An RDF Data Set Quality Assessment Mechanism for Decentralized Systems

Li Huang¹,², Zhenzhen Liu¹,², Fangfang Xu¹,² & Jinguang Gu¹,²,³†

¹Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, Wuhan 430065, China
²College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China
³Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content, State Administration of Radio and Television, Beijing 100038, China

Keywords: Decentralization; Quality assessment; Blockchain; RDF data set

Citation: L. Huang, Z. Liu, F. Xu & J. Gu. An RDF data set quality assessment mechanism for decentralized systems. Data Intelligence 2(2020), 529–553. doi: 10.1162/dint_a_00059
Received: December 27, 2018; Revised: October 25, 2019; Accepted: March 20, 2020

ABSTRACT

With the rapid growth of the linked data on the Web, the quality assessment of the RDF data set becomes particularly important, especially for the quality and accessibility of the linked data. In most cases, RDF data sets are shared online, leading to a high maintenance cost for the quality assessment. This also potentially pollutes Internet data. Recently blockchain technology has shown the potential in many applications. Using the blockchain storage quality assessment results can reduce the centralization of the authority, and the quality assessment results have characteristics such as non-tampering. To this end, we propose an RDF data quality assessment model in a decentralized environment, pointing out a new dimension of RDF data quality. We use the blockchain to record the data quality assessment results and design a detailed update strategy for the quality assessment results. We have implemented a system DCQA to test and verify the feasibility of the quality assessment model. The proposed method can provide users with better cost-effective results when knowledge is independently protected.

1. INTRODUCTION

The current decentralized system has shown explosive growth. A decentralized network structure is a network structure that grants permissions to member nodes without a central authority. Its structural

† Corresponding author: Jinguang Gu (E-mail: simon@wust.edu.cn, ORCID: 0000-0002-8823-8480).

© 2020 Chinese Academy of Sciences. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.
characteristics are: 1) it has many nodes; 2) there is no hierarchical relationship between nodes; 3) each node can run independently without interference from other nodes; and 4) nodes are interconnected with each other. Compared with the centralized structure, the decentralized structure reduces the dependence on the central node and enhances the security and robustness of the entire structure. Decentralized phenomenon-level products such as Bitcoin [1] proposed by Satoshi Nakamoto, and the blockchain technology [2] used by it are a decentralized ledger. At present, the blockchain technology is gradually separated from Bitcoin. As an independent technology, the blockchain is now widely used in many fields [3], and has been proven to perform well in the fields of finance, supply chain and insurance.

The third-generation Internet core content semantic Web technology [4] redefines the way of resource description, and Resource Description Framework (RDF) [5, 6] came into being. The purpose is to describe the resources on the network, and fundamentally solve the problems of data integration difficulties, machine understanding difficulties, etc. caused by the different forms of data description on the Internet. With the continuous increase of decentralized systems using RDF data sets, the quality of the original RDF data sets has also been upgraded and is an important factor that affects the performance of the system. There are many fields that use RDF data structures for transaction processing. For example, in the field of medicine, various hospitals provide RDF data to form an overall pharmaceutical knowledge map for the convenience of users or physicians. Different hospitals may have the same drug information or attributes, and some hospitals have unique information or attributes. How to obtain more cost-effective and accurate query results when users search for drug information becomes the key to the problem. RDF quality assessment in different fields is thus very important. The emergence and development of the blockchain has provided new ideas and inspiration for the issue of RDF data quality assessment. This paper uses the blockchain to store quality assessment results, which can reduce the centralization of authoritative institutions. At the same time, the quality assessment results have characteristics such as immutability. In a decentralized system, RDF data quality assessment adds multiple new dimensions such as completeness and uniqueness. We propose an RDF data set quality assessment mechanism for decentralized systems, and discuss how to realize the partial dimension of RDF data quality assessment under the condition that the knowledge of each node in the blockchain is independently protected. In particular, we have implemented a system to test and verify the feasibility of the quality assessment model and named it DCQA.

2. RELATED WORK

The quality assessment of RDF data sets has received widespread attention. Many researchers have conducted in-depth discussions on the assessment of RDF quality [7,8,9]. Many studies on the quality assessment of RDF data sets have started from different perspectives, and summarized the multiple dimensions of RDF in the production, use and maintenance phases. In addition, with the rapid development of RDF, new dimensions are constantly emerging. Assaf and Senart [10] summarized five types of linked data quality assessment principles: quality of data source, quality of raw data, quality of the semantic conversion, quality of the linking process and global quality. These five principles from the perspective of the data management process, describe the content of the quality assessment work at each stage. The
principles of each stage include several assessment standards, which is the dimension of quality assessment in this paper. Zaveri et al. [11] summarized more than ten linked data quality assessment related papers, introduced the existing quality assessment methods in detail, and divided the quality assessment dimensions into six categories from different assessment aspects. At the same time, they put forward a unified description of terms, dimensions and measurement indicators in terms of data quality. Taking flight information as an example, they are closely combined with quality assessment in the text to reflect the feasibility of the scheme. Bizer and Cyganiak [12] divided the quality dimension into three categories according to the type of information used: content-based, which refers to the content of the information itself; context-based, which refers to the contextual information at the place where the information is declared; and level-based, which refers to the level of the data itself or the information provider. With the development of RDF data quality assessment studies, research on decentralized systems using blockchain technology is increasing. Governments around the world have pursued policies to promote the landing and development of blockchain. For example, the Canadian government intends to use the Ethereum blockchain to track and record government donation information; the British government released a special report on blockchain applications in government work and finance [13]. The Indian government has launched a blockchain ecosystem in cooperation with fund company Covalentfund. Blockchain technology research is already among one of the hot research fields.

