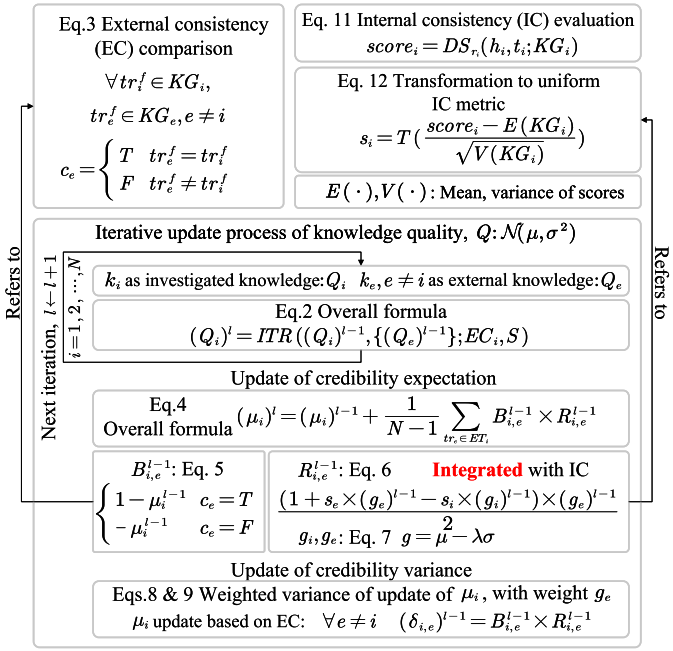


Fig. 4. Logic diagram of main formulas of the proposed method.

A. Problem Definition

Suppose that there are N KGs totally, expressed as $KG = \{\mathrm{KG}_i | i \in \{1, 2, \ldots, N\}\}$. Each KG KG_i contains a set of triples, denoted as $\mathrm{KG}_i = \{\mathrm{tr}_i^1, \mathrm{tr}_i^2, \ldots, \mathrm{tr}_i^n\}$, where tr_i^j represents the triple describing the fact j in KG KG_i. EC comparison evaluates the quality of the investigated triple tr_i^j by comparing it with triples tr_e^j describing the same fact j in other KGs. For concise illustration, the following discussion omits the superscript j representing a certain fact, i.e., tr_i^j and tr_e^j are expressed as tr_i and tr_e, respectively. Denote the investigated triple as tr_i, and all the triples describing the same fact in other KGs are named external triples, expressed as $\mathrm{ET}_i = \{\mathrm{tr}_e | e \neq i\}$. The quality of the investigated triple tr_i and its external triples tr_e could both be represented as a distribution of credibility Q. Given the set of KGs, the goal of this article is to evaluate the quality of triple in KG_i with a distribution of credibility Q. The main parameters involved in the proposed method are summarized in Table I.

TABLE I
MAIN INCLUDED SYMBOLS OF THE PROPOSED METHOD

Symbols	Definitions
KG	Multi-source knowledge graphs
N	Total number of multi-source KGs
KG_i	The i^{th} knowledge graph in KG
tr_i/tr_i^j	The triple describes the fact j in KG_i (j omitted)
ET_i	Set of external triples of tr_i
tr_e/tr_e^j	The external triple of tr_i related to fact j in KG_e (j omitted)
EC_i	The external consistency result of triple tr_i
$Q_i/(Q_i)^l$	The quality of triple tr_i (in the l^{th} iteration)
$Q_e/(Q_e)^l$	The quality of triple tr_e (in the l^{th} iteration)
S	The IC evaluation results of triples $\{tr_i\} \cup ET_i$
s_i	The internal consistency of tr_i
μ_i/μ_i^l	The expectation of Q_i (in the l^{th} iteration)
${\sigma_i}^2/({\sigma_i}^2)^l$	The variance of Q_i (in the l^{th} iteration)
c_e	The external consistency between tr_e and tr_i
$B_{i,e}^{l-1}$	The update upper bound of μ_i^{l-1} for comparison with tr_e.
$R_{i,e}^{l-1}$	The update range of μ_i^{l-1} for comparison with tr_e
$(\delta_{i,e})^{l-1}$	The update amount of μ_i^{l-1} for comparison with tr_e
$(\overline{\delta_i})^{l-1}$	The weighted variance of $(\delta_{i,e})^{l-1}$
$g_i/(g_i)^l$	The numerical indicator of Q_i (in the l^{th} iteration)
$g_e/(g_e)^l$	The numerical indicator of Q_e (in the l^{th} iteration)
$score_i$	The output score of KG embedding model of tr_i.
$E(KG_i)$	The score mean of triples of KG_i by KG embedding model
$V(KG_i)$	The score variance of triples of KG_i by KG embedding model
$T(x)$	The integral of standard normal distribution from $-\infty$ to x

B. Distribution-Based KQ Evaluation

For unverified knowledge, credibility is an important evaluation dimension, which refers to the extent to which unverified knowledge conforms to the correct facts [38], [39]. Therefore, the credibility of knowledge is utilized to model its quality. In the EC-based method, the KQ evaluation may be affected by the unknowns of the quality of external KGs. For example, if a triple's statement of a fact is consistent with that of many triples in other KGs, it may be considered as high quality. However, since the quality of these external triples is unknown, there could exist many incorrect triples, possibly leading to a wrong quality evaluation of the investigated triple.

The unknowns of the KQ may interfere with the KQ evaluation results; moreover, they could also lead to the uncertainty of the estimated quality. To model the KQ rigorously, a normal distribution is utilized to represent KQ. The quality of triple tr_i is expressed as

$$Q_i : \mathrm{cr}_i \sim \mathcal{N}(\mu_i, {\sigma_i}^2) \tag{1}$$

where cr_i represents the credibility of triple tr_i, $\mathcal{N}(\mu_i, {\sigma_i}^2)$ is the normal distribution, $\mu_i (\mu_i \in (0, 1))$ is the expectation, and σ_i^2 is the variance of cr_i.

In fact, other distributions besides the normal distribution could also be utilized to model the KQ in specific scenarios, such as χ^2 distribution and beta distribution. Nevertheless, for

unknown scenarios,[5] the normal distribution is the best choice to approximately estimate the KQ. It could model the KQ properly even if the KQ distribution may not originally obey the normal distribution [40]. Moreover, the normal distribution is of the maximum entropy in all distributions that have the same mean and variance [41], and this characteristic makes the normal distribution a natural choice to model general scenarios.

To evaluate Q_i, this article designs an iterative update method that employs EC and IC for a more proper estimation. The update of the quality of triple tr_i could be expressed as

$$(Q_i)^l = \mathrm{ITR}((Q_i)^{l-1}, \{(Q_e)^{l-1}\}; \mathrm{EC}_i, S) \tag{2}$$

where l represents the lth iteration $\mathrm{ITR}(\cdot)$ denotes the proposed update process in each iteration, Q_e is the estimated quality of the external triples of k_i, EC_i and S represent the EC comparison result and the IC of triples $\{\mathrm{tr}_i\} \cup \mathrm{ET}_i$, respectively. EC_i is obtained by comparing the consistency of the investigated triple with its external triples, denoted as $\mathrm{EC}_i = \{c_e | \mathrm{tr}_e \in \mathrm{ET}_i\}$. c_e is defined as

$$c_e = \begin{cases} \text{True}, & \mathrm{tr}_e = \mathrm{tr}_i \\ \text{False}, & \mathrm{tr}_e \neq \mathrm{tr}_i \end{cases}, \quad \mathrm{tr}_e \in \mathrm{ET}_i. \tag{3}$$

In this way, the EC of the investigated triple is obtained. S represents the IC of triples, which is expressed as $S = \{s_1, s_2, \ldots, s_N\}$. s_i is the IC evaluation result of triple tr_i.

