DCQE: A RDF Dataset Quality Evaluation Mechanism for Decentralized Systems

Li Huang [a,b], Zhenzhen Liu [a,b], Fangfang Xu [a,b] and Jinguang Gu [a,b,c,*]

[a] *Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Hubei, Wuhan 430065, China,*
E-mails: huangli82@wust.edu.cn, 540900559@qq.com, xuff@wust.edu.cn, simon@wust.edu.cn

[b] *College of Computer Science and Technology, Wuhan University of Science and Technology, Hubei, Wuhan 430065, China*

[c] *Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content, SAPPRFT, Beijing 100038, China.*

Editors: First Editor, University or Company name, Country; Second Editor, University or Company name, Country
Solicited reviews: First Solicited Reviewer, University or Company name, Country; Second Solicited Reviewer, University or Company name, Country
Open reviews: First Open Reviewer, University or Company name, Country; Second Open Reviewer, University or Company name, Country

Abstract. The current decentralized system has developed rapidly, especially with the development of blockchain technology. The quality evaluation of RDF data sets in the decentralized system has also received extensive attention. Therefore, from the perspective of data quality evaluation, this paper proposes a RDF data quality evaluation model in decentralized environment, and points out the new dimension of RDF data quality. The blockchain is used to record the data quality evaluation results and the update plan of the quality evaluation results is designed in detail. Finally, the feasibility of the above system is verified and the quality evaluation model is verified. The purpose of this paper is to study how the decentralized system can provide users with better cost performance when the knowledge is independently protected. This paper named this scheme DCQE.

Keywords: Decentralization, Quality Evaluation, Blockchain, RDF dataset

1. Introduction

In recent years, RDF has been widely used in various fields, and there have been many products and applications that use RDF as a data framework. The quantity and quality of RDF data published on the Internet have been greatly improved. The quality of an RDF data set means the correctness and availability of data. RDF data quality evaluation has been favored by many researchers [1]-[3]. Assaf et al. [4] summarized five types of Linked Data quality evaluation principles. Zaveri et al. [5] summarized more than a dozen articles related to the quality of linked data, introduced the existing quality assessment methods in detail, and divided quality assessment dimensions into six categories from different aspects of evaluation. According to different systems with different definitions and explanations, Gu et al. [6] summarized the RDF data quality evaluation dimensions and indicators and adjusted them. Systemically speaking, the quality problems of data are important factors that may affect the performance of the system and even determine whether the system can work.

During the vigorous development of the semantic web, the de centralization system explode [7]-[9]. De centralization of phenomenal products, such as the bitcoin [10] proposed by Satoshi Nakamoto, its blockchain technology [11] is a de centralized book.

[*] Corresponding author. E-mail: simon@wust.edu.cn.

At present, the blockchain technology has gradually separated from Bitcoin, and the blockchain as an independent technology is now widely used in many fields [12]. The blockchain [13] is a decentralized technical solution in which nodes have the characteristics of a consensus mechanism. Point-to-point collaboration or transactions can be performed between nodes without trust credentials. As one of the most important innovations of the blockchain, the super-book can realize the traceability of the digital information in the chain and the indispensable modification of the account information.

Many domains use RDF data structure for transaction processing, so it is very important to carry out RDF quality evaluation in different fields. The emergence and development of the blockchains provide new ideas and inspiration for the quality evaluation of RDF datasets. Compared to RDF data quality evaluation in decentralized systems, previous RDF data set-s must publish RDF data sets to the internet for sharing, and their quality evaluation is expensive to maintain and potentially contaminate Internet data. In this paper, using the blockchain storage quality evaluation results, the centralization effect of the authority can be reduced, and the quality evaluation results have characteristics such as non-tampering. By using these features, we can provide users with better cost-effective results when knowledge is independently protected.

2. Design of quality evaluation model

In the decentralized system, the node quality consists of two parts: node service quality and node data quality. Node service quality refers to the ability of a node to effectively provide services. In general, this indicator is affected by the physical factors of the node itself. Node data quality is a measure of the quality of service provided by a node. Because the quality of service of nodes is limited by the physical, the change of this indicator in different systems is small. Therefore, this paper focuses on the node data quality.

