A Comprehensive Approach to Assess Trustworthiness and Completeness of Knowledge Graphs

Jumana Alsubhi*, Abdulrahman Gharawi, and Lakshmish Ramaswamy

School of Computing, University of Georgia, Athens, GA, USA
Email: jumana.alsubhi@uga.edu (J.A.); abdulrahman.gharawi@uga.edu (A.G.); laksmr@uga.edu (L.R.)
*Corresponding author
Manuscript received September 10, 2023; revised October 21, 2023; accepted November 25, 2023; published March 20, 2024.

***Abstract*—Knowledge Graphs (KGs) have been widely used for many tasks such as question-answering, recommendation, natural language processing, and so on. The quality of KGs is a crucial factor in determining whether a specific KG is appropriate for usage in a certain application. Completeness and trustworthiness are two dimensions that are used to assess the quality of KGs. Estimation of the completeness and trustworthiness of a largescale knowledge graph often requires humans to annotate samples from the graph. How to obtain statistically meaningful estimates for quality evaluation while keeping humans out of the loop to reduce cost is a critical problem. Nowadays, to reduce the costs of the manual construction of knowledge graphs, many KGs have been constructed automatically from sources with varying degrees of trustworthiness. Therefore, possible noises and conflicts are inevitably introduced in the process of construction, which severely interferes with the quality of constructed KGs. Many works have been done to detect noisy triples; however, how to estimate the quality of the entire KG has largely been ignored in prior research. To fill this gap, we propose a new approach to automatically evaluate and assess existing KGs in terms of completeness and trustworthiness. In this paper, we conduct several experiments on three standard datasets, namely, FB15K, WN18, and NELL995 to estimate the quality of KGs and assign a specific score based on the completeness and trustworthiness of the KG. The experimental findings demonstrate the reliability of the proposed scores.**

***Keywords*—assessment, completeness, data quality, knowledge graph, noise detection, trustworthiness**

I. Introduction

Knowledge Graphs (KGs) have been widely used for many tasks such as question-answering, web search, recommendations, natural language processing, and so on. The number of large-scale KGs containing millions of relational data in the form of RDF triples (subject, predicate, object) has increased during the past several years, such as Freebase [1], WordNet [2], NELL [3], etc. The majority of traditional knowledge graph generation techniques often involve extensive human supervision, which is incredibly labor- and time-intensive. Nowadays, many KGs have been automatically constructed from sources with different levels of trustworthiness and completeness [4]. Since these KGs contain incorrect facts or noise or missing facts, it is essential to understand the quality of KG in order to advise downstream applications and assist them in coping with any data quality uncertainty. Despite its significance, the problem of evaluating the trustworthiness and completeness of KG has been largely ignored by prior academic research.

There are several ways to define trustworthiness. For instance, the user's acceptance of the information as right, genuine, real, and credible is defined by its trustworthiness; trustworthiness also refers to an entity's or KG's reputation, which is based on personal experience or third-party recommendations [5, 6]. In this context, the trustworthiness of KG can be defined as the percentage of triples in the KG being correct.

Detecting noises in large-scale KGs that involve extensive human efforts to assess the correctness of facts has been done relying on a worldwide crowd-sourcing effort. This is very expensive and extremely labor-intensive, and time-consuming. Recently, there have been some researchers that focus on automatic KG noise detection. In order to judge the correctness of a triple, some proposed approaches depend on internal or external information, such as CKRL and Knowledge Vault [7, 8]. The former concentrates on the confidence of each triple to detect noises in KGs, focusing only on internal information, while the latter needs prior knowledge derived from existing KGs to judge the correctness of a triple. Here, we consider a triple being correct based on CKRL. Then, we calculate the trust score for the entire KG.

To evaluate the quality of KGs, some papers explore several main evaluation dimensions of KG quality, such as accuracy, completeness, consistency, timeliness, trustworthiness, and availability [5]. Nevertheless, when we improve the accuracy, timeliness, and consistency of a KG, we also increase its trustworthiness. This illustrates the importance of determining the trustworthiness score of a KG to assess its quality. Furthermore, trustworthiness is regarded as a priori perception of an unconfirmed KG; hence, it can be an empirical metric [9].

