Journal of Engineering and Applied Sciences 13 (Special Issue 5): 4705-4711, 2018
ISSN: 1816-949X

RDF (Resource Description Framework) Serialization for Webpage Integrity

Asaad Sabah Hadi and Alyaa Abdul Hussein Alkaabi
Information Technology College, University of Babylon, Babylon, 51001 Hilla, Iraq

Abstract: Integrity of information in LOD (Linked Open Data) is very important issues. The RDF (Resource Description Framework) is a special language for describing data in semantic web. The aim of integrity is checking any change of information. The aim of this study is to improve the integrity of RDF information. The proposed system use ("Woolly mammoth") RDF file from BBC News web page as an example .The system analyze the RDF in order to find all its statements then it will convert these statements into three dimensional array according to RDF representation (subject, predicate, object). The content of the array is binary values (0, 1) where the value '1' represent the existence of the statement in RDF whereas the '0' value means that these statement is not exist. The three dimensional array are converted into a bit stream that represent the message to be checked in order to improve the integrity. Also, the system add the date and time to the bit stream in order to check the validity of the RDF data. To reduce the length of the bit stream we use one of the compression techniques. The message can be saved in the same file (hiding it) or save in an external file that can be checked later. When we want to check that our RDF has the same content and there is no change in its data, we reload it and convert its content into a bit stream and use XOR to match the new stream with the old one to check the integrity in a simplest way.

Key words: Information integrity, data integrity, data quality, LOD, RDF, compression techniques

INTRODUCTION

The new advantage that the semantic web adds to the original web is opening up data and information also, making it possible to use and re-use it Powers (2003). So, the Linked Open Data (LOD) is a growing movement for organizations and other users of web. The robust data language for LOD is the Resource Description Framework (RDF). RDF has features that even if the underlying schemas differ, it support to merging of data (Bizer *et al*., 2009). RDF help to name the relationship between things on web by using URIs (Uniform Resource Identifier) (Joshi, 2013). The usability of the web make the information integrity is an important issue today. So, the management of information is a distinctive discipline and decisions are based on the available information (Hausenblas and Karnstedt, 2010). Then the poor quality information lead to lost productivity. The aim of this study is to improve the integrity of RDF information. The proposed system use ("Woolly mammoth") RDF file from BBC NEWS web page as an example. The system analyzing RDF using regular expression concept then convert the content to three dimension bit predicate matrices called BR-Mat (Binary Representation Matrix). BR-Mat is a binary representation for statements in RDF graph. The bit 1 mean appearance of statement and bit 0 mean not. Because of the sparsity of this matrices, need a large memory to save it. So after, we convert it to 1 dimension bit stream, we use a run length algorithm to compresses it. Also, we add a date and time to know the last update. This bit stream will using by hiding it in the original RDF file to ensure the integrity in our future research but now we check the integrity of RDF by comparing the bit stream with bit stream of the same file after period of time to discover if an update or damage made on it by using the XOR function.

Semantic web: The World Wide Web Consortium (W3C) adds an expansion to the original web called semantic web (WWWCW3C., 2011). The W3C standards most data formats and interchange protocols on the Web by using RDF. According to the W3C, we now can share data and reused it across application, enterprise and community boundaries by using the semantic web (Schmidt and Lausen, 2013).

RDF: The standard language that designed to support the semantic web is the Resource Description Framework (RDF) in the same way that HTML is the language that helped start of the original web (Powers, 2003). RDF based

Corresponding Author: Asaad Sabah Hadi, Information Technology College, University of Babylon, Babylon, 51001 Hilla, Iraq

on the principle that three is a magic number (triple) that define three pieces of information (subject, predicate, object). The subject is a property such as name can belong to animal, book, plant, person, nation, car or any web resources. The predicate is a fact about any individual subject, like a gender, a height, a hair color an eye color, a college degree, relationships and so on. The object is the value associated with that property. The triple's subject is URI, the object can either URI or a string literal and both the subject and object defined a resource. The predicate is represented by URI and specifies the subject and object relationship. The RDF is directed graph contain a set of nodes connected by arcs, forming a pattern of node-arc-node. Several common representation formats of RDF are in use, including, Turtle (Terse RDF Triple Language), N-Triples, RDF/XML, etc. Figure 1 shows the main structure of RDF triple.