2.1. RDF Inspection report

The RDF dataset itself has certain quality problems, such as the number of blank nodes, data redundancy, accessibility of URI (Uniform Resource Identifier) and so on. Therefore, the content of this section is to design and implement a quality reporting model for RDF data, which is used to quantify the quality of RDF data, which is known as the RDF physical examination report. For RDF's own data attribute, this article provides the parameters in Table 1 as the basic parameters:

Redundancy calculation of RDF data sets:

$$Redundancy_{(data1)} = 1 - \frac{DSPO\,(data1)}{SPO\,(data1)} \tag{1}$$

The number of subject average attributes of the RDF dataset:

$$VP_{(data1)} = 1 - \frac{SPO\,(data1)}{DS\ (data1)} \tag{2}$$

The average number of attributes in the RDF dataset indicates the description of a subject in a dataset. The larger the value, the more data sets use the triples to describe the subject, and the increase in the attributes of the subject can make the knowledge more complete. Therefore, the average number of RDF attributes is directly proportional to the quality of the RDF data set.

The accessibility of the URI in the RDF dataset needs to be judged by sampling. The RDF medical report model is given below:

$$\begin{aligned} QRDF_{(data1)} &= k_1 * \left(1 - Redundancy_{(data1)}\right) \\ &+ k_2 * \left(1 - \frac{Blank\,(data1)}{SPO\,(data1)}\right) + k_3 * URI_{(data1)} \\ &+ k_4 * VP_{(data1)} \end{aligned} \tag{3}$$

The *URI* in Eq.(3) is the result of the URI sample visit. That is a certain number of URIs are sampled and accessed. The value is the ratio between the number of accessible and the total number of samples; k_1, k_2, k_3, k_4 are constant greater than 0, and can be adjusted according to different systems. It can be known from the formula that the RDF medical examination report is inversely proportional to the redundancy of the data set and inversely proportional to the number of blank nodes, which is proportional to the accessibility of the data and proportional to the average number of attributes of the subject.

2.2. Verifiability

Verifiability refers to the same query results obtained by multiple identical queries. Frequent changes in the data make it impossible for the user to trust the n-ode's data, but data updates and modifications are nec-

Table 1
RDF data set basic properties

Parameter	Calculation formula	Description
Blank node number	$Blank_{(data1)}$	number of blank nodes in Data1
Number of subjects	$S_{(data1)}$	number of subjects in Data1
The number of unique subjects	$DS_{(data1)}$	number of unique subjects in Data1
Number of predicates	$P_{(data1)}$	number of predicates in Data1
The number of unique predicates	$DP_{(data1)}$	number of unique predicates in Data1
The number of objects	$O_{(data1)}$	number of objects in Data1
The number of unique objects	$DO_{(data1)}$	number of unique objects in Data1
The number of triples	$SPO_{(data1)}$	number of triples in Data1
The number of unique triples	$DSPO_{(data1)}$	number of unique triples in Data1
URI accessibility	$URI_{(data1)}$	URI accessibility in Data1

essary. Therefore verifiability should change as the data set changes. This paper sets the verifiability granularity to the log level, which means that each query will generate logs. For the latest record comparison, the correct rate and error rate are obtained, that is, the same ratio of the comparison results and the ratio of different comparison results. The difference is the verifiable result of the query. The node verifiability is calculated as follows:

$$Verifiability_{(data1)} = \sum_{i=1}^{n}(CorrectRate_{(logi)} - ErrorRate_{(logi)}) \qquad (4)$$

According to the above formula, when the node updates data, its verifiability will be greatly reduced, resulting in the decline of data quality. But with the increase of queries, verifiability will gradually improve. This phenomenon is in line with expectations. Frequent changes make it less verifiable than what the system and users do not want to see.

2.3. Completeness, Relevance and Uniqueness

In order to explain the calculation of the Completeness and other dimensions later, Table 2 gives some symbols and their meanings.

For example, node A does not have entity S_1 in other nodes, and S_1 has a large user demand. For entity S_1, the data set $data1$ in node A contributes a lot to the entire knowledge map. In contrast, part of the entities, all nodes are owned and the contribution of the entity in the node is smaller. The ratio of the attributes in the entity S_1 to the total attributes of the subject in the entire knowledge map is called the completeness of the RDF data in the node. Each subject has its own integrity, and the proportion of the subject to all the subjects is taken into account.