Unlike some approaches that evaluate the quality of a KG based on randomly selected sample triples from the KG to determine their correctness, we evaluate the entire KG to come up with an accurate trust score, which has been shown in the results of our experiments. To the best of our knowledge, this work is among the first to propose a new approach to evaluate KGs and assign a certain trustworthiness factor score to compare these KGs in terms of their degrees of trustworthiness. Recently, remarkable large, cross-domain, and open knowledge graphs have been published such as DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. Despite being widely used, it might be challenging to compare these knowledge graphs to one another in a particular situation. Choosing the ideal knowledge graph for their particular needs is a problem for researchers and developers. Thus, having a specific score for each KG that indicates its quality based on trustworthiness will allow us to compare between various KGs and find the

doi: 10.18178/ijke.2024.10.1.143

most suitable one.

On the other hand, completeness refers to the amount of information present in a particular KG [10]. For example, the instance Barack Obama has a data completeness issue when his birthplace is missing in the KG. Completeness can be subjective because it implies that the quantity of data is adequate for the user's needs, which might vary considerably. In this context, completeness can be measured as the percentage of available data divided by the required data.

That being said there are several evaluation dimensions of KG quality, such as accuracy, completeness, consistency, timeliness, trustworthiness, and availability. However, evaluating the completeness of a KG is of high importance because other dimensions of data quality, such as accuracy, timeliness, and consistency, are influenced by completeness. Unlike other quality dimensions of a KG, the evaluation of KG completeness needs a reference or gold standard to compare results against [10].

In this work, we evaluate KGs and assign a completeness score based on the number of facts that it contains. We conduct experiments to determine the completeness score for each KG. According to what we know, this work proposes a new approach to evaluating KGs to assign a completeness factor score to compare these KGs in terms of their completeness. This approach helps us estimate the completeness score for each KG, so we can determine the best KG that would be ideal for certain requirements.

This paper makes the following contributions:

- Evaluate multiple KGs in terms of trustworthiness and completeness.
- Propose a novel approach to generate specific trust and completeness scores for any KG without human supervision.
- Assess our proposed method using three datasets with different levels of noise and various degrees of completeness.
- Conduct a series of empirical experiments on FB15K, WN18, and NELL995 datasets with 10%, 20%, and 40% negative triples for KG trustworthiness evaluation, and the experimental results show the reliability of the proposed trust scores.
- Conduct a series of empirical experiments on FB15K, WN18, and NELL995 datasets with 90%, 80%, and 60% of the triples for KG completeness evaluation, and the experimental results show the thoroughness of the generated completeness scores.

The rest of this paper is organized as follows: Section II surveys related work, and Section III discusses the design of the experimental study. The results and limitations are explored in Section IV and Section V, respectively. In Section VI, we conclude and discuss our plans for future work.

II. Motivation and Related Work

Noises in KGs appear to be inescapable and can have a significant impact on learning. As a result, knowledge construction and knowledge application depend heavily on noise detection. The majority of knowledge graph noise detection research takes place during knowledge graph construction. Numerous various approaches to KG noise detection have been developed by researchers, such as

CKRL, Knowledge Vault, and PTrustE [7, 8, 11]. However, the evaluation of the trustworthiness of the entire KG after construction did not receive appropriate consideration. CKRL suggests three different sorts of triple confidences based on local triple and global path information. To determine whether the triples are credible or not, it combines the multistep relation path with internal information from the triples [7]. To judge triple confidences, the local triple confidence only concentrates on the inside of a triple based on the translation assumption that $h + r \simeq t$, where h, r, and t are the vectors of a head entity, a relation, and a tail entity. As shown in Fig. 1(A), we can infer that Barack Obama is more likely to write A Promised Land rather than What I know For Sure. Thus, the more a triple fits the translation assumption, the more convincing this triple will be. In addition, the global path confidence takes into account 2-step relation paths. For example, in Fig. 1(B), there are two multi-step relation paths from Barack Obama to A Promised Land. Given the lower path, there will be solid evidence to infer the relation write. Therefore, a triple has high global path confidence if it has more reliable paths from its head to tail entity, and these paths are semantically closer to the corresponding relation. Since CKRL detects noises in KGs focusing only on internal information, we will use this model to help in determining the noisy triples, so we can assess the trustworthiness of the entire KG [7].