As for the utilization of (2), each triple of the ones related to the same fact is separately chosen as the investigated triple, and the other triples are chosen as external triples to process this equation in an iteration. In other words, the quality of all triples is updated in an iteration based on (2). Note that the quality update of all triples related to a fact could be implemented simultaneously in an iteration since the update process of a triple only depends on the estimated KQ in the previous iteration and constant EC comparison and IC metrics.

C. KQ Update Based on the Integration of IC and EC

A method based on incorporating internal with EC is proposed to iteratively update the distribution of triple quality. In each iteration, the quality of the investigated triple is updated by comparison with triples describing the same fact in external KGs. The quality of all triples is updated synchronously in each iteration.

The quality of each triple is in the form of normal distribution, as shown in (1), which consists of two essential parameters, the expectation μ_i and variance ${\sigma_i}^2$. The expectation μ_i of the quality of triple tr_i is updated as follows:

$$\begin{aligned}(\mu_i)^l &= (\mu_i)^{l-1} + \frac{1}{N-1} \sum_{\mathrm{tr}_e \in \mathrm{ET}_i} (\delta_{i,e})^{l-1} \\ &= (\mu_i)^{l-1} + \frac{1}{N-1} \sum_{\mathrm{tr}_e \in \mathrm{ET}_i} B_{i,e}^{l-1} \times R_{i,e}^{l-1}\end{aligned} \tag{4}$$

where l represents the lth iteration, $(\mu_i)^l$ is the expectation of credibility of tr_i after the lth iteration, and $(\delta_{i,e})^{l-1}$ denotes the update amount of μ_i in the comparison with external triple tr_e. $B_{i,e}^{l-1}$ and $R_{i,e}^{l-1}$ represent the upper bound and the update range of μ_i, respectively. $B_{i,e}^{l-1}$ is expressed as follows:

$$B_{i,e}^{l-1} = \begin{cases} 1 - \mu_i^{l-1}, & c_e = \text{True} \\ -\mu_i^{l-1}, & c_e = \text{False} \end{cases} \tag{5}$$

where c_e is the comparison result between the investigated triple tr_i and the external triple tr_e, as shown in (3). If tr_e is consistent with tr_i, $B_{i,e}^{l-1}$ is positive, and the credibility expectation of tr_i increases. On the contrary, μ_i decreases if tr_e is inconsistent with tr_i.

The update range $R_{i,e}^{l-1}$ depends on the quality of both investigated triple tr_i and external triple tr_e, which is expressed as

$$R_{i,e}^{l-1} = \frac{1}{2}(1 + s_e \times (g_e)^{l-1} - s_i \times (g_i)^{l-1}) \times (g_e)^{l-1} \tag{6}$$

where s_i and s_e represent the IC of triple tr_i and tr_e[6] and g_i and g_e are the indicators representing the estimated quality of tr_i and tr_e. Taking g_i as an example, g_i is constructed using μ_i and σ_i of the quality Q_i of triple tr_i, which is expressed as

$$g_i = \mu_i - \lambda \sigma_i \tag{7}$$

where $\lambda \in (0, \infty)$ is a hyperparameter. g_i reflects the lower bound of the credibility of tr_i under a certain probability, and the probability is positively correlated to λ. The design of g is inspired by the "empirical rule" of normal distribution.

The estimated quality g_e is utilized as a weight of the corresponding external triple tr_e during the EC comparison. The external triple with higher estimated quality is considered to be more convincing, thus having more influence on the update of the quality of the investigated triple, i.e., higher weight during the EC comparison. The above situation is summarized in (6), where $R_{i,e}^{l-1}$ is positively correlated with $(g_e)^{l-1}$.

In (6), the IC s_e is also considered, which is utilized as a correction of the weight that tr_e occupies during the EC comparison. Triples with good IC are generally more convincing than those with bad IC. Therefore, factor $s_e \times g_e$ is utilized as a corrected weight of tr_e depending on its IC. Furthermore, $s_e \times g_e - s_i \times g_i$ is utilized to be positively correlated with $R_{i,e}^{l-1}$, which represents the relative weight of the external triple. Note that the correction of IC is only used in the relative weight, whereas it is not used in the second g_e. This design aims to promise the proper quality evaluation when IC is inaccurate.

For the variance σ_i^2 of the credibility distribution Q_i, σ_i^2 represents the statistical dispersion of the credibility of triples. In an iteration, the comparison between each external triple and the investigated triple tr_i could be regarded as a sample for the evaluation of the credibility of tr_i. Therefore, the dispersion of the influences on the credibility of tr_i during the comparison

[5]The unknown scenarios refer to scenarios with unknown conditions such as the trustworthiness of sources of KGs or the accuracy of knowledge extraction methods. These conditions influence the overall quality of knowledge.

[6]The construction for s is illustrated in Section III-D.

could reflect the dispersion of cr_i. Based on the comparison in the lth iteration, the variance $(\sigma_i^2)^l$ is presented as

$$(\sigma_i{}^2)^l = \frac{\sum_{\mathrm{tr}_e \in \mathrm{ET}_i} (g_e)^{l-1} \times ((\delta_{i,e})^{l-1} - (\overline{\delta_i})^{l-1})^2}{\sum_{\mathrm{tr}_e \in \mathrm{ET}_i} (g_e)^{l-1}} \tag{8}$$

where $(\delta_{i,e})^{l-1}$ is the update amount of the credibility expectation μ_i during the comparison with tr_e in the lth iteration. $(\overline{\delta_i})^{l-1}$ is the weighted mean of the $\delta_{i,e}^{l-1}$, denoted as

$$(\overline{\delta_i})^{l-1} = \frac{\sum_{\mathrm{tr}_e \in \mathrm{ET}_i} (g_e)^{l-1} \times (\delta_{i,e})^{l-1}}{\sum_{\mathrm{tr}_e \in \mathrm{ET}_i} (g_e)^{l-1}} \tag{9}$$

where $(g_e)^{l-1}$, the indicator representing the estimated quality of tr_e, is utilized as the weight of the corresponding update amount $(\delta_{i,e})^{l-1}$. The weighted variance corresponds to that the dispersion of the influences caused by high-quality external triples is considered more important for the dispersion evaluation of the estimated quality of tr_i. Note that the weight in (8) and (9) is similar to the weight of tr_e during the update of μ_i in (6), which both assumes that external triples with higher estimated quality occupy higher weight.

The proposed method evaluates the quality of triples based on the comparison with external triples, which incorporates the IC. As the number of iterations increases, the quality of triples is prone to agree with the EC of triples. To balance the EC and IC, a restriction term is added to the update of μ_i

$$(\mu_i)^l = (\mu_i)^{l-1} + l^{-2+(\sigma_i)^{l-1}} \times \frac{1}{N-1} \sum_{\mathrm{tr}_e \in \mathrm{ET}_i} B_{i,e}^{l-1} \times R_{i,e}^{l-1} \tag{10}$$

where $l^{-2+(\sigma_i)^{l-1}}$ is the restriction term, which degrades the update amount of μ_i as the number of iteration increases. When the variance of the credibility of the investigated triple is low, it represents that the estimated credibility expectation $(\mu_i)^{l-1}$ is of good confidence, thus degrading the update amount of μ_i to a greater extent. On the contrary, if the variance is high, it shows that the estimated credibility expectation is not convincing enough and the update amount of μ_i is not restricted as much as the former case.

D. Triple Credibility Evaluation Based on IC

For the IC s_i of tr_i in (6), it evaluates the extent to which other knowledge in KG supports the investigated knowledge, which reflects an aspect of knowledge credibility and could be used as a basis for assessing credibility [1]. Using the embedding of triples to evaluate the IC of knowledge is an important IC evaluation method [11]. This method embeds knowledge into a continuous vector space and uses a scoring function $\mathrm{DS}(\cdot)$ to measure consistency while retaining the inherent structure of KG. $\mathrm{DS}(\cdot)$ provides a score_i according to whether the knowledge is consistent with KG, which could be expressed as

$$\mathrm{score}_i = \mathrm{DS}_{r_i}(h_i, t_i; \mathrm{KG}_i) \tag{11}$$

where h_i, r_i, and t_i are the head entity, relation, and tail entity of the triple tr_i, respectively. KG_i is the KG containing tr_i. The larger score_i is, tr_i is evaluated as better IC. The KG embedding methods support many KG quality control tasks based on IC, such as error detection and completion of KGs [20], [42]–[44], which select high-quality triples to complete KGs or filter out low-quality triples to enhance the quality of KGs.