The formula for calculating the Completeness of the entity S1 in the data set data1 is:

$$I_{(data1,S_1)} = \frac{P\,(data1, S_1)}{BP\,(S_1)} \qquad (5)$$

The Eq.(5) shows that the sum of the completeness of all nodes is not 1. Because the attributes of the same subject in different nodes may be duplicated. Correlation degree is the sum of the proportion of repeated attributes calculated by entity S_1 in different nodes. Correlation degree represents the similarity between a data set and other nodes. To a certain extent, it indicates the importance of the node to knowledge map. The calculation formula is as follows:

$$Relationship_{(data1)} = \sum_{i=1}^{n}(\frac{P_{(data1,S_i)}}{BP_{(S_i)}})$$
$$(S_i \in Data1 \cap Eachother) \qquad (6)$$

Relatively, Uniqueness refers to the degree of knowledge that a node has no other nodes, which greatly improves the value of nodes. These subjects or attributes are the uniqueness knowledge of the node, and its calculation formula is shown in Eq. 7.

$$Uniqueness_{(data1)} =$$
$$\sum_{i=1}^{n}(\frac{P_{(data1,S_i)} - PublicP_{(data1,S_i)}}{BP_{(data1,S_i)}})\,(S_i \in Data1) \qquad (7)$$

Table 2

Partial symbols and their significance in model calculation

Symbol	Description
$P_{(data1,S_1)}$	The number of predicates in the Data1 with the subject of S1
$DP_{(data1,S_1)}$	The only number of predicates in Data1
$BP_{(S_1)}$	The number of predicates with the subject of S1 in the entire knowledge map
$PublicP(data, S_1)$	The number of predicates and BP intersections in the subject of S1 in Data1

Based on the uniqueness model, the node data contribution model is proposed. This model refers to the degree of knowledge provided by a node to the whole decentralized network knowledge.

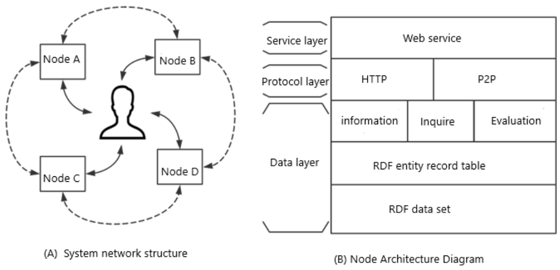
$$Contribution_{(data1)} = \sum_{i=1}^{n}(Uniqueness_{(data1,S_i)})* \left(SC_{(S_i)} - SF_{(S_i)} * k\right) \left(S_i \in Data1, SF_{(S_i)} \in [0,1], k > 0\right) \tag{8}$$

In Eq. (8), $SC_{(s_i)}$represents the query frequency of the entity S_i, and represents the effect of a transaction on the quality of the data, which belongs to one of the user's behavior; $SF_{(S_i)}$ represents the user feedback for the entity S_i, which greatly affects the degree of data contribution to the node, The smaller the item, the more reliable the data. k is the user feedback adjustment factor. According to the formula, if the value of k is 2, if the user feedback is 50 percent or more of the query frequency, a negative contribution will occur.

2.4. Node Data Quality

The data quality of the node is composed of several parts, including the dimensions of the RDF medical examination report, data completeness, verifiability, user feedback, etc. The comprehensive model of data quality evaluation is given here:

$$QD_{(data1)} = k_1 * QRDF_{(data1)} + k_2 * Contribution_{(data1)} + k_3 * Verifiability_{(data1)} (k_1 > 0, k_2 > 0, k_3 > 0) \tag{9}$$



(A) System network structure

(B) Node Architecture Diagram

Fig. 1. DCQE system network structure

In Eq. (9), $QRDF_{(data1)}$ represents the physical examination report of RDF data Data1; $QRDF_{(data1)}$ represents the degree of contribution of data set Data1; $Verifiability_{(data1)}$ represents Validity of data set Data1. The uniqueness (second item) in the formula accounts for a large proportion. The reason is that if the uniqueness knowledge possessed by a node is widely demanded by users (the number of queries is large), it means the indispensability of the knowledge and the higher the data quality of the node.

3. DCQE system design

DCQE communicates through P2P protocol, uses RDF data storage, and forms a multi-node alliance information exchange system. The meaning of alliance information interaction is that each node possesses independent knowledge and interacts with other nodes while the knowledge is protected. The user accesses from any node and the access effect is consistent. At the same time, the system does not have a central node and belongs to a decentralized network structure. The system network structure is shown in Figure 1(A). The logical structure of the node is shown in Figure 1(B).