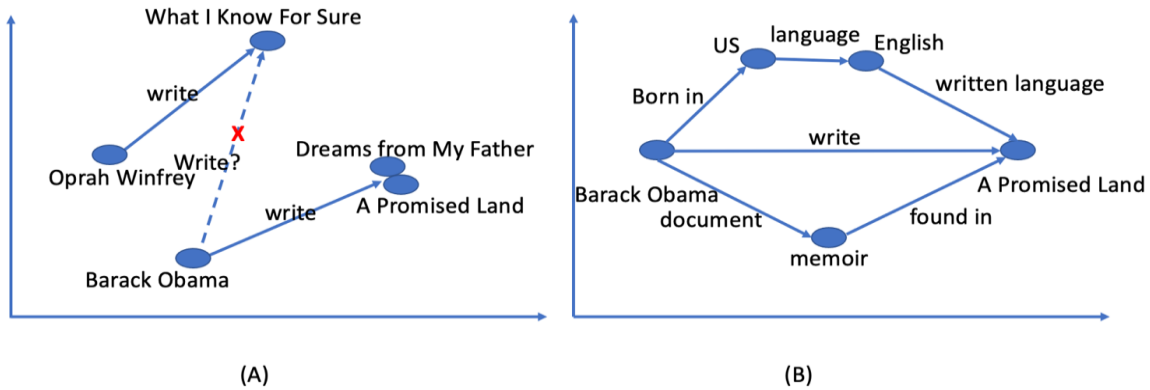

Fig. 1. CKRL mechanism of local triple confidence and global path confidence (A) Local Triple confidence (B) Global path confidence.

There is a common approach for adding synthetic noise to KGs that we adopted in this paper. In order to evaluate the performance of our proposed method to assess the trustworthiness of various KGs with different levels of noise, noise needs to be added to the datasets, which has been done using a common approach in which the head, tail, or relation of a triple is changed randomly [7, 11].

Likewise, there are many KG completion techniques proposed in the literature to predict the missing entities or relationships in the knowledge graph. KGRL, PRA, and NELL have proposed knowledge graph completion techniques that seek to complete the structure of the knowledge graph [12–15]. On the other hand, there is not enough research done to assess the completeness of KGs. To evaluate the completeness of a KG, we randomly removed 10%, 20%, and 40% of the triples from each dataset.

Therefore, in this research paper, we first identify the noisy triples using CKRL, then we assess the trustworthiness of KGs by calculating the percentage of correct triples in the KG. Furthermore, we evaluate the completeness of knowledge graphs by measuring the percentage of found triples divided by the queried triples.

III. Design of Experimental Study

In this empirical study, we designed two different experiments to evaluate the trustworthiness and completeness of KGs. We assign specific trust or completeness scores that correspond to the level of noise or completeness, respectively. We use multiple datasets with different levels of noise and completeness to illustrate the effectiveness of this approach.

Table 1. A sample of WN18 triples

Subject	Predicate	Object
Toy	Hyponym	Swing
Land reform	Hyponym	Reform
Variation	Derivationally related form	Vary
Physics	Member of domain topic	Relativistic
Ear	Part of	Head
Write	Also see	Write up

Specifically, in these experiments, we use three standard datasets, namely, FB15K, WN18, and NELL995. Table I shows samples of the triples from the WN18 dataset. Moreover, Table II displays some statistics of these datasets.

Table 2. Statistics of datasets

Dataset	#Rel	#Ent	#Train	#Valid	#Test
FB15K	1,345	14,951	483,142	50,000	59,071
WN18	18	40,943	141,442	5,000	5,000
NELL995	200	75,492	149,678	543	3,992

A. Assessment of Trustworthiness

The first experiment is conducted to evaluate the trustworthiness of a KG. In this experiment, we obtained three KGs FB15K-N1, FB15K-N2, and FB15K-N3 with 10, 20, and 40 percent of added noise, respectively. The original dataset is FB15K, which is a typical benchmark KG derived from Freebase. Since there are no explicitly labeled noises or conflicts in FB15K, [7] generated new datasets with different levels of noise based on FB15K. Noisy triples are generated automatically by replacing the head, relation, or tail of a triple. Given a genuine triple (Barack Obama, Nationality, American), for instance, (Barack Obama, Nationality, England) is a plausible negative example as opposed to the obviously illogical (Barack Obama, Nationality, Google), as England and American are more prevalent as the tails of Nationality. Thus, the selected entity should have previously appeared in the same position since the majority of errors in real-world KGs result from the misunderstanding between similar entities, as described in [7, 11]. In other words, to construct a negative triple, they randomly alter one of the head or tail entities for a given positive triple in KG. It is required for the formation of negative triples that the new head or tail exists in the head or tail position with the same relation in the KG to make it harder and more confusing.

Likewise, the noise was added to the other datasets, which are WN18 and NELL995. To be more precise, to add a noisy triple from an initial positive triple (h, r, t) in KG, either h or t was randomly switched to generate a negative triple. Following this idea, three noisy KGs based on the aforementioned datasets were acquired, with noisy triples making up 10%, 20%, and 40% of positive triples, respectively.