For KQ evaluation in multisource KGs, the entities, relations, and link modes of different KGs may be various, which may lead to quite different evaluation results (i.e., the score of triples) when an embedding model is applied to different KGs to evaluate IC of their containing triples. To solve the above problems, an IC transformation method is designed. This method eliminates the possible bias and normalizes the fluctuations of the scores of different KGs to construct a unified consistency evaluation system.

For the investigated knowledge, the proposed method uses the ranking of its score in scores of all the knowledge in KG_i to evaluate the IC s_i. The evaluation formula is expressed as

$$s_i = T\left(\frac{\mathrm{score}_i - E(\mathrm{KG}_i)}{\sqrt{V(\mathrm{KG}_i)}}\right) \tag{12}$$

where $E(\mathrm{KG}_i)$ is the average of the scores of all knowledge in KG_i, $V(\mathrm{KG}_i)$ is the variance of these scores, and $T(x)$ calculates the proportion of knowledge whose score is lower than the investigated knowledge. $E(\mathrm{KG}_i)$, $V(\mathrm{KG}_i)$, and $T(x)$ are expressed as follows:

$$E(\mathrm{KG}_i) = \sum_{\mathrm{tr}_j \in \mathrm{KG}_i} \mathrm{score}_j / |\mathrm{KG}_i| \tag{13}$$

$$V(\mathrm{KG}_i) = \sum_{\mathrm{tr}_j \in \mathrm{KG}_i} (\mathrm{score}_j - E(\mathrm{KG}_i))^2 / |\mathrm{KG}_i| \tag{14}$$

$$T(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \tag{15}$$

where score_j represents the score of each triple contained in KG_i from IC evaluation models and $|\mathrm{KG}_i|$ is the number of triples that KG_i contains.

Equations (12)–(15) are a process of ranking calculation of IC scores based on normal distribution. Equation (12) first calculates the degree to which the score of the investigated knowledge exceeds the average value, modeling the exceeding degree with a standard normal distribution through (13) and (14). Then, $T(x)$ is used to transform the score ranking to a uniform metric space, ranging from 0 to 1 (higher ranking when closer to 1). Through this transformation, the output scores of different KGs are scaled to the same order of magnitude.

Equation (12) produces a more effective metric by the rank-based scaling. The numerical difference of scores in different KGs could not be accurate in reflecting the difference of IC. Identical numerical difference of two scores reflects a larger gap in a KG where the scores are close to each other in total. Even if these scores are scaled to the same numerical space by min-max normalization, the results are still easily affected by extreme values losing effectiveness. Using rank-based scaling could avoid this issue. In addition, the normal distribution is used to model scoring ranking, which corresponds to the characteristic of KG embedding models that the outputs could be more concentrated around the threshold. To sum up, the

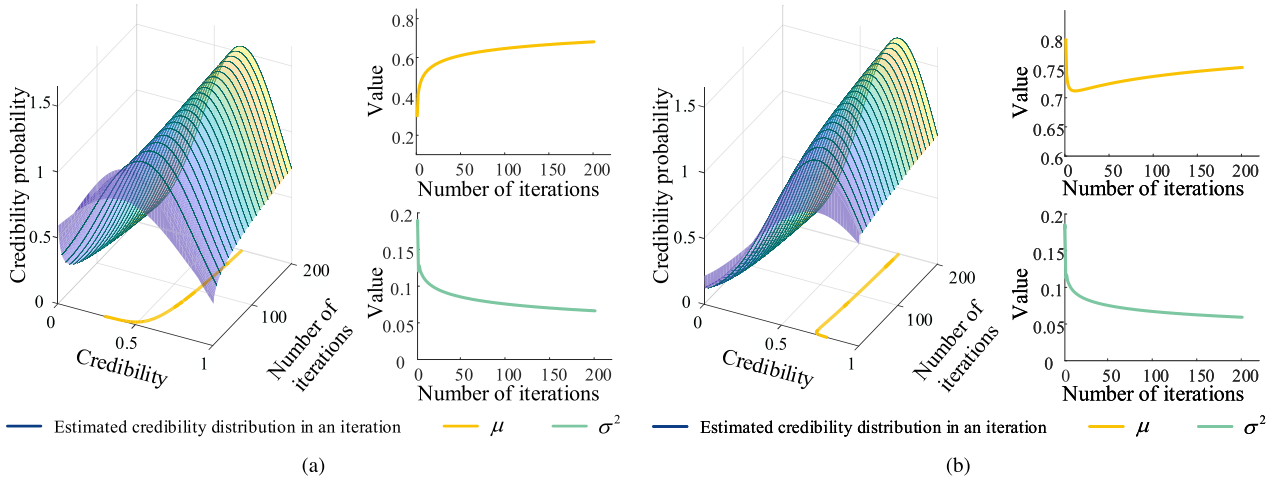


Fig. 5. Change of credibility distributions of triples in the iterations. (a) Credibility distribution of a triple in Dataset A. (b) Credibility distribution of a triple in Dataset B. For (a), five external triples are initialized with $(\mu_e)^0 = 0.8, 0.8, 0.8, 0.8$, and 0.5. The comparison results of EC are correspondingly C, C, C, I, and I (C/I represents that the investigated triple is consistent/inconsistent with the external triple). For (b), five external knowledge elements are initialized with $(\mu_e)^0 = 0.8, 0.3, 0.5, 0.8$, and 0.8. The comparison results of EC are correspondingly I, C, I, C, and C. Note that the investigated triples in (a) and (b) are true, initialized with $(\mu_i)^0 = 0.3$ in (a) and $(\mu_i)^0 = 0.8$ in (b). External triples with C are true (correct) and those with I are fake (incorrect). In the graph of changing trend of σ^2, the initialized value is not displayed since it is a hyperparameter that is not calculated through the proposed method and hardly affects the evaluation results in reality.

ranking of the score uniformly reflects the IC of triples in different KGs.

IV. Experimental Results

A. Datasets

The utilized multisource KGs are generated based on four real-world KGs. Original triples in these KGs are taken as ground truth (correct triples). According to [45], fake (incorrect) triples are added to a generated KG by changing the relation[7] of the original triple to another one with a uniform probability p. Different KGs could be set with different p's to simulate real-world multisource KGs. The utilized datasets generated on the four real-world KGs are introduced as follows.

Dataset A: FB15K237 [46] is a subset of Freebase [47], which contains 237 relations and 14 541 entities, and consists of 310 116 triples in total.

Dataset B: WN18 [33] is a subset of Wordnet, which contains 18 relations and 40 943 entities, and there are 151 442 triples totally.

Dataset C: NELL-995 [48] contains 75 492 entities, 200 relations, and 154 213 total triples.

Dataset D: YAGO3-10 [49] contains 123 182 entities, 37 relations, and 1 089 040 total triples.

B. Experimental Setup

To evaluate the IC of triples, TransE is utilized to construct the embeddings of the entities and relations, in which the scoring function (distance metric to evaluate triple embeddings) $\text{score}_i = -\|h_i + r_i - t_i\|_{1/2}$ is used to evaluate the IC of triple. The training data provided by the FB15K237, WN18, NELL-995, and YAGO3-10 are used to train the TransE model, and the testing data of them are used to generate KGs for experiments. To initialize the quality of triples, 10% triples in a KG are validated by comparing them with the corresponding ground truth. Its mean accuracy is utilized as $(\mu_i)^0$ for all the triples contained in the KG and $(\sigma_i^2)^0$ is uniformly initialized with 0.05. The hyperparameter λ in (7) is assigned with 1 if not specifically illustrated.