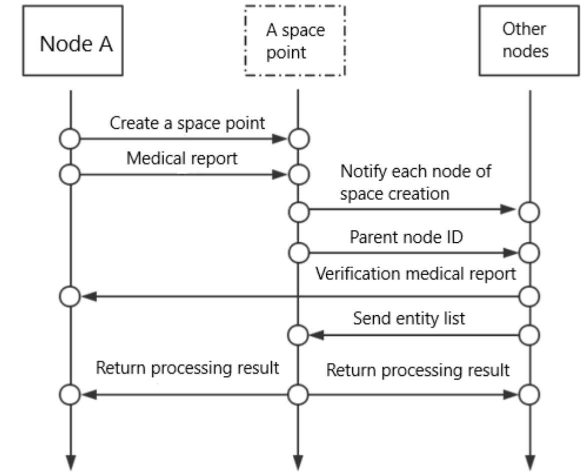


Fig. 2. Blank node verification process

3.1. Node connectivity calculation

For different nodes in a decentralized system, each node has its own knowledge and the knowledge independence within the node is protected. In order to calculate the uniqueness and other dimensions in the data quality evaluation of each node, information needs to be communicated between each node. So here we propose a node-communicative calculation method.

When it is necessary to perform the connection calculation, the initiator establishes a temporary node (hereinafter referred to as an blank node) to join the decentralized system. Blank nodes have independent flag bits indicating that they are temporarily creating nodes. This node provides services that can verify its identity. Figure 2 shows the blank node's authentication method:

In order to ensure that the knowledge of each node is independently protected, each node will not output complete triplet information to blank nodes, but rather a combination of <subject-predicate>. Finally, after the blank node calculates the task result, each result is returned to the corresponding node, and each node adds the result to the calculation of the quality evaluation.

With regard to security, after the destruction of an blank node, two things need to be ensured: the blank node no longer exists; the parent node A who created the blank node does not steal information from other nodes. For the first point, each node accesses the destructor information of an blank node, and accesses it, and updates the routing information if it is unreachable. For the second point, using the query log in section 4.2, each node makes its own query for the RDF data of node A, and node A provides a service that only calculates the hash of the query result and does not return the result. When the blank node is destructed, the node A is queried to obtain the hash value, and when the blank node is created, the hash is checked and verified. The current and late comparisons indicate that the content of the node has not changed; if the comparison results are inconsistent, the node may illegally steal other node information. If the hash results are inconsistent, node A will be removed to decentralize the network. Since each node provides different queries, the more nodes there are, the less likely it is that node A steals.

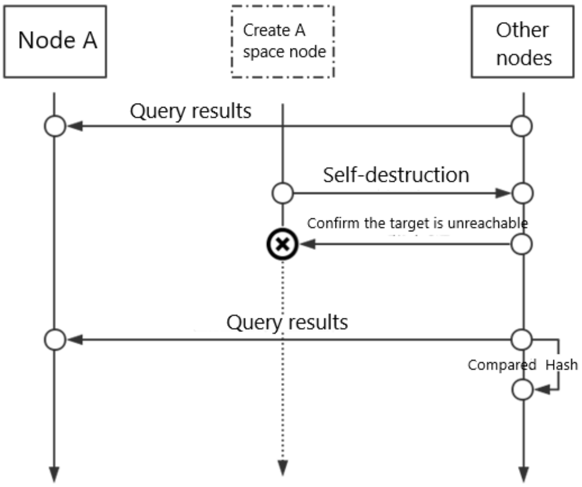


Fig. 3. Blank node destruction process

3.2. Advantages of Using Blockchain Record Quality Evaluation Results

In the previous studies, the quality evaluation was that the authoritative center issued the quality certificate to guarantee the quality of the RDF data set. The center did not exist in the decentralized system. The disadvantage of this is the need for authoritative central agencies. When the authoritative center has problems or the degree of trust is declining, it is difficult to ensure that the quality evaluation results are true. Blockchain technology solves the trust problem in the decentralized network structure. Therefore, blockchain is used to record transaction information and quality evaluation result information. There are two main conditions for using blockchain Technology: 1, the attached system is decentralized system; 2, there is consensus mechanism. This system belongs to a centralization. The consensus mechanism is described in detail later.