In a decentralized system, the RDF knowledge of each node is independent and protected. Therefore, the quality assessment of RDF data is also very different from centralized distributed systems, and many new dimensions and updated methods have appeared in recent years. In the quality assessment of RDF data combined with the blockchain, we should not only pay attention to the data quality of the RDF data set itself, but also pay attention to the quality issues caused by the interaction between RDF data sets. The previous studies on quality assessment are based on a single data set. In response to these problems, this paper proposes an RDF data set quality assessment mechanism for decentralized systems, which aims to provide users with better services in a decentralized system.

3. DESIGN OF QUALITY ASSESSMENT 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. Since the quality of service of nodes is limited by the physical nature, the indicator does not change much. Therefore, this paper focuses on the node data quality.

3.1 RDF Inspection Report

An RDF data set has certain quality factors including the number of blank nodes, data redundancy and accessibility of Uniform Resource Identifier (URI). Combining the above three RDF data quality issues and the number of average subject attributes of the RDF data set, in this section we designed and implemented...
an inspection report model to quantify the quality of the RDF data. For RDF data attribute symbols used in this section, please refer to Table 1.

Table 1. RDF data set basic attributes.

| Attribute description              | Symbol   |
|-----------------------------------|----------|
| Blank node number                 | Blank\(_{\text{data}}\) |
| Number of subjects                | \(S_{\text{data}}\) |
| The number of unique subjects     | \(US_{\text{data}}\) |
| Number of predicates              | \(P_{\text{data}}\) |
| The number of unique predicates   | \(UP_{\text{data}}\) |
| The number of objects             | \(O_{\text{data}}\) |
| The number of unique objects      | \(UO_{\text{data}}\) |
| The number of triples             | \(SPO_{\text{data}}\) |
| The number of unique triples      | \(USPO_{\text{data}}\) |
| URI accessibility                 | \(URI_{\text{data}}\) |

RDF data set redundancy calculation equation:

\[
Redundancy_{\text{(data)}} = 1 - \frac{USPO_{\text{(data)}}}{SPO_{\text{(data)}}}\tag{1}
\]

Equation for calculating the average number of subjects in an RDF data set:

\[
VP_{\text{(data)}} = \frac{SPO_{\text{(data)}}}{US_{\text{(data)}}}\tag{2}
\]

The average number of attributes in an RDF data set indicates the description of a subject in a data set, and the larger value indicates that the data set uses more triples to describe the subject. The increase of the subject’s attribute can make the knowledge more complete. So the number of RDF average properties is proportional to the quality of the RDF data set.

The RDF inspection report model is given below:

\[
QRDF_{\text{(data)}} = k_1 \times \left(1 - Redundancy_{\text{(data)}}\right)
+ k_2 \times \left(1 - \frac{Blank_{\text{(data)}}}{SPO_{\text{(data)}}}\right)
+ k_3 \times URI_{\text{(data)}}
+ k_4 \times VP_{\text{(data)}} \tag{3}
\]

Sample access to the URIs in the RDF data set, and then \(URI_{\text{(data)}}\) in Equation (3) is the ratio between the number of accessible URIs and the total number of random checks. \(k_1, k_2, k_3\) and \(k_4\) are positive weight coefficients which can be adjusted according to different systems.
3.2 Verifiability

Verifiability refers to the same query results obtained by multiple identical queries. Frequent changes in data can make users unable to trust the data of this node, but updates and modifications of data are necessary. Therefore verifiability is various as the data set changes. This paper sets the verifiability granularity to the log level, which means that each query will generate logs. The latest record is compared with the previous record result to obtain the accuracy and error rate, and the difference value is the verifiability result of the query. Node verifiability is calculated as follows:

\[
Verifiability_{\text{data}} = \sum_{i=1}^{n} (\text{CorrectRate}_{\text{logi}} - \text{ErrorRate}_{\text{logi}}) 
\]

(4)

According to Equation (4), 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.

The granularity of verifiability is log. Although the granularity is not refined to the entity, the effect is more practical. For example, the entity in node A has five attributes, and new attributes need to be added. Due to the low user demand, the change of the entity has little effect on the entire knowledge graph, and its verifiability fluctuations are not large. If the entity granularity level is adopted, it consumes more space, but the effect is not obvious. We design the log of each query. The format is shown in Figure 1.