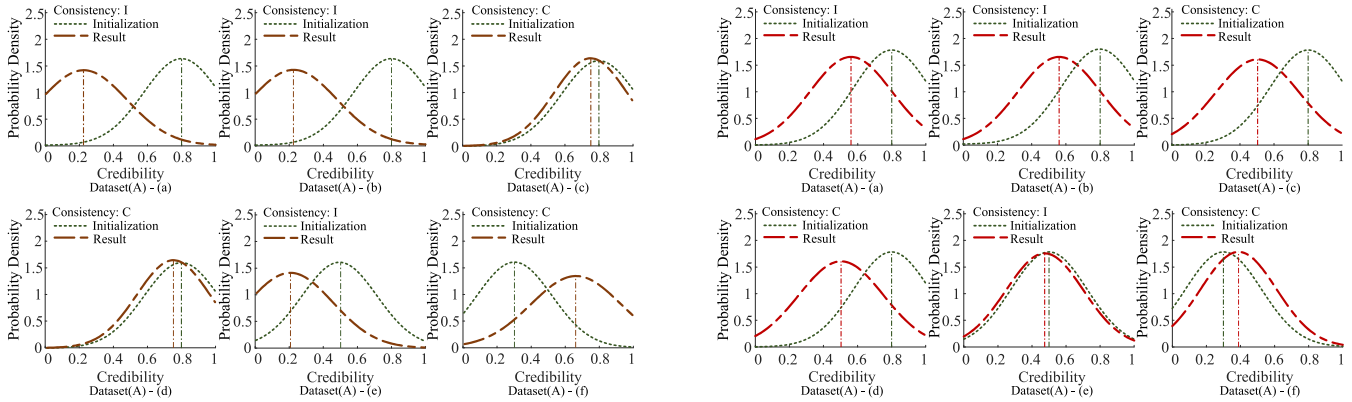
[7]To be concise, the triples having the same head and tail entities are considered to describe the same fact and utilized for EC comparison.

C. Results and Analysis

The characteristics of the proposed method are analyzed through the changing trend of credibility distribution in the iterations and representative cases of the quality evaluation of triples. Moreover, to demonstrate the effect of the proposed method, ablation experiments are conducted, which compares the proposed method with the methods that only utilize EC or IC. In addition, the influence of the number of external KGs and λ of the quality indicator on the performance of the proposed method is also presented.

1) Changing Trend of Credibility Distribution: To analyze the iterative process of updating the quality of triples, experiments are conducted on Datasets (A) and (B) and representative results for the change of credibility distribution are reported in Fig. 5, and the corresponding investigated triples are both true ones. It suggests that both μ and σ change drastically in the first few iterations and then tend to change slower in the following iterations, which indicates that the estimated quality of triples tends to conform to the EC and the IC as the iteration progresses.

For the change of σ^2 in Fig. 5(a) and (b), they are both high in the first few iterations and then degrades, which corresponds to the fact that the obtained credibility is of higher confidence as the iteration progresses, implying the rationality of the proposed method to some extent.



(a) (b)

Fig. 6. Comparison between the proposed method and the method with only EC. (a) Triple quality evaluated through the proposed method. (b) Triple quality evaluated only by EC. The scores of the IC of the six triples are 0.27,0.27,0.91,0.91,0.27, and 0.91. The property "consistency" in subfigures refers to EC with the ground truth, where C (or I) represents the knowledge is consistent (or inconsistent) with the truth, respectively. Knowledge with C is correct and knowledge with I is incorrect. Note that knowledge with the same "consistency" property is consistent with each other.

For the change of μ in Fig. 5(a), it increases from a low value to a high one, for which the external triples with high estimated quality are mostly consistent with the investigated example, thus improving its credibility expectation.

For the change of μ in Fig. 5(6), it first degrades and then increases, for which the initialized quality of some fake triples is high and thus degrades μ of the investigated triple, whereas the incorrect quality of the fake triples is corrected through the comparison of EC. Therefore, the weight of the fake triples becomes lower for the quality update of the investigated triple and its μ increases subsequently.

2) Ablation Experiments: To demonstrate the effect of the proposed method, the proposed method is compared with the methods that only utilize EC (denoted as external-only method) and IC (denoted as internal-only method). For the external-only method, the influence of IC is ignored by assigning all scores with a uniform value, i.e., $\forall i, e\ s_i = s_e = 1$. For the internal-only method, the degree of IC of triples is taken as the quality of triples, i.e., $\forall i\ \ Q_i = s_i$.

a) Comparison with external-only method: To show the effect of incorporating IC in the proposed method, the quality of triples in Dataset A is evaluated through the proposed method and the external-only method. Six KGs are generated for this experiment and the representative results are shown in Fig. 6. Fig. 6(a) represents the quality of six triples evaluated through the proposed method and Fig. 6(b) represents the quality of the same triples evaluated by the external-only method.

For the external-only method, it suggests that the obtained quality[8] of the true triples and the fake triples is very close, even a fake triple [Dataset A in Fig. 6(b)] has the highest expectation of credibility. The reason is that the initialized quality of some fake triples is incorrect. For example, fake triples [Dataset A in Fig. 6(a) and (b)] are initialized with high quality. The incorrect initialization corresponds to the case that there exist low-quality triples in the external KG, which makes the weights[9] of the fake triples high during the comparison of EC, thus leading to the obtained quality of fake triples higher than the quality of true ones.

TABLE II
PERFORMANCE OF THREE METHODS WITH DIFFERENT ICS

Datasets		Methods		
		the internal-only method	the external-only method	the proposed method
Dataset A	20%		**79.7%**	69.7%
	40%		78.7%	**80.2%**
	60%		79.3%	**85.1%**
	80%		79.2%	**88.9%**
Dataset B	20%		**78.7%**	65.3%
	40%		78.3%	**80.5%**
	60%		77.8%	**84.2%**
	80%		80.3%	**89.2%**

For the proposed method, it indicates that the quality of the six triples is evaluated properly, where the obtained quality of fake triples becomes low and that of true triples becomes high. The proper evaluation is due to the IC incorporated in the proposed method. For example, even if fake triples [Dataset A in Fig. 6(a) and (b)] are initialized with high quality, their scores of IC are much lower than those of true triples, and their weights during the comparison of EC are restricted correspondingly, as shown in (6). Thus, the fake triples occupy less weight than the true ones and are prone to get low estimated quality.

The above results demonstrate that the proposed method could reduce the harm of low-quality triples existing in external KGs by incorporating IC.

b) Influence of IC: To analyze the influence of the IC on the proposed method, comparative experiments are conducted on Datasets (A) and (B) with three methods: internal-only method, external-only method, and the proposed method.

[8]The quality here could be regarded as the credibility expectation of triples to show their differences visibly and concisely, which is rational since the variance of credibility of different triples is close in this case.

[9]The weight here represents the degree of the influence of an external triple on the quality update of the investigated triple, expressed as (6).