The advantages of using blockchain to record quality are as follows:

- Structural security. Quality evaluation certificate is not issued without authority center. Nodes are mutually authenticated, and each node has backup of other nodes' quality evaluation. The quality evaluation result that prevents central node collapse or being attacked is not credible.
- Information security. Prevent the node from tampering the quality evaluation results. The use of consensus mechanism prevents Byzantine attacks. The whole network node maintains the super account of the quality evaluation results. If malicious nodes want to falsification of quality information, they can be effectively screened through the whole network. Malicious nodes need to control more than 50 percent nodes to launch forgery attacks, but the cost is very high.
- Support the quality result update mechanism. Previously, RDF data needs to be recertification and published after each update. Using blockchain to record can be updated in the node itself, and then the update log is written to the block and recorded in the quality evaluation book.
- Update log track can be found. Each update of the update record has a basis. The quality evaluation can be reviewed from the first generation to the last change.

Due to the above advantages, it has become possible to achieve record and dynamic update quality evaluation in the decentralized system. The result of quality evaluation is not only to make users feel comfortable using RDF data sets, but also to guide the system to provide users with more cost-effective query results and to maintain system operation, which becomes an important factor affecting the operation of the system.

3.3. Using Blockchain Record Quality Evaluation Results

Because the nodes are dynamically added to the system, this paper also uses incremental builds. When the new node joins the system, it evaluates the quality of the node and synchronizes it to the quality evaluation books of each node. This section mainly introduces the specific process of generating and synchronizing quality evaluation results when a new node joins the decentralized system.

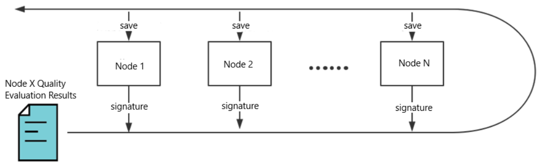


Fig. 4. Node quality evaluation record generation process

First of all, when a new node joins a decentralized system, quality assessment is required. Here is the result of the quality evaluation of new node 1 is generated.

- The new node enters the centralization system, synchronizing the quality evaluation account.
- New blank node and notifies other nodes.
- Provide RDF data physical examination report to the blank node.
- Each node obtains the RDF physical examination report Hash value of the new node for later comparison.
- The Hash value of the parent node's physical examination report and the Hash value provided by the parent node compared with the blank nodes calculated by each node to verify the legality of the blank node.
- The combination of each node's subject predicate combination.
- Blank nodes compute the completeness, relevance, uniqueness and other related attributes of each node and return them to each node.
- Each node generates a RDF entity record table.
- The result of the calculation of the quality evaluation.
- The quality evaluation results are broadcast to other nodes for signature.
- Write the initial record of the quality evaluation hash into Merkel tree.
- Empty block self-destruction
- Each node verifies whether the blank node is fully configured and updated.

According to the above process, the initial quality evaluation of the new node in the de centralization system is completed. The quality evaluation result is the first item in the update log of the quality evaluation result, and the renewal of the quality evaluation result is based on the record.

Figure 4 shows the process of node quality evaluation record generation. When a node generates a quality evaluation result, it broadcasts to other nodes for signature. The signature here is the chain signature. That is, in order to join the system, each node will sign and return to the blank node. Blank nodes are distributed to each node for recording. After obtaining all node signatures, the first quality evaluation record is generated.

The second condition of using blockchain: consensus mechanism. Users need to spend tokens for each query. Here, k Percent ($k \in [0, 100]$) of all transactions in each T-hour will be deducted. This will be used to pay the billing node for account records, and workload verification will be used. The proof of work is consistent with the amount of work in mining in Bitcoin. The detailed design can be found in literature [9].

The storage and recording of transaction records in the blockchain can form an account book so that the transaction information cannot be forged and cannot be modified. Similarly, storing quality evaluation results in the same way, each update has the same effect. The following is an update of the quality evaluation results. Each update is based on the results of the last quality assessment. The update log structure includes Block ID, Update, Signature and Timestamp.

Among them, an update term refers to a dimension that is reduced or being promoted in an update. The space occupied by the record update item occupies more space than the record-only increment result, but if the record quality evaluation result is incremented, the update of the quality evaluation result cannot be completely represented. Therefore, the results are updated, and each block calculates its own new quality evaluation result. The signature refers to the signature information of each block, indicating that each block knows the existence of the record, and the signature item is a list. The block ID is used to illustrate blocks that update the quality evaluation results. According to the log structure above, the time tree can be formed according to the timestamp, and each update aims at what items are updated for the results of the last quality evaluation, and the relationship chain between the updates is realized. It can effectively prevent the renewal and deception of quality evaluation.