Linked open data: Linked open data is a way of publishing structured data that allows metadata to be connected and enriched. In Linked open data the different representations of the same content can be found and there are links made between related resources. The linked data infrastructure can be used not only as a means of publishing open data but also as a general mechanism for managing distributed graph data. LOD used RDF (resource description framework) to connect different type of data that use URIs to provide useful RDF information. Linked data provide one main addition to the semantic web principles that all entity URIs should be dereference able to make a reliable RDF representation. Berners-Lee resume a set of "rules" that help to promulgation data on the web like that the published data becomes a section of one general data space:

- The standards using URIs to identify resources
- Use HTTP URIs addresses so that people can search that identified resources
- The criterions (RDF, SPARQL) that helps to give a useful information to the user how search a URI
- The URIs use links to each other to enable users discovers more information

These rules known as the "linked data principles" and provide a basic form for promulgation and connecting data using the basic physical and organizational structures and facilities of the web while make a commitment of its standards and architecture (Bizer *et al*., 2009). The linked open data has been growing so fast in the last few years from its first published to nowadays. Figure 2 shows the linked open data of 2014.

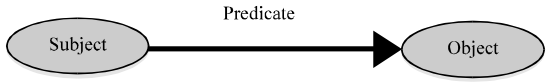

Fig. 1: Resource Description Framework (RDF)

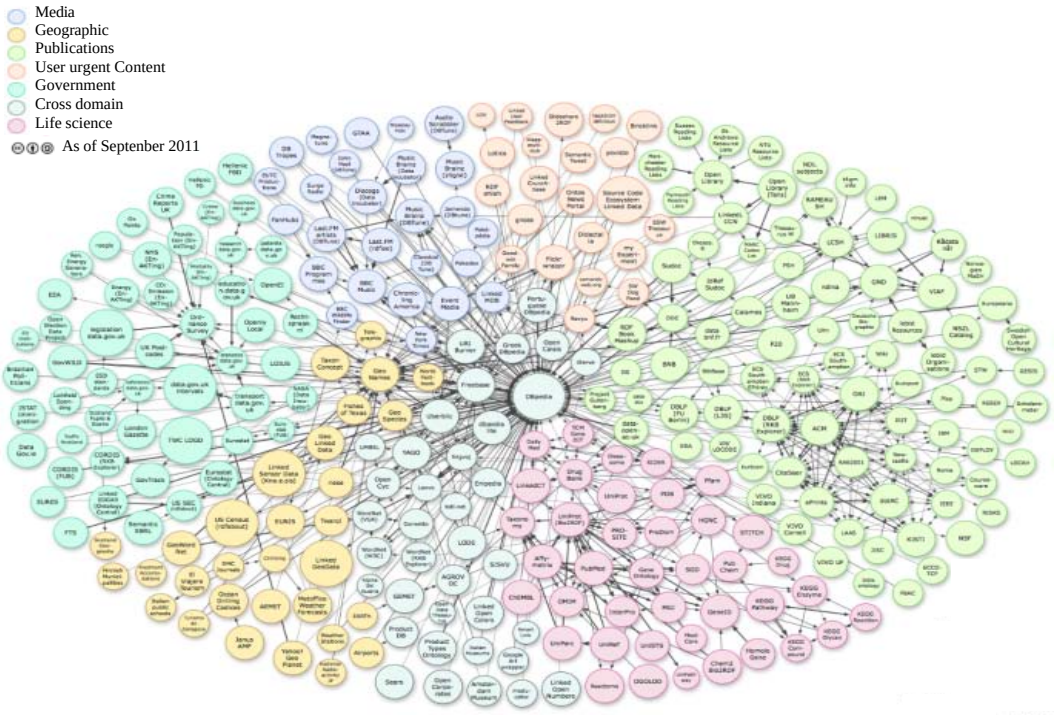

Fig. 2: Linked open data (2014)