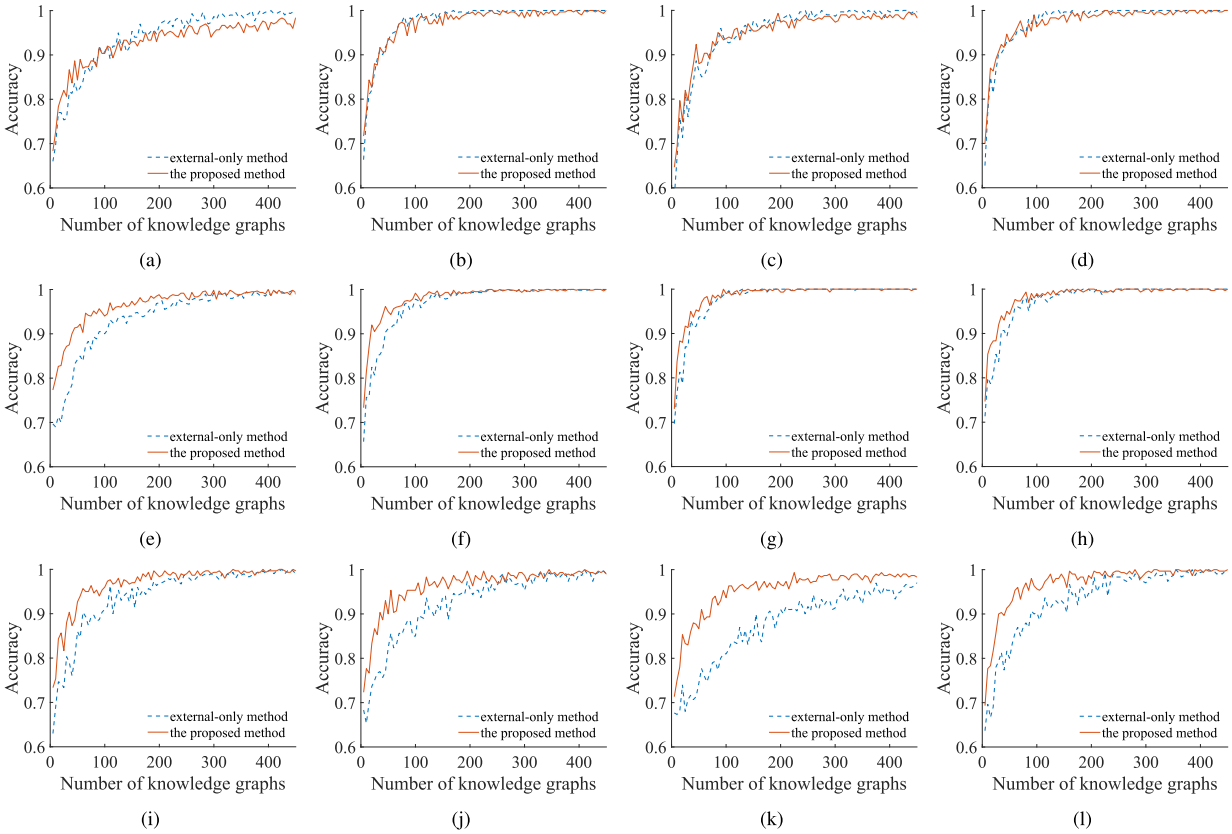


Fig. 7. Performance of two methods under different numbers of KGs, with a different accuracy of IC. (a) 60% IC, Dataset A. (b) 60% IC, Dataset B. (c) 65% IC, Dataset A. (d) 65% IC, Dataset B. (e) 70% IC, Dataset A. (f) 70% IC, Dataset B. (g) 75% IC, Dataset A. (h) 75% IC, Dataset B. (i) 80% IC, Dataset A. (j) 80% IC, Dataset B. (k) 85% IC, Dataset A. (l) 85% IC, Dataset B. In (a), 60% IC represents that the accuracy of the corresponding IC evaluation is fixed with 60%, similar to the other subfigures.

The obtained quality is utilized to detect error relations. For triples that have the same head entity and tail entity, the triple that has the highest credibility expectation and triples consistent with it is considered true, whereas the triples inconsistent with it are considered fake. The results are validated with the ground truth and the accuracy is reported in Table II. Note that the accuracy of IC is with different values[10] to analyze its influence on the proposed method, and ten KGs are generated.

It suggests that as the accuracy of the IC increases, the accuracy of the proposed method increases accordingly, and the proposed method outperforms the internal- and external-only method in most cases. In particular, when the accuracy of the IC is 20%, the performance of the proposed method degrades compared to the external-only method. In this case, since there are too many incorrect evaluation results of IC, lots of low-quality external knowledge may occupy higher weights than the high-quality knowledge, thus leading to more erroneous estimates of KQ. When the accuracy of IC is no less than 40%, it is seen that the proposed method has an obvious performance improvement over the external-only method. It indicates that the proposed method has good resistance to the interference of bad IC accuracy since the IC is properly utilized to correct weights of knowledge in (6).

[10]To obtain different accuracies of IC, triples are chosen by a specific ratio from the correct testing results and the error testing results of TransE, which detects error relations. These triples are gathered into a triple set on which the internal-only method has a fixed accuracy according to the ratio.

3) Analysis on the Number of External KGs: To analyze the influence of the number of external KGs on the proposed method, experiments are conducted on Datasets (A) and (B) with different numbers of external KGs. KGs are generated with a random probability p ranging from 0.3 to 0.8 to produce fake triples. To analyze the influence of the number of KGs more clearly, experiments are conducted under different conditions of fixed IC accuracy, and the corresponding performance of the external-only method is reported together with the proposed method in Fig. 7.

It suggests that the performance of the proposed method is improved as the number of external KGs increases, drastically at the beginning and then more slowly. The changing trend of the performance of the proposed method is similar to that of the external-only method, for which the comparison with external knowledge is more sufficient with more external KGs, thus leading to more accurate quality evaluation. When the number of KGs is small, the sampling of the contained knowledge may be extremely uncertain, where there may be a large proportion of incorrect knowledge describing certain facts greatly interfering with the quality evaluation. In this case, it is seen that the IC plays a significant role, which obviously improves the performance based on insufficient EC comparison.

For the condition of IC accuracy, it indicates that under the condition of low IC accuracy [as shown in Fig. 7(a)–(d)], the proposed method outperforms the external-only method

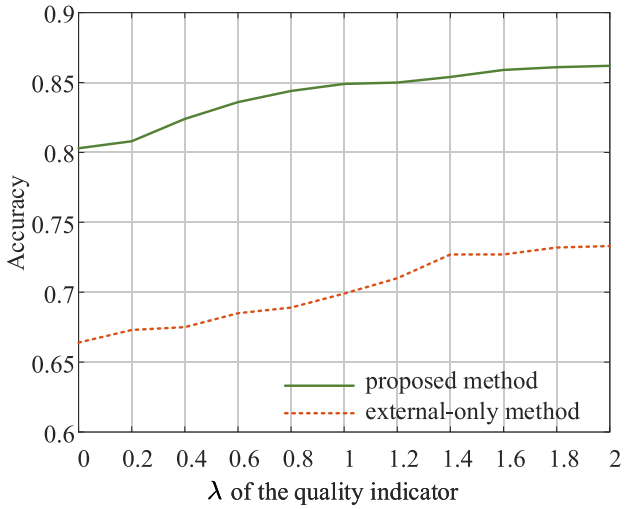


Fig. 8. Accuracy of the proposed method and external-only method with different λ's of the quality indicator.

with no more than 100 KGs, whereas the performance of the external-only method may exceed that of the proposed method with more KGs. The reason is that the EC comparison is sufficient with lots of external KGs; however, there exist many wrong evaluations of IC of knowledge under the condition of low IC accuracy, which cannot provide effective enough information for the sufficient EC comparison. In addition, it suggests that with higher IC accuracy, the proposed method outperforms the external-only method to a greater extent, which indicates the effectiveness of the IC integrated in the proposed method.

The above results indicate that the effect of the proposed method is positively correlated with the number of external KGs due to more sufficient EC comparison with more KGs. This experiment also proves the effectiveness of the integrated IC in the proposed method from a different aspect. Moreover, the performance improvement produced by the integrated IC is especially significant when the number of external KGs is not enough to provide sufficient EC comparison.

4) Analysis on λ of the Distribution-Based Quality Indicator: To analyze λ of (7), experiments are conducted with different λ's ranging from 0 to 2. Note that λ value cannot be too large since $g = \mu - \lambda\sigma$ could be negative with large λ, thus making (8) ineffective to update the variance of the distribution-based quality. The accuracy of the proposed method of external-only method on Dataset A is reported in Fig. 8.