As shown in Figure 5. After the update log is generated, it is released to each block for signature. After all signatures, the results are recorded. Merkel tree storage quality evaluation results are used here to effectively prevent the results from falsifying. Any tampering will cause the hash value stored in the root node to change. When it is inconsistent with other nodes, it is considered that the quality evaluation result in the block is not trustworthy.

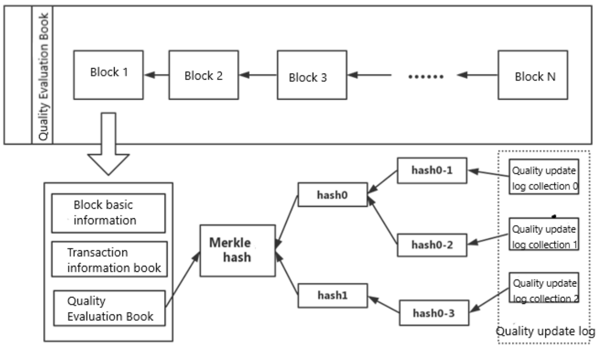


Fig. 5. Account schematic of quality evaluation

4. Experiment Analysis

Experimental environment: MacBook Pro computer model, 2.6 GHz Intel Core i7 processor, 16GB 2133 MHz LPDDR3 memory. In this environment simulation, a decentralized system that implements six nodes is built. The agreement includes the Http protocol and the P2P protocol. The experimental data sets use the ArchiveHub data set. The size of the data set file is 71.8M, the number of entities is 106919, the number of unique subjects is 51,411, the number of unique predicates is 141, the number of unique objects is 104408, and the number of triples is 431088. In order to highlight the importance of the quality assessment dimensions such as verifiability, completeness, and uniqueness in the model, this paper divides the data set into 6 copies and records it as AH1 to AH6. Table 3 is the basic information of these 6 data sets. To increase the completeness judgment, this paper selectively replicates some node information in AH1 AH6.

The RDF physical examination report in Table 3 is calculated according to Eq.(3), where $k_1 = 1, k_2 = 1, k_3 = 1, k_4 = 1$.

As the system does not activate verifiable and monopolistic in the model (Eq. 9) before accepting the query, the result of the RDF medical examination report is the initial quality evaluation result of the node.

4.1. Verification of Quality Evaluation Model

4.1.1. Verifiable model validation

Firstly, verify the verifiability of the model. The activation process is to make 100 queries after the decentralized system is set up. The calculation updates the new quality assessment results for each node.

Statement1:select* { ?s?p?o } where { <http://api.talis.com/stores/locah/items/1305283343810#self>?p?o. }

Table 3

Basic information of the experimental data set

Data set	The number of entities	The number of triples	The number of unique masters	RDF medical report	
AH1	5.7	16748	28361	6945/89	4.05
AH2	6.7	20514	39705	9271/46	4.28
AH3	9.9	20876	56722	2676/43	21.1
AH4	15.1	28366	79410	4576/66	17.35
AH5	16.3	40803	102099	12983/29	7.86
AH6	19.3	43590	124796	14970/101	8.34

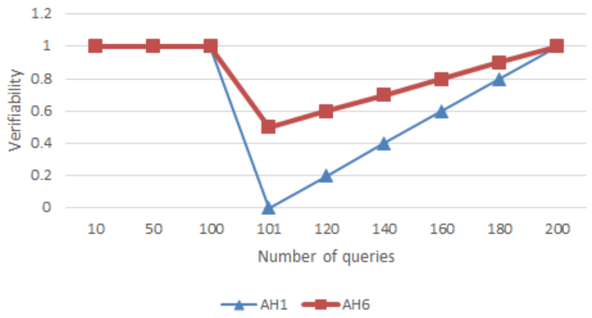


Fig. 6. Verifiability impact on quality assessment

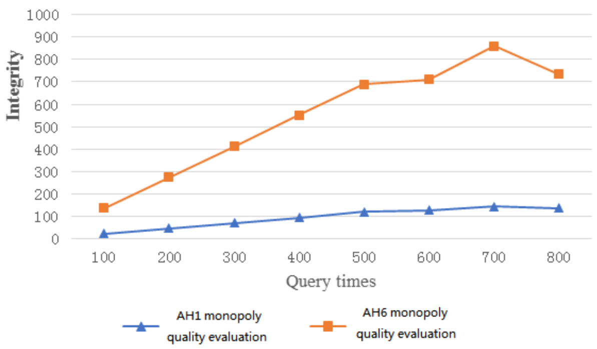


Fig. 7. AH1, AH6 verification uniqueness model

Hit result: AH1, AH2, AH6

Statement2:select*｛?s?p?o where ｛< http://data.archiveshub.ac.uk/id/perso-n/aacr 2/martindorothyfree-born> ?p ?o . ｝