![Figure 1. Query log structure.](image)

The query hash is used to record the hash of each query result for comparison with the next hash value. The transaction ID corresponds to the query transaction ID and is used to calculate transaction information. The last part of the query log is the time of a transaction. For example, a log list of node A is shown in Table 2.

| Transaction ID | Query hash | Query result hash | Query delay |
|----------------|------------|-------------------|-------------|
| 1              | MrEzFx8ECd5R1-BoJLNjhWd4GcUNGv | 200ms |
| 2              | MrEzFx8ECd5R1-BojLNjhWd4GcUNGv | 200ms |
| 3              | MrEzFx8ECd5R1-njbPpd2WP6i78GcG | 300ms |
| 4              | edy8vvAmltbgD2XK-s9Djcr0jZhWI14C6 | 600ms |
| 5              | MrEzFx8ECd5R1-njbPpd2WP6i78GcG | 350ms |
| 6              | edy8vvAmltbgD2XK-s9Djcr0jZhWI14C6 | 600ms |
An RDF Data Set Quality Assessment Mechanism for Decentralized Systems

The verifiability of node $A$ is 1. The results of log 2 and log 3 in the log list are different, which means a low verifiability. This leads to a decline in data quality. The verifiability of a stable node is $Dlog_{data}$, which is the number of unique query hash. If the number is smaller than this value, the verifiability of the node decreases (see Equation (5)).

$$
Verifiability_{(A)} = CorrectRate_{(\text{MdEzfxxw}8\text{Ec}d5\text{Sr}1-\text{r})} - ErrorRate_{(\text{MdEzfxxw}8\text{Ec}d5\text{Sr}1-\text{r})} + \\
CorrectRate_{(\text{edy}8\text{vv}Am\text{tbg}D2\text{XX})} - ErrorRate_{(\text{edy}8\text{vv}Am\text{tbg}D2\text{XX})} \\
= 2/4 - 2/4 + 2/2 - 0 = 1
$$

3.3 Completeness, Relevance and Uniqueness

In order to explain the calculation of the completeness and other dimensions later, Table 3 gives a list of symbols and their meanings.

| Symbol               | Description                                                                 |
|----------------------|-----------------------------------------------------------------------------|
| $P_{(data,S_i)}$     | The number of predicates in the Data with the subject of $S_i$              |
| $UP_{(data,S_i)}$    | The number of unique predicates in Data                                      |
| $BP(S_i)$            | The number of predicates with the subject of $S_i$ in the entire knowledge graph |
| $PublicP_{(data,S_i)}$ | The number of predicates and BP intersections in the subject of $S_i$ in Data |

Completeness, relevance, and uniqueness are new quality assessment indicators to ensure that the knowledge of each node in the decentralized system has its own intellectual property rights. Each node owns a part of the entire system knowledge graph, so the proportion of each part in all the attributes of the entire knowledge graph affects the degree to which the node contributes to the entire knowledge graph.

If node $A$ has an entity $S_i$ that does not exist in other nodes, and the user has a greater demand for $S_i$. For entity $S_i$, the data set in node $A$ contributes a lot to the entire knowledge graph. In contrast, if each node owns an entity, the contribution of that entity in that node is small.

The equation for calculating the completeness of entity $S_i$ in the data set data is:

$$
l_{(data,S_i)} = \frac{P_{(data,S_i)}}{BP(S_i)}
$$

According to Equation (6), we know that the sum of the integrity of all nodes is not 1, because the attributes of the same subject in different nodes may be duplicated. The correlation is the sum of the proportion of the repeated attributes calculated by entity $S_i$ in different nodes, which represents the similarity between the node data sets. Its calculation equation is:
Relationship_{(data)} = \sum_{i=1}^{n} \frac{P_{(data,S_i)}}{BP_{(S_i)}}(S_i \in Data \cap Eachothernode)

(7)

In contrast, uniqueness refers to the degree to which a node has knowledge that other nodes do not have. This attribute will greatly increase the value of the node. The calculation equation is:

Uniqueness_{(data)} = \sum_{i=1}^{n} \left( P_{(data,S_i)} - \frac{PublicP_{(data,S_i)}}{BP_{(S_i)}} \right)(S_i \in Data)

(8)

Based on the uniqueness model, the node data contribution model is proposed. The model refers to the degree to which a node provides knowledge relative to the entire decentralized network.

Contribution_{(data)} = \sum_{i=1}^{n} \left( Uniqueness_{(data,S_i)} \times \left( SC_{(S_i)} - SF_{(S_i)} \times k \right) \right)

\left( S_i \in Data, SF_{(S_i)} \in [0,1], k > 0 \right)

(9)

In Equation (9), $SC_{(S_i)}$ represents the query frequency of the entity $S_i$, and it also represents the impact of a transaction on data quality, which is one of user behaviors; $SF_{(S_i)}$ represents user feedback for the entity, which will greatly affect the degree of node data contribution. The smaller the item is, the more reliable the data are, and $k$ is the user feedback adjustment coefficient.

In order to compute and update these dimensions, an RDF data set entity record table is presented. This table is used to record uniqueness coefficient, query frequency, user feedback and other information of the entity. Its structure is shown in Figure 2

![Figure 2. RDF data set entities record table structure.](image)

Uniqueness coefficient represents the degree of uniqueness of an entity in the entire knowledge graph. Query frequency refers to the number of times that the entity is retrieved from the query result, which reflects the impact of transaction information on quality assessment. User feedback refers to the user’s comprehensive assessment of the availability, correctness and other dimensions of the entity, with the default value of 0. Query frequency and user feedback are the embodiment of the key role of user behavior in quality assessment. Contribution_{(data)} can be calculated through the log indicators.