It suggests that the performance of the proposed and external-only method grows first fast and then slower as λ increases. This result indicates that the distribution-based KQ evaluation is effective to provide more accurate results when the uncertainty σ of the estimated credibility weights higher in the iterative process. Note that when $\lambda = 0$, the distribution-based quality degenerates into a quality metric. Therefore, it also demonstrates the effectiveness of introducing distribution to KQ evaluation.

V. Discussion

A. Comparison With Other Methods

In order to show the effect of the proposed method, comparative experiments are conducted on the four datasets to determine true triples in multiple KGs and the proposed method is compared with TransE- andTransD-based method, MV, and EM algorithm. TransE and TransD are representative internal methods, which only utilizes the knowledge contained in the KG itself to evaluate the quality of triples. Specifically, they train the embedding representation of triples by learning the structure of the KG and evaluate the rationality of triples through the scoring function. TransE uses the scoring function $\|h + r - t\|_{1/2}$. TransD uses the scoring function $\|h_\perp + r - t_\perp\|_{1/2}$, where $h_\perp = (r_p h_p^T + I)h$ and $t_\perp = (r_p t_p^T + I)t$. r_p, h_p^T, and t_p^T are rejection vectors. Majority vote and EM algorithm evaluate the quality of triples through the external comparison between knowledge of different KGs. Majority vote chooses the most occurring knowledge as the correct knowledge. The EM algorithm provides the correct possibility of each triple, which is taken as its estimated quality. The triple with the highest estimated quality[11] in the triples that share the same head entity and tail entity is considered true.

For the generated KGs, 1000 original triples are randomly chosen from FB15K237, WN18, NELL-995, and YAGO3-10 as true triples, and KGs are generated with different settings of probability p to produce fake triples. For the setting of p to produce incorrect triples, Rodrigues *et al.* [45] provided a choice, and it is utilized in the performance comparison. Moreover, to show the performance of the methods in different conditions of interference of incorrect knowledge, three additional settings of p are also utilized in the comparative experiment for a more comprehensive performance comparison. The accuracy of the proposed method and the four comparison methods as well as the four utilized settings of p is reported in Table III.

To statistically analyze the comparison results, the Friedman test with the Wilcoxon signed-rank *post hoc* tests [53] is utilized on the accuracy of the five methods. They are conducted based on the rank of each method under every p setting of each dataset (16 in total). The average ranks of each method are shown in Table III. According to the Friedman test, there is statistically a significant difference in the effectiveness of the five methods, $\chi^2(15) = 127.2$ and $p < 0.001$. The Wilcoxon signed-rank *post hoc* tests show that there are significant improvements in performance for the proposed method over the four comparative methods, i.e., TransE-based method ($Z = -3.70$ and $p < 0.001$), TransD-based method ($Z = -3.63$ and $p < 0.001$), majority vote ($Z = -3.70$ and $p < 0.001$), and EM algorithm ($Z = -3.67$ and $p < 0.001$). The statistical analysis of the results indicates that the proposed method outperforms both the internal methods (TransE- and TransD-based methods) and external methods (MV and EM algorithm), which demonstrates that the

[11]For the proposed method, g_i is utilized as the indicator for the estimated quality. For TransE- and TransD-based methods, their output scores for each triple are utilized as its estimated quality. For MV and EM algorithm, their original criteria for selecting correct knowledge are utilized.

TABLE III
ACCURACY OF FOUR OTHER METHODS AND THE PROPOSED METHOD

Dataset	FB15K237/%				WN18/%				NELL-995/%				YAGO3-10/%				**Average rank**
Setting of p	p_1	p_2	p_3	p_4	p_1	p_2	p_3	p_4	p_1	p_2	p_3	p_4	p_1	p_2	p_3	p_4	
TransE-based method [33]	84.0	82.9	83.4	82.3	68.0	68.2	67.4	67.7	71.2	70.8	69.3	70.4	73.4	72.9	72.6	73.1	4.25
TransD-based method [50]	87.6	86.8	86.1	85.6	82.5	82.2	81.9	81.8	75.8	74.1	73.6	74.2	88.1	88.3	87.8	87.1	2.75
Majority Vote [51]	81.4	67.7	56.1	59.6	82.1	70.1	58.8	62.1	81.8	65.1	58.3	59.3	80.8	68.8	55.7	61.5	4.69
EM algorithm [52]	95.6	89.6	85.4	82.9	95.2	90.3	86.1	84.2	96.7	88.7	85.6	82.8	86.6	90.5	82.3	83.9	2.31
Proposed method	**98.5**	**93.8**	**91.6**	**88.9**	**97.9**	**94.1**	**89.3**	**87.2**	**97.4**	**89.6**	**86.7**	**84.6**	**97.8**	**92.0**	**88.5**	**88.3**	**1.00**

*For all datasets, KGs are generated based on the p_1 setting to produce incorrect triples provided in [45] and three additional settings p_2, p_3, p_4 for a more comprehensive performance comparison, which are $p_1 = \{0.01, 0.1, 0.3, 0.5, 0.7\}$, $p_2 = \{0.1, 0.2, 0.3, 0.5, 0.7\}$, $p_3 = \{0.1, 0.3, 0.4, 0.5, 0.7\}$ and $p_4 = \{0.2, 0.2, 0.4, 0.5, 0.7\}$, respectively.
*TransE and TransD models are trained on the training data of FB15K237, WN18, NELL-995 and YAGO3-10. Triples of the generated KGs are from the testing data of them.

incorporation of IC with EC is effective to evaluate the quality of triples properly.

B. Motivation of the Uniform Metric of IC

An IC transformation method is proposed to provide uniform IC metrics for triples in different KGs, as introduced in Section III-D. Equation (12) is the core step for this metric transformation. Its design is motivated by the "empirical rule" of normal distribution. The IC evaluation scores of triples from an embedding-based model on a KG are regarded as samples with certain distribution (assumed as normal distribution in this article). Even if these scores of triples in different KGs may differ in numerical magnitude, their ranking in the corresponding score distribution could represent the IC level of a triple under the criteria of the embedding model on the KG. To assess this ranking, the empirical rule is an effective approach that could estimate the probability of a sample in a certain range around the mean value of samples, thus deriving the ranking of each sample. As for a known distribution, the ranking of a sample could be explicitly estimated by calculating the integral over the distribution from $-\infty$ to its value. Therefore, (12) is designed as this integral to provide a uniform metric for IC scores in different KGs.

C. Utilization of IC Assessment Models

TransE is utilized to produce IC scores of triples for experiments in this article. In addition to TransE, other embedding methods are also feasible for the proposed method only if they could provide a quantitative assessment of triple quality merely based on the KG structure. Some of these methods process with different distance functions to obtain embeddings, such as TransD, TransH, and TransR [43], [50], [54]. Besides, NN-based methods are also utilized for KG embedding, which utilize the hidden layer of NNs to encode entities and relations by training with different architectures. The architecture of an NN-based KG embedding model represents the scoring function it uses. Representative models are semantic matching energy (SME), neural tensor network (NTN), and neural association model (NAM) [55]–[57].

D. Limitations

The proposed method explores a way that incorporates the IC of KGs with EC comparison to evaluate the triple quality. However, the combination approach of the two aspects is implemented in a serial way, i.e., the evaluation of IC and the triple quality is independent, which may suffer from the accumulated errors. The errors in the IC evaluation may mislead the quality evaluation, for example, a high-quality triple may be misjudged as low IC, whose weight may be greatly reduced and become lower than the incorrect triples in the EC comparison, thus leading to a wrong quality evaluation result (actually, this kind of errors can already be avoided to some extent in the proposed method). To address this issue, an end-to-end way could be explored to fuse the IC and EC and eliminate the accumulated errors.