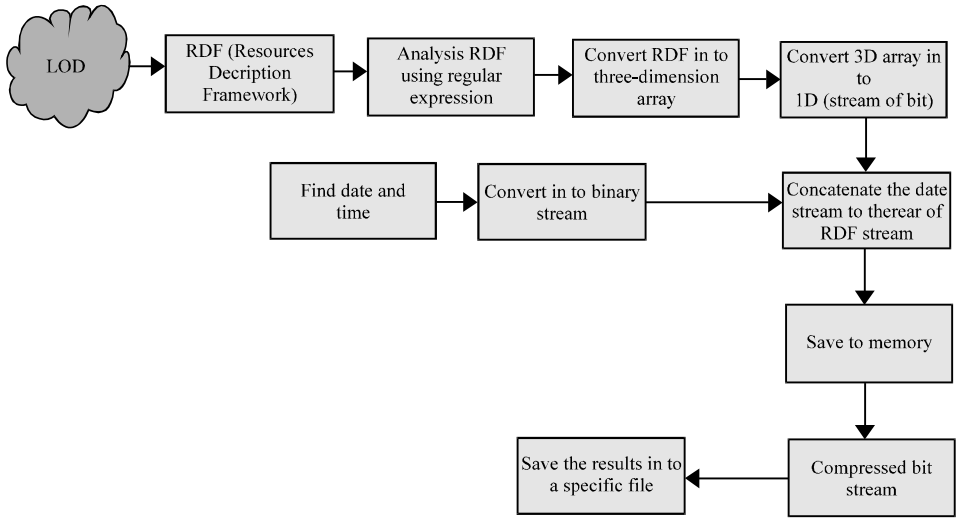

Fig. 3: Analyzing RDF phase-1

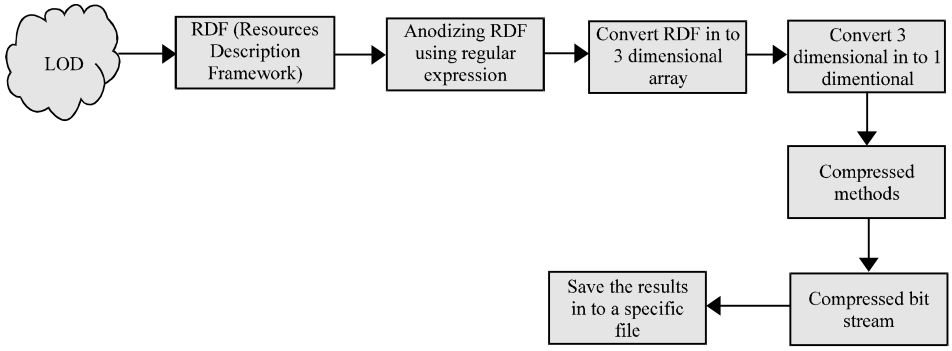

Fig. 4: Integrity checking phase 2

MATERIALS AND METHODS

Proposed system: The proposed system contains three main phases as shown in Fig. 3-6.

Analyzing phase: RDF analyzing using regular expression (Goyvarets and Steven, 2009). Regex or regexp also, called pattern is an expression used to characterize a set of string needs for a particular purpose. Regular expression consists of constant that denote a set of string and operator symbols that represent the operations on these sets. Our proposed algorithm read RDF-XML file and Extract the file as text. After prepare a pattern for the search, the algorithm segments the content and use the pattern to search through the text (using regular expressions to draw a pattern). After extract the data from the text into the storage structure we declare lists of predicate list, subject list, object list object list as shown in Table 1.