For example, node A and node B have the triples in Table 4, and the method of uniqueness influence on nodes is given here. First, calculate the uniqueness coefficient of each subject, and generate an RDF data set entity record table (the default value of the query frequency is 1 and the default value of user
feedback is 0). Then receive the user query and feedback to update the RDF data set entity record table. Finally, the calculation of the uniqueness model is introduced.

Table 4. Entities in node A and node B.

| Node ID | Triple       | Node ID | Triple       |
|---------|-------------|---------|-------------|
| A       | <s1,p1,o1>  | A       | <s1,p2,o2>  |
| A       | <s1,p3,o3>  | B       | <s2,p1,o2>  |
| A       | <s3,p2,o2>  | B       | <s1,p3,o5>  |

For the subject \( s_1 \) in node \( A \), a calculation example of entity uniqueness is given here:

\[
Uniqueness_{A,s_1} = \frac{P_{(A,s_1)} - PublicP_{(A,s_1)}}{BP_{(s_1)}} = \frac{3 - 0}{4} = 75\%
\] (10)

The received enquiries and feedback: Query statement: select? S where ? S S P o2.

Query results: <s1,p2,o2>, <s2,p1,o2>, <s3,p2,o2> User feedback: s3 results are unreliable and update the RDF data set entity record table to get the results shown in Figure 3 (B).

![Figure 3. RDF data set entities record the table transformation process.](image-url)
The influence of the uniqueness, transaction information and user feedback on the whole is calculated as follows:

\[ \text{Uniqueness}_{(\text{data}_n)} \times (SC_{(\text{data}_n)} - SF_{(\text{data}_n)}) = 1.0 \times 2 \times \frac{2 - 1 \times k}{2} = 0 \quad (k = 2) \] (11)

The effect of user behavior is increased by adjusting the value of \( k \) in Equation (9). The \( k \) value is taken as 2 to calculate the overall impact model of uniqueness on node \( A \): Contribution\(_{(\text{data})} = 3.5 \).

### 3.4 User Behavior

In data quality assessment, user behavior is indispensable, and only the quality assessment combined with user behavior can be meaningful. User behavior can dynamically and comprehensively adjust the quality assessment of data from the outside world, which is a common method to make up for the inappropriateness of model computing quality assessment under different scenarios and conditions. User behavior can affect the accuracy, credibility and other dimensions of data quality assessment. Furthermore, with the combination of blockchain, every transaction belongs to a user behavior and becomes an important part of the assessment of data quality.

For data correctness, this dimension can be influenced by user feedback. Although we have the entire data set, we cannot judge whether a triple is true or false. User feedback can be used to determine whether the triple is available or not. The user's feedback on the right and wrong triples in a query will greatly affect the quality of the entire data set.

The transaction information adds value to the entities in the query results. Assuming that node \( A \) has entity \( s_1 \), which is not found in other nodes, and that the entity query frequency is high, the node will contribute more to the entire knowledge graph. Similarly, if node \( A \) and other nodes not only have the same entity \( s_2 \), but also have the same entity attribute, then entity \( s_2 \) in node \( A \) contributes less to this knowledge graph. In this paper, the query distribution in a query is combined with subject relevance, uniqueness and the final overall impact of data quality assessment results.

User behavior consists of two parts in this system: user query (that is, a transaction) and user feedback. These two parts will affect the results of quality assessment. Different from the results of initial calculation node quality assessment, the initial calculation is static calculation, and the user behavior impact is dynamic impact. This is because user behavior triggers much faster than quality updates. Therefore, this paper proposes query logs and RDF entity record tables to cache the results of user actions.

### 3.5 Node Data Quality

The data quality of the node is composed of several parts, including the dimensions of the RDF inspection report, data completeness, verifiability, user feedback, etc. The comprehensive model of data quality assessment is given here:
\[ QD_{(data)} = k_1 \times QRDF_{(data)} + k_2 \times \text{Contribution}_{(data)} + k_3 \times \text{Verifiability}_{(data)} \]

In Equation (12), \( QRDF_{(data)} \) represents the inspection report of Data; \( QRDF_{(data)} \) represents the degree of contribution of the data set Data; \( \text{Verifiability}_{(data)} \) represents validity of the data set Data. The second item in the equation accounts for a large proportion. The reason is that if the unique knowledge possessed by a node is widely demanded by users, it means the indispensability of the knowledge and a higher data quality of the node.

4. DCQA SYSTEM DESIGN

DCQA is an alliance information interaction system based on P2P protocol and designed for RDF data storage and multi-node communication. An important feature of P2P is to change the current state of the Internet that is centered on large websites, return to “decentralization” and return the power to users. At present, the biggest advantage is improved network utilization for users. Because multiple nodes are connected with each other, the network bandwidth of users will be used to the maximum extent. 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. The system does not have a central node and belongs to a decentralized network structure. The system network structure is shown in Figure 4(A). The logical structure of the node is shown in Figure 4(B). Nodes are logically divided into three layers: service layer, protocol layer and data layer. The service layer is mainly node Web services, including query interface, node basic information interface, node quality information interface, etc. The protocol layer includes the HTTP protocol and P2P protocol after encapsulation. The data layer includes RDF data sets, basic node information, query logs, quality assessment ledgers and RDF entity record tables. The basic information of the nodes refers to the basic information of the nodes in the entire network, including routing information and quality assessment information. The quality assessment ledger is a concrete implementation of the quality assessment blockchain, which records transaction information and quality assessment updates information. The query log is a supplement to transaction information combined with knowledge quality assessment.
4.1 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 assessment of each node, information needs to be communicated between each node. So we propose a node-communicative calculation method.

When it is necessary to perform the connection calculation, the initiator establishes a temporary node to join the decentralized system. Temporary node has independent flags, indicating that they are temporary creating nodes. This node provides services that can verify its identity. Figure 5 shows the verification method of temporary nodes.
To ensure that the temporary node is indeed created by a certain node in the system, the parent node sends the RDF medical report to the temporary node when the node is created. Each node compares the hash value of the parent node RDF medical report provided by the parent node and the temporary node. The comparison is successful, indicating that the temporary node was created legally. At the same time, in order to ensure that each node in the decentralized system is knowledge-independent and protected, nodes will not transfer complete triplet information to temporary node, but rather a combination of <subject-predicate>. Finally, after the temporary node calculates the result of the task, it returns the result to each node in the system and adds the result to the calculation of quality assessment temporary node self-destruction.

Figure 6 illustrates the process of ensuring temporary node destruction. Considering security, after the destruction of a temporary node, two things need to be ensured: the temporary node no longer exists; the parent node A who created the temporary node does not steal information from other nodes. For the first point, each node accesses the destructor information of a temporary node, and accesses it, and updates the routing information if it is unreachable. If the temporary node no longer exists, each node makes its own query for the RDF data of node A, and node A then provides a service that only calculates the hash of the query result and does not return the result. When the temporary node is destroyed, the node A is queried to obtain the hash value. And when the temporary 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.
4.2 Advantages of Using Blockchain Record Quality Assessment Results

This paper uses the blockchain to record transaction information and quality assessment result information. The advantages of using blockchain to record quality assessment are as follows:

- **Structural security.** Quality assessment certificate is not issued without authority center. Nodes are mutually authenticated, and each node has backups of other nodes’ quality assessment. The quality assessment result that prevents central node collapse or being attacked is not credible.

- **Information security.** Prevent the node from tampering with the quality assessment results. The use of consensus mechanism prevents Byzantine attacks. A consensus mechanism was used to prevent Byzantine attacks. The entire network node maintains a super-book of quality assessment results. If a malicious node wants to forge quality information, it can be effectively screened through the entire network. Malicious nodes need to control more than 50 percent of the nodes to launch forgery attacks, but with a high cost.

- **Quality result update mechanism.** Previously, RDF data need to be recertified 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 assessment book.

- **Update log track can be found.** Each update of the update record has a basis. The quality assessment can be reviewed from the first generation to the last change.

With all the advantages stated above, the structure of DCQA system is robust. Nodes in the DCQA system are mutually authenticated, and the quality assessment results of the nodes are backed up to each other.
There are no central node in DCQA. This effectively prevents the problem that the quality assessment result is not credible when the central node collapses or is attacked. At the same time, the model uses the consensus mechanism to effectively prevent malicious nodes from tampering with the quality assessment results and ensuring the security of the information. For each RDF data update of the node, the DCQA model records the quality assessment result and the update log into the quality assessment ledger, which effectively implements the dynamic update mechanism of the quality assessment result and can refer to the corresponding update record. Such quality assessment results not only enable users to use the RDF data set with confidence, but also guide the system to provide users with more cost-effective query results.

4.3 Quality Assessment Blockchain Construction Method

This paper also discusses the application of incremental builds, because nodes are dynamically added to the system. When a new node joins the system, the node quality assessment is performed and the result is synchronized to the quality assessment book of each node. This section mainly introduces the specific process of generating and synchronizing quality assessment results when a new node joins the decentralized system.

First, when a new node joins the decentralized system, a quality assessment is required. Here is the process:

- The new node enters the decentralization system, synchronizing the quality assessment account.
- Create a new temporary node and notify other nodes.
- RDF data inspection report is provided to the temporary node.
- Each node in the network record the RDF inspection report hash value of the new node for later comparison.
- The hash value of the parent node’s inspection report is calculated by the temporary node. Each node compares the value with the hash value provided by the parent node to verify the validity of the temporary node.
- The combination of each node’s subject predicates combination.
- Temporary node computes the completeness, relevance, uniqueness and other related attributes of each node and returns them to each node.
- Each node generates an RDF entity record table.
- The result of the calculation of the quality assessment.
- The quality assessment results are broadcast to other nodes for signature.
- Write the initial record of the quality assessment hash into the Merkel tree.
- Temporary node self-destruction.
- Each node verifies that the temporary node is fully destructed and then updates the route.

According to the above process, the initial quality assessment of the new node of the decentralized system is completed. The quality assessment result is the first item of the quality assessment result update logs, and the update of the quality assessment results starts with this record.
Figure 7 shows the process of node quality assessment record generation. The node generation quality assessment result is broadcast to other nodes for signature. The signature here is the chain signature. That is, according to the order of joining the system, each node signs, and finally returns to the temporary node. After obtaining all node signatures, the first quality assessment record is generated.

![Figure 7. Node quality assessment record generation process.](image)

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 assessment results in the same way, each update has the same effect. Each update is based on the results of the last quality assessment. The update log contains block ID, update items, signature and timestamp.

An update item is a dimension that is reduced or promoted in one update. The space occupied by the record update item is more than the space occupied by only the incremental result. However, if the incremental quality assessment result is recorded, the update of the quality assessment result cannot be fully represented. Therefore, the results are then updated, and each block calculates a new quality assessment result by itself. 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. Normally, the signature list should include the signatures of all blocks. The block ID is used to describe a block that updates the quality assessment result. We can form a time tree according to the timestamp. Each update is based on which items of the last quality assessment result are updated, and the relationship between the updates is realized, which can effectively prevent the update fraud of quality assessment.

After the update log is generated, it is released to each block for signature in Figure 8. After all nodes have been signed, the results are recorded. Here we use the Merkel tree [14] storage quality assessment results. Merkel tree is a tree in which every leaf node is labelled with the hash of a data block and every non-leaf node is labelled with the cryptographic hash of the labels of its child nodes. Hash trees allow efficient and secure verification of the contents of large data structures. Use the Merkel tree can effectively prevent the result of tampering. Any tampering will cause a change to the hash value that is stored by the root node. When the hash value is inconsistent with other nodes, the quality assessment result in the block is considered to be unreliable. The node will then be removed from the decentralized network.
4.4 Interaction between User Behavior and Quality Assessment

There is a bidirectional relationship between user behavior and quality assessment. User behavior in this paper includes two parts: the conclusion of a transaction and user feedback. Each query is essentially a transaction. Here we take query as an example to explain how query and quality assessment interact.

According to Equation (9), query will affect quality assessment. The higher the query frequency is, the higher the quality assessment will be. The following is the processing flow of a query.

- Receive the query statement and query.
- Feedback the results to users.
- Update the query log.
- Update the RDF data set entity record table.
- Generate transaction information.
- Wait for T time for user feedback.
- Update the data entity record table and query log if user feedback is available.
- End this query, after the operation will not affect this query.
It can be concluded from the above process that the maintenance of the query log and the entity record table of the RDF data set does not occupy the query response time, but is performed after the query result is fed back to the user. This process reflects how queries and user feedback affect quality assessment. Likewise, quality assessment guides the query mechanism to return more cost-effective results. The following are the basic rules of the system:

- When the cost of query results is the same, the data of nodes with high service quality are preferred.
- When the query results are different, the data of nodes with a higher ratio of the number of result data to the price are recommended.
- When part of the query results are the same, but part of the nodes have uniqueness entities, higher quality node data of query results are recommended.

5. EXPERIMENT ANALYSIS

The objectives of the experiment are to verify the verifiability of the model, the integrity of the data, the impact of the user’s feedback on the quality assessment results, and whether the quality assessment book is safe. This section proposes a decentralized system based on the third and fourth sections of the model and system design. A series of experiments are carried out.

Experimental environment: MacBook Pro laptop with a specification of 2.6 GHz Intel Core i7 processor, and a 16GB 2133 MHz LPDDR3 memory, constructed in this environment simulation to implement a decentralized system with 6 nodes. The protocols used include the Http protocol and the P2P protocol. The experiment uses the Archives Hub data set. It is a sample data set of descriptions of archive collections held on the Archives Hub, a UK aggregator, and output as linked data. The Hub linked data provide a perspective on the people, organizations, subjects and places connected with the archives that are described. The size of the data set file is 71.8M, with a number of entities 106,919, a number of unique subjects 51,411, a number of unique predicates 141, a number of unique objects 104,408, and a number of triples 431,088. In order to highlight the importance of the quality assessment dimensions such as verifiability, completeness, and uniqueness in the model, divide the Archives Hub data set into six parts and record it as AH1 to AH6. Table 5 shows basic information of the six data sets. To increase the completeness judgment, we selectively replicate some node information in AH1 to AH6.

© Archives Hub data set source http://data.archiveshub.ac.uk/.
Table 5. Basic information of the experimental data set.

| Data set | File size (M) | The number of entities | The number of triples | The number of unique subject and predicate | RDF inspection report |
|----------|---------------|------------------------|----------------------|-------------------------------------------|----------------------|
| AH1      | 5.7           | 16,748                 | 28,361               | 6,945/89                                   | 4.05                 |
| AH2      | 6.7           | 20,514                 | 39,705               | 9,271/46                                   | 4.28                 |
| AH3      | 9.9           | 20,876                 | 56,722               | 2,676/43                                   | 21.1                 |
| AH4      | 15.1          | 28,366                 | 79,410               | 4,576/66                                   | 17.35                |
| AH5      | 16.3          | 40,803                 | 102,099              | 12,983/29                                  | 7.86                 |
| AH6      | 19.3          | 43,590                 | 124,796              | 14,970/101                                 | 8.34                 |

The RDF inspection report in Table 5 is calculated according to Equation (3), where $k_1 = 1$, $k_2 = 1$, $k_3 = 1$ and $k_4 = 1$. The coefficient values are all set to 1 because in this system we assume that the weights of the four factors affecting RDF data quality mentioned in the previous paragraph are equal.

Before the system accepts the query, the system does not activate the verifiability and uniqueness in the model (Equation (9)), so the result of the RDF medical examination report is the initial quality assessment result of the node.

5.1 Quality Assessment Model Verification

5.1.1 Verifiable Model Validation

Firstly, verify the verifiability of the model. Use statement 1 and statement 2 to activate the model. The activation process is to make 100 queries after the decentralized system is set up. The language uses SPARQL [15], which is a query language and data acquisition protocol developed for RDF. Calculate and update the new quality assessment results of each node.

Statement 1: select*{?s?p?o}where{<http://api.talis.com/stores/localh/items/1305283343810#self>?p?o.}

Hit result: AH1, AH2, AH6

Statement 2: select*{?s?p?o} where {<http://data.archiveshub.ac.uk/id/perso-n/aacr2/martindorothyfreeborn>?p?o.}

Hit result: AH6

Once the quality assessment results are updated, update the entities hit by statement 1 in the data set after the 100th query and then execute statement 1 again. The query result record is shown in Figure 9.
Figure 9. Impact of verifiability on quality assessment.

According to the records in Figure 9, 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 time of 101st, verifiable landslides occurred, resulting in a sharp decline in quality assessment results. The reason is that the data in AH1 and AH6 were updated, which caused the query results to be inconsistent with the previous query ones. For users, frequent changes in data are undesired, so a large-scale decline in quality assessment is in line with expectations. At the same time, with the increase of the number of queries, the quality assessment has steadily increased linearly. If data updates occur during this process, data quality will still be degraded. After comparing the quality assessment results between AH1 and AH6, we found the reason why AH6 quality degradation is lower than AH1 after 100 queries is that statement 2 was executed during the activation model process, resulting in the fact that the AH6 query log contains two parts, and the update data do not affect the statement 2 Hit results. Therefore, the quality of AH6 declines less. By comparing AH6 with AH1, it is verified that the granularity of verifiability model is at the log level.

5.1.2 Impact of Data Completeness and User Feedback on Quality Assessment

Table 6 is an RDF entity record, which is generated after the model is activated. As the number of queries increases, the quality of the node data gradually increases.

| Data set | Entity | Uniqueness coefficient | Query frequency |
|----------|--------|------------------------|-----------------|
| AH1      | self<sup>a</sup> | 0.24                  | 100             |
| AH6      | Martindorothyfreeborn<sup>b</sup> | 1                     | 100             |
| AH6      | self   | 0.378                 | 100             |

Note: <sup>a</sup>self stands for entity at http://api.talis.com/stores/localh/items/1305283343810#self.

<sup>b</sup>Martindorothyfreeborn stands entity at http://data.archiveshub.ac.uk/id/perso-n/aacr2/martindorothyfreeborn.
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. Figure 10 shows the quality change in AH1 and AH6 during this process.

![Figure 10. AH1 and AH6 verification uniqueness model.](image)

In Figure 10, 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 multiple subjects. After 500 inquiries, less than 50 percent of users added feedback to interference, and the increase in quality became slow. After 700 inquiries, the interference was adjusted to more than 50 percent, and the quality was reduced. We can see that the quality is related to user feedback coefficient. So we can use this coefficient to adjust the impact of user behavior. We also find the increase and decrease of AH6 are 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.

### 5.2 Quality Assessment Book Safety

Although the blockchain solves the problem of security, the 51% attack can still happen. That is, if someone has mastered the computing power of more than 51% of the whole network, they can preempt a chain of longer, forged transactions like a race. This paper calculates the difficulty of controlling each node and analyzes which nodes will launch attacks from the perspective of interests. If node A, B, C and D contain subject s1, the number of attribute is 9, 5, 31 and 7, respectively, among which the common attribute is 5. 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:
An RDF Data Set Quality Assessment Mechanism for Decentralized Systems

\[ \text{Cost}_{(B)} \geq \sum_{i=0}^{n} \sum_{j=1}^{m} \text{price}_{(\text{node}_{i})} \times \text{Uniqueness}_{(\text{data}, S_j)} \]  

(13)

Where \( node_i \) is the number of nodes that you intend to bribe. The relationship between the number of nodes and the cost value is shown in Figure 11. It can be known that the more bribe nodes cost, the higher the data quality of the nodes that intend to bribe, the higher the cost. To get 51% of support, the cost is much greater than itself. As the network grows, the probability of launching a 51%-attack approaches 0.

![Figure 11. Relationship between node number and cost value.](image)

6. CONCLUSION

This paper proposed a method on assessing and updating the quality of RDF data in the context of the rapid developing Web and decentralized systems. Firstly, it indicates the significance of this research and the importance of achieving RDF data quality assessment in a decentralized system. Then an RDF data quality assessment model in a decentralized system is proposed. Finally, based on the node quality assessment model, a decentralized quality assessment system DCQA is designed and implemented. At the same time, we also proposed and discussed in detail the use of blockchain storage quality assessment results. The obtained RDF data quality assessment results combined the advantages of the blockchain without having to issue a quality assessment certificate by an authoritative center, and effectively prevented tampering and supported dynamic updates of the record query. The purpose of this paper is to promote the research and development of RDF data quality assessment in a decentralized system, and to provide users...
with better services. In the future, we will carry out research in the following aspects: 1) the new dimension of quality assessment in a decentralized environment and 2) the impact of price factors on the model.

**AUTHOR CONTRIBUTIONS**

All of the authors made great contributions to the work. L. Huang (huangli82@wust.edu.cn) is the leader of this work. She designed the RDF data quality assessment mechanism and DCQA system framework in a decentralized system. Z. Liu (540900559@qq.com) and F. Xu (xuff@wust.edu.cn) summarized the methodology part of this paper. J. Gu (simon@wust.edu.cn) as the corresponding author summarized and drafted this paper. All authors have made meaningful contributions to the revision of the paper.

**ACKNOWLEDGEMENTS**

This work was supported by the National Natural Science Foundation of China (Grant No: U1836118, Grant No: 61602350), the Key Projects of National Social Science Foundation of China (Grant No: 11&ZD189) and the Scientific Research Project of Education Department of Hubei Province (Grant No: B2019008). Thanks to graduate students Yansong Wang and Ren Hui for their professional guidance in the process of implementing the DCQA system in this paper, as well as Professor Lynda Hardman from Information Access Research Group, Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands, for her suggestions for modification and improvement of this paper.

**REFERENCES**

[1] S. Nakamoto. Bitcoin: A peer-to-peer electronic cash system. Available at: https://bitcoin.org/en/bitcoin-paper.
[2] M. Swan. Blockchain: Blueprint for a new economy. Sebastopol, CA: O'Reilly Media, 2015. isbn: 9781491920497.
[3] R. Beck, J. S. Czepluch, N. Lollike & S. Malone. Blockchain—The gateway to trust-free cryptographic transactions. In: The Twenty-Fourth European Conference on Information Systems (ECIS), 2016, pp. 1–14. Available at: http://aisel.aisnet.org/ecis2016_rp/153.
[4] T. Berners-Lee, J. Hendler & O. Lassila. The semantic web. Scientific American 284(5) (2001), 28–37. doi: 10.1038/scientificamerican0501-34.
[5] S. Decker, S. Melnik, F. van Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann & I. Horrocks. The semantic web: The roles of XML and RDF. IEEE Internet Computing 4(5)(2000), 63–73. doi: 10.1109/4236.877487.
[6] A. Newitz. Web 3.0. New Scientist 197(2647)(2008), 42–43. doi: 10.1016/S0262-4079(08)60674-0.
[7] C. Bizer, R. Cyganiak & T. Heath. How to publish linked data on the web. Available at: http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/LinkedDataTutorial/.
[8] A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres & S. Decker. An empirical survey of linked data conformance. Journal of Web Semantics 14(2012), 14–44. doi: 10.1016/j.websem.2012.02.001.
[9] J. Kandari. Information quality on the World Wide Web: A user perspective. PhD dissertation, University of Nebraska-Lincoln, 2010. Available at: doi: 10.1504/IJIQ.2011.043784.
An RDF Data Set Quality Assessment Mechanism for Decentralized Systems

[10] A. Assaf & A. Senart. Data quality principles in the semantic web. In: IEEE Sixth International Conference on Semantic Computing, 2012, pp. 226–229. doi: 10.1109/ICSC.2012.39.

[11] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann & S. Aue. Quality assessment for linked data: A survey. Semantic Web 7(1)(2016), 63–93. Available at: http://www.semantic-web-journal.net/system/files/swj773.pdf.

[12] C. Bizer & R. Cyganiak. Quality-driven information filtering using the WIQA policy framework. Journal of Web Semantics, 7(1)(2009), 1–10. doi: 10.2139/ssrn.3199414.

[13] D. Magazzeni, P. McBurney & W. Nash. Validation and verification of smart contracts: A research agenda. Computer 50(9)(2017), 50–57. doi: 10.1109/MC.2017.3571045.

[14] R.C. Merkle. Protocols for public key cryptosystems. In: 1980 IEEE Symposium on Security and Privacy. IEEE, 1980: 122–122. doi: 10.1109/SP.1980.10006.

[15] SPARQL. Available at: https://www.w3.org/TR/sparql11-query/.
AUTHOR BIOGRAPHY

Li Huang is currently an associate professor of computer science of the College of Computer Science and Technology, Wuhan University of Science and Technology. She received her PhD in computer science from Huazhong University of Science and Technology in 2011. Her research interests include data management and semantic Web and knowledge.
ORCID: 0000-0001-5702-2321

Zhenzhen Liu is currently a postgraduate student in the College of Computer Science and Technology, Wuhan University of Science and Technology. He received a bachelor’s degree from Wuhan University of Science and Technology in 2017. His research interests include semantic Web, knowledge graph and linked data quality assessment.
ORCID: 0000-0003-2066-312X

Fangfang Xu received her B.S. degree from the College of Computer Science and Technology at Wuhan University of Science and Technology (WUST) in 2012 and the M.S. degree from College of Computer Science and Technology at WUST in 2015. Now she is an engineer at the College of Computer Science and Technology, WUST. Her research interest is semantic computing.
ORCID: 0000-0003-1057-6444
Jinguang Gu is currently a professor of the College of Computer Science and Technology, Wuhan University of Science and Technology (WUST) and vice dean of Institute of Big Data Science and Engineering, WUST. He received his PhD in software theory from Computer School of Wuhan University in 2005. His research interests include knowledge graph, semantic Web, and distributed and service computing.
ORCID: 0000-0002-8823-8480