E. Future Work

This article utilizes the IC of KGs as auxiliary information to identify low-quality knowledge, integrating it with external comparison among knowledge for proper KQ evaluation. Except for the internal structure of KGs, there also exist other types of information capable to support KQ evaluation. Among them, a type that has recently attracted much research attention is the causality inference from data [58]–[61]. The causality could be considered as a special form of semantic relation of two entities, similar to triples of the KG. In comparison to IC, the causality is inferred from real data, which could validate the knowledge created by human experiences with the objective facts. The causality might be taken as stronger evidence to detect low-quality triples or even discover missing knowledge, which has great potential to be further used in KG improvement tasks. For example, based on the causal analysis of head and tail entities, error correction and KG completion could be processed. This idea is also interesting since it introduces knowledge mining from data in the number form to the KQ evaluation, which has been seldom researched yet. In addition, the tuning process of hyperparameters could be time-consuming, such as λ and hyperparameters in the used embedding model. Therefore, self-adaption of hyperparameters could be integrated to improve the work in future [62].

VI. CONCLUSION

This article explores a combination approach to evaluate the quality of triples, which incorporates IC of a triple with its EC compared to triples of other KGs. To properly utilize IC, an IC transformation method is proposed to transform the IC scores of triples in different KGs into a uniform

framework. Experiments are conducted on different generated multisource KGs to analyze the characteristics and effect of the proposed method. It demonstrates that the incorporation of IC in the proposed method is effective to alleviate the interference caused by incorrect external knowledge, which outperforms both the external methods and internal methods. Furthermore, the experimental results show that the proposed method has a high tolerance for the accuracy of IC evaluation. In addition, it indicates that the proposed method has a significant performance improvement produced by IC in the case of insufficient EC comparison, i.e., when there is a small number of KGs.

The proposed KQ evaluation method has great potential to assist related applications of KGs. For example, the quality of knowledge could be used to select high-quality knowledge to solve the conflict between knowledge in the task of knowledge fusion or to detect and filter wrong knowledge in the KGs to improve the accuracy of KGs. Moreover, other types of auxiliary information for KQ evaluation are to be explored, especially for causality inference from data. In the future, attempts will be made to integrate knowledge mining from data with the KQ evaluation to provide a more comprehensive and practical framework.

ACKNOWLEDGMENT

The authors appreciate the comments from anonymous reviewers, which helped to improve this article.

REFERENCES

[1] X. Wang *et al.*, "Knowledge graph quality control: A survey," *Fundam. Res.*, vol. 1, no. 5, pp. 607–626, Sep. 2021.
[2] X. Wu *et al.*, "Knowledge engineering with big data," *IEEE Intell. Syst.*, vol. 30, no. 5, pp. 46–55, Sep./Oct. 2015.
[3] X. Zhao, X. Feng, and H. Chen, "A background knowledge revising and incorporating dialogue model," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Nov. 18, 2021, doi: 10.1109/TNNLS.2021.3123128.
[4] X. Zhao, L. Chen, and H. Chen, "A weighted heterogeneous graph-based dialog system," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Nov. 18, 2021, doi: 10.1109/TNNLS.2021.3124640.
[5] X. Zhao, H. Chen, Z. Xing, and C. Miao, "Brain-inspired search engine assistant based on knowledge graph," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 5, 2021, doi: 10.1109/TNNLS.2021.3113026.
[6] A. Bordes and E. Gabrilovich, "Constructing and mining web-scale knowledge graphs," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2014, pp. 1195–1197.
[7] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer, "Quality assessment for linked data: A survey," *Semantic Web*, vol. 7, no. 1, pp. 63–93, Mar. 2015.
[8] M. Färber and A. Rettinger, "Which knowledge graph is best for me?" 2018, *arXiv:1809.11099*.
[9] C. Akkaya, A. Conrad, J. Wiebe, and R. Mihalcea, "Amazon mechanical Turk for subjectivity word sense disambiguation," in *Proc. Workshop Creating Speech Lang. Data Amazon's Mech. Turk (NAACL HLT)*, 2010, pp. 195–203.
[10] S. Kubler, W. Derigent, A. Voisin, J. Robert, and Y. L. Traon, "Knowledge-based consistency index for fuzzy pairwise comparison matrices," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, Jul. 2017, pp. 1–7.
[11] H. Paulheim, "Knowledge graph refinement: A survey of approaches and evaluation methods," *Semantic Web*, vol. 8, no. 3, pp. 489–508, 2017.
[12] H. Chen, G. Cao, J. Chen, and J. Ding, "A practical framework for evaluating the quality of knowledge graph," in *Proc. China Conf. Knowl. Graph Semantic Comput.*, 2019, pp. 111–122.
[13] X. Zhao, F. Xiao, H. Zhong, J. Yao, and H. Chen, "Condition aware and revise transformer for question answering," in *Proc. Web Conf.*, 2020, pp. 2377–2387.
[14] S. Lyu and H. Chen, "Relation classification with entity type restriction," in *Proc. Findings Assoc. Comput. Linguistics (ACL-IJCNLP)*, 2021, pp. 390–395.
[15] H. Lin, J. Li, X. Zhang, and H. Chen, "Grammatical error correction with dependency distance," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, 2021, pp. 1018–1027.
[16] Z. H. Syed, M. Röder, and A.-C. N. Ngomo, "FactCheck: Validating RDF triples using textual evidence," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2018, pp. 1599–1602.
[17] N. Q. V. Hung *et al.*, "Answer validation for generic crowdsourcing tasks with minimal efforts," *VLDB J.*, vol. 26, no. 6, pp. 855–880, Dec. 2017.
[18] A. Zaveri *et al.*, "Quality assessment methodologies for linked open data," *Semantic Web J.*, vol. 7, no. 1, pp. 63–93, 2015.
[19] J. Liu, Z. Liu, and H. Chen, "Revisit word embeddings with semantic lexicons for modeling lexical contrast," in *Proc. IEEE Int. Conf. Big Knowl.*, Aug. 2017, pp. 72–79.
[20] Y. Lin, Z. Liu, H. Luan, M. Sun, S. Rao, and S. Liu, "Modeling relation paths for representation learning of knowledge bases," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 705–714.
[21] B. Xue and L. Zou, "Knowledge graph quality management: A comprehensive survey," *IEEE Trans. Knowl. Data Eng.*, early access, Feb. 10, 2022, doi: 10.1109/TKDE.2022.3150080.
[22] D. Gerber *et al.*, "DeFacto: Temporal and multilingual deep fact validation," *J. Web Semantics*, vol. 35, pp. 85–101, Dec. 2015.
[23] X. Yin, J. Han, and P. S. Yu, "Truth discovery with multiple conflicting information providers on the web," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 6, pp. 796–808, Jun. 2008.
[24] J. Pasternack and D. Roth, "Generalized fact-finding," in *Proc. 20th Int. Conf. Companion World Wide Web*, 2011, pp. 99–100.
[25] J. Pasternack and D. Roth, "Making better informed trust decisions with generalized fact-finding," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 2324–2329.
[26] J. Pasternack and D. Roth, "Latent credibility analysis," in *Proc. 22nd Int. Conf. World Wide Web*, 2013, pp. 1009–1020.
[27] J. Le, A. Edmonds, V. Hester, and L. Biewald, "Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution," in *Proc. Workshop Crowdsourcing Search Eval. (SIGIR)*, vol. 2126, 2010, pp. 22–32.
[28] J. Wu, C.-W. Wong, X. Zhao, and X. Liu, "Toward effective automated content analysis via crowdsourcing," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2021, pp. 1–6.
[29] G. Foody *et al.*, "Increasing the accuracy of crowdsourced information on land cover via a voting procedure weighted by information inferred from the contributed data," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 3, pp. 80–91, 2018.
[30] S. K. Ng, T. Krishnan, and G. J. McLachlan, "The EM algorithm," in *Handbook of Computational Statistics*. New York, NY, USA: Springer, 2012, pp. 139–172.
[31] M. Färber, F. Bartscherer, C. Menne, and A. Rettinger, "Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO," *Semantic Web*, vol. 9, no. 1, pp. 77–129, 2018.
[32] K. Dentler, R. Cornet, A. ten Teije, and N. de Keizer, "Comparison of reasoners for large ontologies in the OWL 2 EL profile," *Semantic Web*, vol. 2, no. 2, pp. 71–87, 2011.
[33] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 2787–2795.
[34] Z. Sun, Z.-H. Deng, J.-Y. Nie, and J. Tang, "RotatE: Knowledge graph embedding by relational rotation in complex space," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–18.
[35] H. Paulheim and C. Bizer, "Improving the quality of linked data using statistical distributions," *Int. J. Semantic Web Inf. Syst.*, vol. 10, no. 2, pp. 63–86, Apr. 2014.
[36] A. Melo and H. Paulheim, "Detection of relation assertion errors in knowledge graphs," in *Proc. Knowl. Capture Conf.*, Dec. 2017, pp. 1–8.
[37] D. Wienand and H. Paulheim, "Detecting incorrect numerical data in DBpedia," in *Proc. Eur. Semantic Web Conf.* Heidelberg, Germany: Springer, 2014, pp. 504–518.
[38] R. Zhang, M. Indulska, and S. Sadiq, "Discovering data quality problems," *Bus. Inf. Syst. Eng.*, vol. 61, no. 5, pp. 575–593, Oct. 2019.
[39] L. P. English, *Information Quality Applied: Best Practices for Improving Business Information, Processes and Systems*. Hoboken, NJ, USA: Wiley, 2009.
[40] M. Ahsanullah, B. G. Kibria, and M. Shakil, "Normal distribution," in *Normal and Student st Distributions and Their Applications*. Paris, France: Atlantis Press, 2014, pp. 7–50.
[41] J. N. Kapur and H. K. Kesavan, "Entropy optimization principles and their applications," in *Entropy and Energy Dissipation in Water Resources*. Berlin, Germany: Springer, 1992, pp. 3–20.