We used the CKRL model, which identifies potential noise in KGs with some level of confidence for each triple in the KG to identify noises. If the triple has a confidence score that is less than 0.5, we count the triple as noise. On the other hand, triples with a confidence score equal to or greater than 0.5 are considered trusted facts. To calculate the trust score for the entire KG, we calculate the number of trusted triples over the total number of triples, including noise.

B. Assessment of Completeness

In order to assess the completeness of KGs, we obtained three KGs with different levels of completeness 90%, 80%, and 60% of data from each KG. As we mentioned before, the original datasets are FB15K, WN18, and NELL995. We removed 10, 20, and 40 percent of the triples to have three KGs with different levels of completeness. Then, we run random queries on these three KGs. If there is a matching triple for the query in the knowledge graph, we increase the completeness score. We query 40% of the triples to get a great estimate of the completeness of each KG. To calculate the completeness score for each KG, we divide the number of found triples by the total number of queried triples. The results show that the calculated completeness score mirrors the level of completeness of a knowledge graph.

IV. Results and Discussion

That being said we first used various datasets with different levels of noise to evaluate the trustworthiness of each dataset and determine the trust factor accordingly. The results of this experiment will be discussed in Section IV-A. Secondly, we conducted some experiments using the same datasets with various degrees of completeness to assess the completeness of each KG and assign a completeness score based on that. The results for the assessment of completeness are examined in Section IV-B.

A. Experiment 1

Table III shows that the trust score varies depending on the level of noise in each KG. When the noise is only 10% in the KG, we can see that the calculated trustworthiness score is between 86.72% and 91.65%. On the other hand, when the noise increases to 40% in a KG, the trustworthiness score dropped by roughly 30%. The trustworthiness score can represent the amount of noise found in a KG. In other words, the quantity of noise presented in a KG is reflected in the trustworthiness score.

Table 3. Trust scores for multiple datasets with different levels of noise

Noise	FB15K	WN18	NELL995
10	91.65	88.93	86.72
20	80.17	79.54	77.94
40	58.94	57.81	56.13

B. Experiment 2

Table IV shows the generated completeness scores for each dataset. When querying the dataset with 90% of data, the resulting completeness score is around 90%. On the other hand, running the same queries on the KG that only has 60% of the data results in a completeness score of about only 60%. This means that these scores show how many triples are found in each of these KGs. The more complete the KG, the better score will be assigned.

In order to evaluate the KG, we do not indicate that a certain KG is of high quality in terms of trustworthiness and completeness, but rather we assign a specific score for the entire KG. Therefore, in this paper, we subjectively generate a completeness or trust score for a certain KG. Then, based on the application domain, the users of the data can determine if the KG is good enough for their use based on the generated completeness score.

Table 4. Completeness scores for datasets with various degrees of completeness

Datasets	90%	80%	60%
FB15K	88.49	81.36	58.94
WN18	91.2	82.73	59.68
NELL995	86.01	74.73	52.92

V. Limitations

There are many noise detection methods that can be utilized to measure the confidence of each triple; however, in this paper, we only used one model for error detection, namely, CKRL. In future research, we can investigate other models to see if they will improve the produced scores. We can also consider the accuracy of the utilized model by multiplying the output trust score by the accuracy of the model.

For the completeness experiments, even though we conducted each experiment using one dataset with different levels of completeness by removing some triples, we can still apply the same technique to different KGs that include information about the same domain. In other words, although we have created these KGs with different degrees of completeness, we can still apply this approach to different real-world KGs used in the same domain, which was challenging for us to obtain. This can be addressed in future research using techniques that help in matching entities across different Knowledge Graphs such as [16].

Additionally, to make sure that we have a representative set of queries, we can utilize other methods that assist us to estimate node importance in Knowledge Graphs to query such nodes instead of using purely random queries, which also gives promising results [17, 18].

VI. Conclusion

This work focuses on evaluating the trustworthiness and completeness of KGs, which are two of the most important dimensions for Knowledge Graph quality assessment. Through multiple experiments, we developed new approaches to assess the quality of KGs focused on completeness and trustworthiness. We were able to explore the quality of KGs thoroughly by generating a specific score for each dimension.

We automatically obtained statistically meaningful estimates for KG completeness and trustworthiness evaluation with no cost related to expert annotators. The experimental results show that there is a great correlation between the level of noise found in a KG and the assigned trust score. Additionally, there is a strong link that exists between the number of triples available in a KG and the generated completeness score. This indicates the reliability of this comprehensive approach to evaluating the quality of KGs.

Having the proper metrics to evaluate and enhance the quality of Knowledge Graphs is crucial since data of high quality guarantees its suitability for usage in a variety of applications. For instance, determining if a KG with a trust or completeness score of 70% is of high or low quality is subjective to different applications. We do not really say that a KG with a 70% completeness score is of high quality because this level of detail could be sufficient, for example, for the description of a movie but insufficient for a use case in medicine. Rather, this approach aims to generate a specific score for the assessment of trustworthiness and completeness of a KG.

In future work, to calculate the trust score, we can use different noise detection techniques with higher accuracy to determine noises in a KG. Additionally, we can combine multiple noise detection techniques to vote for noisy facts in a KG, which can improve accuracy that will result in a better estimate of the trust score.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

JA contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript; AG and LR gave technical advice and reviewed the paper; all authors had approved the final version.

References

[1] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in *Proc. the 2008 ACM SIGMOD International Conference on Management of Data*, 2008, pp. 1247–1250, http://dx.doi.org/10.1145/ 1376616.1376746.

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

[3] W. Xiong, T. Hoang, and W. Y. Wang, Deeppath: A reinforcement learning method for knowledge graph reasoning. arXiv preprint arXiv:1707.06690, 2017.

[4] Q. Wang, Z. Mao, B. Wang, and L. Guo, L. "Knowledge graph embedding: A survey of approaches and applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 12, pp. 2724–2743, 2017.

[5] X. Wang, L. Chen, T. Ban, M. Usman, Y. Guan, S. Liu, and H. Chen, "Knowledge graph quality control: A survey," *Fundamental Research*, 2021.

[6] M. Gamble and C. Goble, "Quality, trust, and utility of scientific data on the web: To- wards a joint model," in *Proc. the 3rd International Web Science Conference*, 2011, pp. 1–8.

[7] R. Xie, Z. Liu, F. Lin, and L. Lin, "Does William Shakespeare really write Hamlet? Knowledge representation learning with confidence," in *Proc. the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018, April.

[8] X, Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, and W. Zhang, "Knowledge vault: A web-scale approach to probabilistic knowledge fusion," in *Proc. the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 601–610.

[9] C. Bizer, "Quality-driven information filtering in the context of web-based information systems," 2007 Ph.D. thesis.

[10] S. Issa, O. Adekunle, F. Hamdi, S. S. S. Cherfi, M. Dumontier, and A. Zaveri, "Knowledge graph completeness: A systematic literature review," *IEEE Access*, vol. 9, pp. 31322–31339, 2021.

[11] J. Ma, C. Zhou, Y. Wang, Y. Guo, G, Hu, Y. Qiao, and Y. Wang, "PTrustE: A high-accuracy knowledge graph noise detection method based on path trustworthiness and triple embedding," *Knowledge-Based Systems*, vol. 256, 109688, 2022.

[12] Z. Chen, Y. Wang, B. Zhao, J. Cheng, X. Zhao, and Z. Duan, "Knowledge graph completion: A review," *IEEE Access*, vol. 8, pp. 192435–192456, 2020.

[13] Y. Wei, J. Luo, and H. Xie, "KGRL: An OWL2 RL reasoning system for large scale knowledge graph," in *Proc. 12th Int. Conf. Semantics, Knowl. Grids (SKG)*, Beijing, China, Aug. 2016, pp. 83–89.

[14] N. Lao and W. W. Cohen, "Relational retrieval using a combination of path-constrained random walks," *Mach. Learn.*, vol. 81, no. 1, pp. 53–67, Oct. 2010.

[15] 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.

[16] M. Azmy, P. Shi, J. Lin, I. F. Ilyas, "Matching entities across different knowledge graphs with graph embeddings," arXiv preprint arXiv:1903.06607, 2019.

[17] N. Park, A. Kan, X. L. Dong, T. Zhao, and C. Faloutsos, "Estimating node importance in knowledge graphs using graph neural networks," in *Proc. the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, 2019, July, pp. 596-606.

[18] H. Huang, L. Sun, B. Du, C. Liu, W. Lv, and H. Xiong, "Representation learning on knowledge graphs for node importance estimation," in *Proc. the 27th ACM SIGKDD Conference on Knowledge Discovery Data Mining*, 2021, August, pp. 646–655.