Table 1: Predict list subject list object list example

ID	Values
1	Predicate 1
2	Predicate 2
3	Subject 1
4	Subject 2
5	Subject 3
6	Object 1
7	Object 2
8	Object 3
9	Object 4

BR-Mat (Binary Representation Matrix): It is a binary representation for RDF triples (Atre *et al*., 2009). To represent each triple in two dimensional bit matrix we slice it along the predicate dimension which give us a predicates matrices of size (P*S*O). Also, the matrices can be slicing along subject or object dimension but the predicate dimension represent the most favorable structure in the term of compactness. The bit matrices extracted from storage, Table 2 for RDF data. Each column represent object and each row represent subject in this

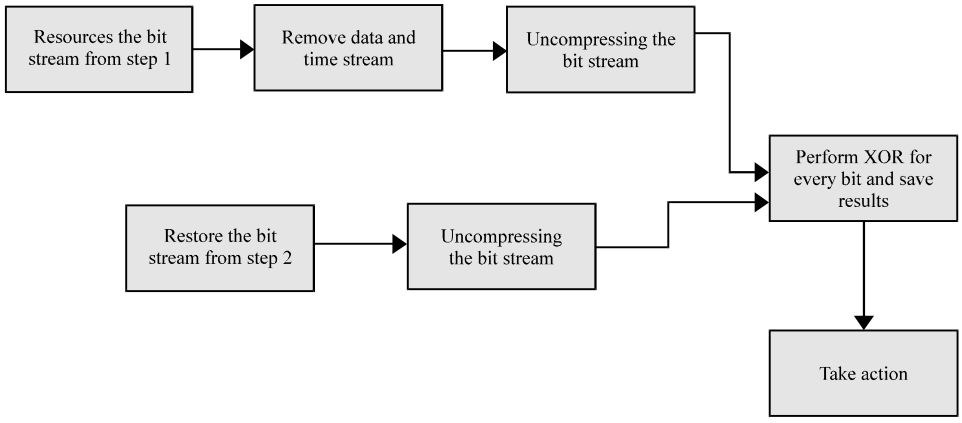

Fig. 5: Comparison phase 3

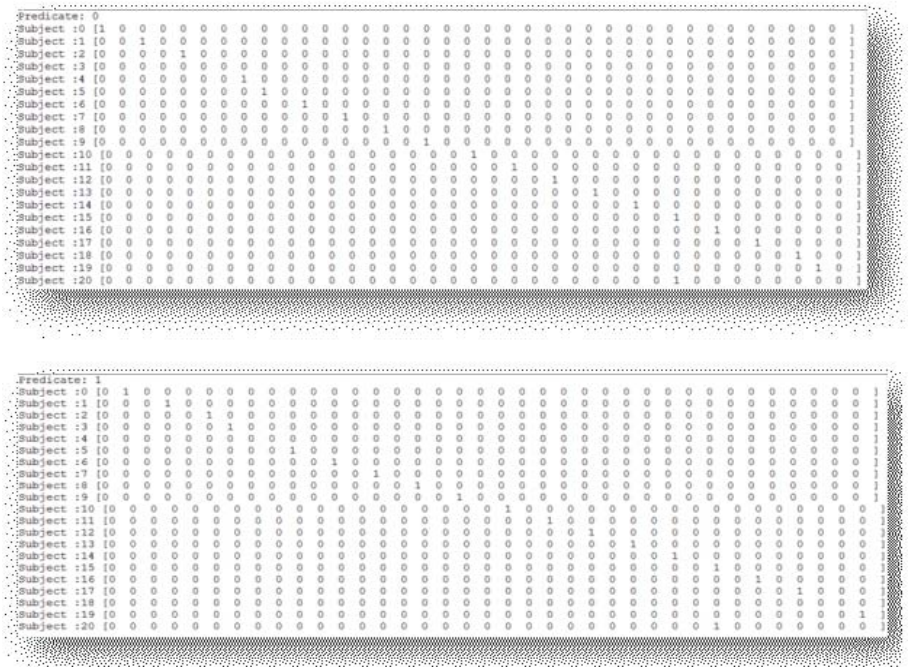

Fig. 6: Bit matrices for woolly mammoth

Table 2: XOR truth table

Input		
A	B	Output
0	0	0
0	1	1
1	0	1
1	0	1
1	1	1

predicate matrices. The 1 bit mean that S∩0 in RDF graph and 0 bit otherwise. Then, we convert these matrices into one dimensional bit stream to use it for integrity matching. Also, we take the date and time of last update for LOD page information and convert it to a bit format then concatenate it to the rear of RDF bit stream. The final result is a one-dimension sparse bit vector that contain a large number of zeros, so, in order to reduce the memory storage used for save this stream we use run length coding compression method.

Compression technique: In order to reduce the size of binary stream, we use compression technique. In this study, we use RLE (Run Length Encoding) algorithm which is a lossless compression technique that used to eliminate redundant data. The decompressed data are identical to the original uncompressed data (Salomon, 2013). RLE is characterized by compress any type of repeating data sequence and give high compression ratio. In RLE algorithm the consecutive sequences of the same

data value (runs) are stored or transmitted as a single data value and count. Rather than the original individual data elements. RLE algorithm used to compress the sparsity bit matrices to reduce the size of memory that used to store large RDF data by storing the compressed bit vector. This compression technique research very well if the input stream contains of tow symbol (0, 1). For example, the binary stream:

```
0000001000000000000001111111000
```

The compressed bits:

```
6011407130
```

The size need for save the input stream is 31 byte while the size for compressed data is (10) byte which saving (31)/(10) = (3.1) (Algorithm 1). The decompressed step:

```
0000001000000000000001111111100
```

Algorithm 1; The decompressing ratio is (100%):
Run length encoding algorithm
Loop: count = 0
Repeat
Get next symbol
Count = count+1
Until (symbol not equal to next one)
Output count and single value Go to loop

Integrity checking: By increasing the users of web and use it as a source for information that make the improvement of information a very important task (Flowerday, 2007). There are a number of problems that face the users of information like, download personal information like (CV) and they want, after a period of time to check if that information has been changed or not. Also, download scientific web page like (DB pedia) and use it as a reference and after some time they need to check if any updates or damage made on that information. Also, maintains process doing by authorized people to check and detect any attack and unauthorized update. So, to check the integrity of our RDF information we download the same RDF in another time and perform the algorithm in phase-1 by reading the file and analyzing it then convert it to three dimension of predicate matrices, after that, we convert it to one stream of bit then perform our compression method and save the file to use it in comparison step to know if any change made on that information.

Comparison: For comparison step we apply the following. Restore the compressed bit stream that, we have from step 1 after removing the date and time and uncompressing it. Restore the compressed bit stream from step 2 and uncompressing it. After that we have two 1 dimension uncompressed binary stream. Because these bit vectors as we mentioned before, represent the appearance or not appearance of the statements in RDF graph, therefore any change or attack made on RDF statements will change these bits. For example, if one of these statements has been deleted the bit that represent the statement will change from 1-0. After that our proposed system use the XOR function to check the integrity using these bits. Where XOR is a logical operation that its outputs is true only when inputs differs (one bit is true, the other is false) (Davies) and give the output '0' if these bits identical to each other. The input to this function is the two uncompressed bit streams from above step, after we do this operation we have one-bit stream as a result. If the result from this operation all zeros (0) then this will prove the integrity of the information else if the result ones (1) in any bit then the information was exposed for update.

If the system return that the information was exposed for update, either in a legal or illegal way. If the authorized person, who's owen the web page, detect the problem then he can upload the original information. If unauthorized person (users of linked open data) then he can download the new version of the updated webpage.

RESULTS AND DISCUSSION

In order to test our proposed system, we take (woolly mammoth) RDF file from BBC NEWS web page as a sample. This RDF file contain (179) triples. The no. of subject is 21, the number of predicate is 2 and the number of object is 37. Algorithm 2 shows statement from (woolly mammoth) RDF/XML file and its subject, predicate and object. Whereas, Fig. 6 shows the bit matrices for woolly mammoth.

Algorithm 2; RDF/XML statement:
<rdf: Description rdf: about = "/nature/species/woolly-mammoth">
<foaf: Primary topic rdf: resources = "/nature/species/woolly-mammoth#species"/>
<rdfs: See also rdf: resources = "/nature/species"/>
</rdf: Description>
Subject http://www.w3.org/nature/species/Woolly_mammoth
Predicate http://xmlns.com/foaf/0.1/primaryTopic
Object http://www.w3.org/nature/species/Woolly_mammoth#species

The bit matrices shown in Fig. 6 give as a large sparse matrices and using the compression method for this matrices give us a high compression ratio according to compression ratio equation:

$$r = \frac{\text{100-size of new matricx}}{\text{size of orignal matrix}}$$

Fig. 7: One dimension stream of bit array-1, array-2

Fig. 8: Compression of bit streams array-1, array-2

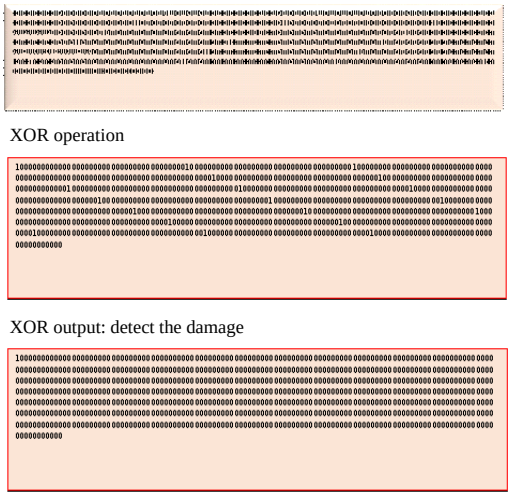

Fig. 9: Integrity checking for array-1

Figure 7 shows the matrices after convert it to 1 dimension stream of bit whereas, Fig. 8 shows the compression of these stream bits.

Fig. 10: Improving integrity for Array-2 from

After, we decompress the vectors then, we use the XOR function on the two matrices after the RDF file exposed to damage in some statements for predicate-1. Figure 9 and 10 show the detection of Integrity in the RDF File.

CONCLUSION

This study displays a suggestion to improve the information integrity in linked open data by analyzing

RDF file using regular expression and convert it to binary representation and using a compression method to reduce the size of storage using to save RDF. To check the Integrity, we apply the same algorithm to RDF in another time and decompress the binary stream from step 1 and 2 then do comparison step using XOR function.

RECOMMENDATIONS

When the output all zeros that mean the information has no change else we detect the change if the outputted bit is 1. The future research will use the compressed binary stream as message to hide it inside the original RDF file.

REFERENCES

Atre, M., V. Chaoji, J. Weaver and G.T. Williams, 2009. Bitmat: An in-core RDF graph store for join query processing. Master Thesis, Rensselaer Polytechnic Institute, Troy, New York.

Bizer, C., T. Heath and T. Berners-Lee, 2009. Linked Data the Story so Far. In: Semantic Services, Interoperability and Web Applications: Emerging Concepts, Amit, P.S. (Eds.). IGI Global, Hershey, Pennsylvania, USA., pp: 205-227.

Flowerday, R., 2007. What constitutes information integrity?. South Afr. J. Inf. Manage., 9: 1-19.

Goyvarets, J. and L. Steven, 2009. Regular Expression Cookbook. O'Reilly Media, Inc, Sebastopol, California, USA., ISBN:978-0-596-52068-7, Pages: 479.

Hausenblas, M. and M. Karnstedt, 2010. Understanding linked open data as a web-scale database. Proceedings of the 2010 2nd International Conference on Advances in Databases Knowledge and Data Applications (DBKDA), April 11-16, 2010, IEEE, Menuires, France, ISBN:978-1-4244-6081-6, pp: 56-61.

Joshi, A.P., 2013. Linked Data for Software Security Concepts and Vulnerability Descriptions. University of Maryland, College Park, Maryland, Pages: 166.

Powers, S., 2003. Practical RDF: Solving Problems with the Resource Description Framework. O'Reilly Media, Inc, Sebastopol, California, USA., ISBN: 978-0-596-00263-7,.

Salomon, D., 2013. A Guide to Data Compression Methods. Springer, Berlin, Germany, ISBN:978-0387-95260-4, Pages: 287.

Schmidt, M. and G. Lausen, 2013. Pleasantly consuming linked data with RDF data descriptions. Proceedings of the 4th International Conference on Consuming Linked Data Vol. 1034, October 22, 2013, ACM, Aachen, German, pp: 62-73.

WWWC W3C., 2011. W3C semantic web activity. World Wide Web Consortium W3C, Cambridge, Massachusetts, USA.