Hit result: AH6

After completing the update of the quality evaluation result, the entity of the statement 1 hit in the data set is updated after the 100th query. Then execute statement 1. Record the query results as follows:

According to the records in Figure 6, the quality of AH1 and AH6 remains stable for the first 100 queries. The reason is that the data set itself has not changed, and the consistency of query results is stable and consistent. At the 101st time, verifiable landslides occurred, resulting in a sharp decline in quality assessment results. The reason is that the data in AH1 and AH6 have been updated, resulting in inconsistent query results and previous query results. For users, frequent changes in data are undesired, so a large-scale decline in quality evaluation is in line with expectations. At the same time, with the increase of the number of queries, the quality evaluation has steadily increased linearly. If data updates occur during this process, data quality will still be degraded. Compared the quality evaluation results between AH1 and AH6, the reason why AH6 quality degradation is lower than AH1 after 100 queries is that the statement 2 was executed during the activation model process, resulting in the AH6 query log contains two parts, and the update data does not affect the statement 5.2 Hit results. Therefore, the quality of AH6 declines less. By comparing AH6 and AH1, the verifiable model granularity was verified as log level.

4.1.2. *Impact of data completeness and user feedback on quality assessment*

After the system is activated, each node generates its own RDF entity record table. As the query increases, the quality of the node data gradually increases. Table 4 is an RDF entity record table that is generated after the model is activated.

In Table 4, Self representative entity<http://api.talis.com/stores/locah/item-s/1305283343810#self>.Marti-ndorothy freeborn representative entity < http://data.archiveshub.ac.uk/id/perso-n/aacr2/martindorothyfre-eborn> Since the entity self exists in other data sets, the uniqueness coefficients of AH1 and AH6 are lower. This paper executes statements 1 and 2 multiple times and randomly adds user feedback after 500 executions. Fig. 7 shows the quality change in AH1, AH6 during this process.

In Fig. 7, the quality of AH1 and AH6 has steadily increased due to the increasing frequency of queries. Among them, AH6 has grown rapidly because it has

Table 4
Record Table for RDF Entities in AH1 and AH6

Data set	Entity	Uniqueness coefficient	Query frequency
AH1	self	0.24	100
AH6	Martindorothy freeborn	1	100
AH6	self	0.378	100

multiple subjects. After 500 inquiries, less than 50 percent of users added feedback to interfere, and the increase in quality became slow. After 700 inquiries, the interference was adjusted to more than 50 percent, and the quality was reduced. So we can see that the quality is related with user feedback .Different systems can adjust the user behavior through this parameter. Figure 7 shows that the increase and decrease of AH6 is significantly higher than that of AH1. The reasons are as follows: 1. AH6 involves more subjects; 2. AH6 has a subject with higher uniqueness coefficient.

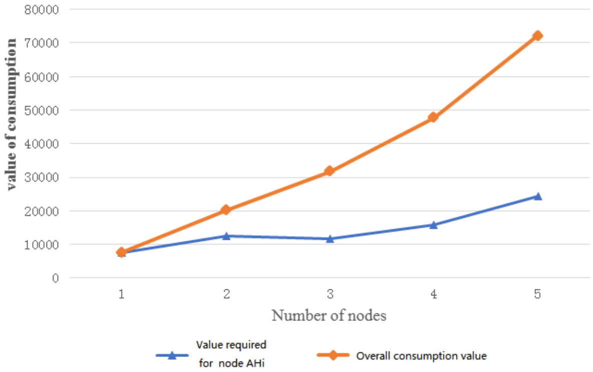


Fig. 8. Relationship between node number and cost value

4.2. *Quality Evaluation Book Safety*

Although the blockchain solves the problem of trust, 51 percent of attacks can still be launched. This paper calculates the difficulty of controlling each node and analyzes which nodes will launch attacks from the perspective of interests. If the nodes A, B, C, and D contain the subject s1, the attributes are 9, 5, 31, and 7 respectively. There are 5 common attributes. For each query, the initial contribution of A is 0.105, 0, 0.763, and 0.05. Assume that the prices of A, B, C, and D are the same. The cost for node B to launch an attack is:

$$Cost_{(B)} \geqslant \sum_{i=0}^{n}\sum_{j=1}^{m} price_{(node_i)} \\ *Uniqueness_{(data1,S_j)} \qquad (10)$$

$node_i$ is the number of nodes that you want to bribe. The Fig.8 shows the relationship between the number of nodes and the cost value.

5. Conclusion

This paper discusses how to evaluate and update RDF data in the context of the rapid development of Semantic Web and decentralized systems. First of all, it shows the significance of the research and the importance of the RDF data quality evaluation in the decentralized system. Then, it pointed out that the RDF data quality is calculated by the RDF medical examination report, credibility, and completeness dimensions, which can reflect the quality of the RDF data service provided by the node and the contribution of the node to the entire decentralized system. Finally, according to the node quality evaluation model, this paper designs and implements the decentralized quality assessment system DCQE. Descrip the use of the system construction process, user behavior impact on quality evaluation, and quality evaluation results in detail. For quality evaluation results, this paper proposes and discusses the use of blockchain storage quality evaluation results. The blockchain techniques and quality assessment results have the following advantages:

- No authoritative center is required to issue a quality evaluation certificate. The nodes authenticate each other, and each node has backups of other node quality evaluations.
- Prevent nodes from falsifying quality evaluation results. In accordance with the Byzantine attack model, only nodes controlling more than 50 percent of nodes can perform forged results, but this is expensive.
- Support quality result update mechanism. Every update has a basis, and the quality evaluation can refer to the change log from the first time to the last time.

The model presented in this paper still has areas for improvement and research:

- The model does not consider the price factor. Because the price factor is different in different systems, but as a core part of cost-effectiveness, each system can propose and implement more complex models to achieve a model solution more in line with the real environment by providing quality evaluation on the system.
- In this paper, blockchain technology is used to solve the consensus problem in the decentralized system, but multiple ledgers and multiple sets of protocols are stored and this can be considered to be integrated into one account. At the same time, you can consider using consensus mechanisms that are more consistent with quality assessment.
- Moreover, in addition to the completeness and other dimensions, it is also one of the future tasks to look for new dimensions of the quality evaluation of other decentralized network datasets and improve the de-centralized system RDF data quality assessment mechanism.

Acknowledgements

This work has been supported by the National Natural Science Foundation of China (61602350) and the Key Projects of National Social Science Foundation of China(11&ZD189). Thanks the graduate students Yansong Wang and Hui Ren for their excellent work in this paper.

References

[1] D.T Heath, How to publish linked data on the web,*Proc Iswc*, 2009.

[2] A. Hogan, J.Umbrich and A.Harth,et al, An empirical survey of linked data conformance,*Agents on the World Wide Web* 14(3) 2012, pp.14–44.

[3] J.Kandari , An empirical survey of linked data conformance, 2016.

[4] A. Assaf and A. Senart, Data quality principles in the semantic web, *2012 IEEE Sixth International Conference on Semantic Computing*, 2012, pp. 226–229.

[5] D. Wilson and G. Ateniese, From pretty good to great: enhancing PGP using Bitcoin and the blockchain, *Proceedings of the 9th International Conference on Network and Sys- tem Security*, 2015, pp.368-375.

[6] J. Gu, T.Zhu and H.Li,et al,Overview of link data quality assessment in knowledge mapping ,*Journal of Wuhan University* 63(1) 2017, pp.22–38.

[7] Y. Jia, B. Wang ,Centralization of secure distributed storage system ,*Computer Engineering* 38(3) 2012, pp.126–129.

[8] T. Yang, X.liu and S.Liu,To centralization finance and block chain ,*Financial Expo: fortune* (12) 2016, pp.18–19.

[9] D. Fan, Research and design of website architecture de centralization on cloud computing platform ,*Shanghai Jiao Tong University* 2013.

[10] S. Nakamoto, Bitcoin: A peer-to-peer electronic cash system ,*Consulted* 2008.

[11] M. Swan, Blockchain: Blueprint for a New Economy ,*O'Reilly Media, Inc.* 2015.

[12] Y. zhang,The impact of block chain technology on the development of China's financial industry ,*International finance* (5) 2016, pp.41–45.

[13] R. Beck, J. Czepluch, N. Lollike,et al. BLOCKCHAIN ÅąC THE GATEWAY TO TRUST-FREE CRYPTOGRAPHIC TRANSACTIONS ,*Twenty-Fourth European Conference on Information Systems* 2016.