[42] R. Xie *et al.*, "Representation learning of knowledge graphs with hierarchical types," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 2965–2971.
[43] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *Proc. AAAI Conf. Artif. Intell.*, vol. 28, no. 1, Jun. 2014, pp. 1112–1119.
[44] M. Nickel, V. Tresp, and H.-P. Kriegel, "A three-way model for collective learning on multi-relational data," in *Proc. 28th Int. Conf. Int. Conf. Mach. Learn.*, 2011, pp. 809–816.
[45] F. Rodrigues, F. Pereira, and B. Ribeiro, "Sequence labeling with multiple annotators," *Mach. Learn.*, vol. 95, no. 2, pp. 165–181, May 2014.
[46] M. Nickel, L. Rosasco, and T. Poggio, "Holographic embeddings of knowledge graphs," in *Proc. AAAI Conf. Artif. Intell.*, vol. 30, no. 1, 2016, pp. 1955–1961.
[47] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in *Proc. ACM Special Interest Group Manage. Data Int. Conf. Manage. Data*, 2008, pp. 1247–1250.
[48] D. Nathani, J. Chauhan, C. Sharma, and M. Kaul, "Learning attention-based embeddings for relation prediction in knowledge graphs," 2019, *arXiv:1906.01195*.
[49] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel, "Convolutional 2D knowledge graph embeddings," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 1811–1818.
[50] 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.
[51] W. Tang and M. Lease, "Semi-supervised consensus labeling for crowdsourcing," in *Special Interest Group Inf. Retr. Workshop Crowdsourcing Inf. Retr.*, 2011, pp. 1–6.
[52] D. R. Karger, S. Oh, and D. Shah, "Iterative learning for reliable crowdsourcing systems," in *25th Annu. Conf. Neural Inf. Process. Syst.*, 2011, pp. 1–9.
[53] W. Min, E. Y. Ha, J. Rowe, B. Mott, and J. Lester, "Deep learning-based goal recognition in open-ended digital games," in *Proc. 10th AAAI Conf. Artif. Intell. Interact. Digit. Entertainment*, 2014, pp. 37–43.
[54] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 2181–2187.
[55] A. Bordes, X. Glorot, J. Weston, and Y. Bengio, "A semantic matching energy function for learning with multi-relational data," *Mach. Learn.*, vol. 94, no. 2, pp. 233–259, Feb. 2014.
[56] 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.*, vol. 26, 2013, pp. 926–934.
[57] Q. Liu *et al.*, "Probabilistic reasoning via deep learning: Neural association models," 2016, *arXiv:1603.07704*.
[58] X. Wu, B. Jiang, K. Yu, and H. Chen, "Separation and recovery Markov boundary discovery and its application in EEG-based emotion recognition," *Inf. Sci.*, vol. 571, pp. 262–278, Sep. 2021.
[59] X. Wu, B. Jiang, K. Yu, H. Chen, and C. Miao, "Multi-label causal feature selection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 4, 2020, pp. 6430–6437.
[60] X. Wu, B. Jiang, K. Yu, C. Miao, and H. Chen, "Accurate Markov boundary discovery for causal feature selection," *IEEE Trans. Cybern.*, vol. 50, no. 12, pp. 4983–4996, Dec. 2020.
[61] X. Wu, B. Jiang, Y. Zhong, and H. Chen, "Tolerant Markov boundary discovery for feature selection," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2020, pp. 2261–2264.
[62] X. Luo, Y. Yuan, S. Chen, N. Zeng, and Z. Wang, "Position-transitional particle swarm optimization-incorporated latent factor analysis," *IEEE Trans. Knowl. Data Eng.*, early access, Oct. 23, 2020, doi: 10.1109/TKDE.2020.3033324.

Taiyu Ban received the B.Sc. degree in computer science and technology from the School of the Gifted Young, University of Science and Technology of China, Hefei, China, in 2020. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, University of Science and Technology of China.

His current research interests include machine learning and knowledge engineering.

Xiangyu Wang received the B.Sc. degree from Donghua University, Shanghai, China, in 2015. He is currently pursuing the Ph.D. degree with the School of Data Science, University of Science and Technology of China, Hefei, China.

His research interests include knowledge engineering and machine learning.

Lyuzhou Chen received the B.Sc. degree from the University of Science and Technology of China, Hefei, China, in 2020, where he is currently pursuing the Ph.D. degree with the School of Data Science.

His current research interests include ensemble learning, knowledge engineering, and causal learning.

Xingyu Wu received the B.Sc. degree from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2018. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, University of Science and Technology of China (USTC), Hefei, China.

He has published some scientific papers in prestigious journals and conferences. His research interests include causal learning and causal inference.

Mr. Wu served as a Reviewer for IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, *International Journal of Information Security*, and IEEE/CAA Journal of Automatica Sinica (JAS) and a PC Member of Association for the Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence, and Conference on Empirical Methods in Natural Language Processing (EMNLP).

Qiuju Chen received the B.Sc. degree from the School of Information Science and Technology, University of Science and Technology of China, Hefei, China, in 2004, and the Ph.D. degree from the Institute of Electronic Engineering, Hefei, China, in 2016.

She is currently an Associate Research Fellow with the University of Science and Technology of China. Her research interests include machine learning and knowledge engineering.

Huanhuan Chen (Senior Member, IEEE) received the B.Sc. degree from the University of Science and Technology of China (USTC), Hefei, China, in 2004, and the Ph.D. degree in computer science from the University of Birmingham, Birmingham, U.K., in 2008.

He is currently a Full Professor with the School of Computer Science and Technology, USTC. His current research interests include neural networks, Bayesian inference, and evolutionary computation.

Dr. Chen received the 2015 International Neural Network Society Young Investigator Award, the 2012 IEEE Computational Intelligence Society Outstanding Ph.D. Dissertation Award, the IEEE TRANSACTIONS ON NEURAL NETWORKS Outstanding Paper Award (bestowed in 2011 and only one paper in 2009), and the 2009 British Computer Society Distinguished Dissertations Award. He is also an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS and the IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE.