# Improving data quality using techniques from human computation

Vikram Kumar Kirpalani and Syed Saif ur Rahman
SZABIST
Karachi, Pakistan

***Abstract:*** **The DBpedia is an open data repository extracted from a crowd sourced Knowledge base Wikipedia, because of which the information available there is more vulnerable to inconsistency, grammatical errors, structures and, data type problems. These are just a few issues that existing data is prone to. In this research, our prime focus would be on Data type problems, particularly, the problem of one attribute containing multiple facts. Proposed is the approach in which the issue of Inconsistent behavior of the desired output is addressed and improved, on the idea based on concept hierarchy i.e. association of Parent - Child relationship by employing Human computation and the confidence, and trust of the output has been calculated, out of which the Hierarchies of the entities could be maintained in the form of Triples which could be used to mapped on other data to led the data out of the problem of Implicit type of relationship between attributes..**

*Keywords: Human computation; Crowd sourcing; Data quality; Linked data; DBpedia.*

## I. INTRODUCTION

This research work is based upon the idea of the improvisation and the enhancement of the quality of the data, by employing Human computation as the Prime key for the same. For the purpose of creating a base of this research work the author would like the reader to go through the concepts of Human computation, Crowd Sourcing and DBpedia [Linked Data] to have a fine understanding of paradigm of this research work.

### *1.1 Human Computation*

When talking about data quality there are certain attributes (Table 1. of Section 5.2) which are to be met, which could be different in different domains, like if a table exist for wrong answers the validity cannot be based on facts, for example if there exist an statement within that table that sun rises in the north, the quality attribute of Validity and Accuracy still exist if it was meant to be wrong in that particular table and context, but in real world it is not valid, now if a computer is said to validate it, based on facts, it will pick it up as wrong. This is where Human computation can easily be understood, Humans can do those tasks very comfortably which computers would feel difficult to do, rather it would be peculiar to make computer understand what is going wrong for every single thing, usually the tasks like translation, brainstorming and writing are much easier for a human to do rather than a computer. Keeping in mind these facts, the idea is creation of a framework where data quality can be improved using human cognition to achieve better results in a lesser time with achieving quality by verification of data quality by user feedback [1, 2, 3]. In this study we will be working with the data sets from DBpedia which provides us with the data of Wikipedia and is open to use, and we will make use of techniques from human computation to improve data quality. Even if computers are set to automate the process by defining rules [4], humans are still needed to verify that what the computer has completed accomplishing; we can make this process a bit semi-automated for better results.

### *1.2 Crowd Sourcing*

The term Crowd Sourcing is derived from two different words, which are crowd, i.e., group of people homogeneous in interests, and outsourcing. The term is usually meant when there is a group of people working over some problem or issue, together in parallel, or even in iterative manner i.e., based on the task completed by one, another version of the same based on preceding is created. freelancing is now being replaced by crowd sourcing as in freelancing the requester asks for something made based on the requirements the requester have, and the worker bids as per the nature of the task, and builds on it after the work is complete the requester could ask for modifications if required or as agreed upon contract. But in Crowd Sourcing a requester can ask for something based on it multiple users can work on, finally the requester will pay to the one he/she likes the work of. To make it more clear let's consider a very common example of a products review magazine. The magazine would offer in-depth reviews for a set of products which have been launched recently in the market. Now each review would be by a single expert who has used the product. If crowd sourcing is implemented here, reviewers can login to the site and post reviews relating to the product and in this way a large number of reviews would be generated. The positivity of the reviews would let a consumer know what product if better for them and what pros and cons are there for that product. Apart from this more products can be added as reviews on different products are requested by consumers. All reviewers would be login in using some social media platform so the reviews won't be anonymous and people would know who has reviewed and can ask the reviewer further questions.

### *1.3 Inter HC and CS Application*

While the sad news of Malaysian airline MH-370, shocked the whole world, somewhere within the world thought a very different and workable way, to Crowd Source the task of searching the lost entity, using human computation, such thing also happened to prove it reliable and time savvy back in 2011 United Nations Refugee agency named as UNHCR to evaluate the number of homes that got fled in the incident occurred, by searching for signs like temporary shelter and this job was completed in approximately 120 hours, and similar job 2 years back took two months to get completed the other traditional

way [6]. Tomnod [7] is a web-based application/portal that employees' crowd sourcing and human computation to solve real world problems, right now featuring the search for the missing plane of Malaysian Airline MH-370, they are actually using satellite images for the exploration of the earth, so that you can point out anything you feel fishy, in the required context, so that could led to similar analysis, that's what human computation and Crowd Sourcing can do together, such empowerment can be thought of breaking complex tasks into several units and let the human cognition get involved with the computational power to bring up miraculous results by miraculous it is meant the complex task in lesser time.

### *1.4 Linked Data*

Linked data is about publishing, connecting and correlating piece of data, information, and knowledge in such a way that one data could be linked to other in a uniform manner [in contrast to pure compilation of datasets, it is formed in a standardized way], to induct semantics in that piece of data, so that if interlinked it could be made more useful and more meaningful in that particular context, previously the data was meant to be more for Humans only, but to make it understandable, Linkable, comprehensive for computers too, the concept of Linked data came to an existence, so that when queried a computer can understand and respond more in an intelligent manner, and by algorithmic understanding, it is based on stake of RDFs [Resource description framework], HTTP [Hyper Text Transmission Protocol] and URIs [Uniform Resource Identifier]. Tim Berners Lee Founder of W3C [World Wide Web Consortium] gave 4 principles that a Linked data is expressed in are as follows:

- For the purpose of Denotation and referring the, URIs should be used
- Use of the HTTP URIs, so that searchers can look up or refer
- When someone looking up via URI, provide useful information regarding same, Via RDF and SPARQL like Web standards/technologies
- Inclusion of other and external URIs for the sake of discovery of new and more relevant information, He also states that it is not mandatory to have them always, but to have an exposure of full potential all rules must be fulfilled.

### *1.5 Linked Data*

As discussed in the problem taxonomy [5], in the dimension of accuracy and category of implicit relationships between attributes, there-exist a problem of multiple facts being stored or encoded in single attribute and is one of the highest occurring problems. While querying data I found out the problem that the birth place is occurring more than one against each entity usually, which is going to mess up if one needs to search for the birthplace of say, an entity X and the result tells you that X belongs to A, B, C and D as well, even if it is true as say for example A is a country, B is an State and C is a city, D is a town. When I queried for entities with predicate of birth place in SPARQL via the query mentioned below [Query 1], it got figured out, it is one of the most confusing thing, all true but still not satisfying the result set.

1) QUERY 1

```
SELECT * WHERE?s ?p ?o;

rdfs:label ?name;

rdfs:comment ?description.

FILTER (regex (?p,
"http://dbpedia.org/ontology/birthPlace","i"))

FILTER ( lang(?name) = "en" )

FILTER ( lang(?description) = "en" )

LIMIT 100
```

The Query 1. Line 6 filters out data, having attribute of Birthplace in them, Query 1. Line 7, Restricts the data i.e. which is available in English language only

## II. IMPLEMENTATION

For the proof of concept, I have chosen PHP, as the base language, the Online DBpedia's SPARQL query editor is chosen to query the record, for reducing the time of querying the data, the data that was returned as a result of the query was imported in a CSV Comma Separated Value file, that would be used for this experiment. After the import of the RDF format i.e. the triple format (Subject, Predicate, and Object) is imported the file is given to the application that would import that particular CSV file, store it in an array of size based on the total no. of rows of the imported CSV. After this process, that array is broken into the name of the entity and the birth place that entity belongs to like if ABC belongs to P, Q, R. it would be saved in 3 attributes, having attribute 1 as P, 2 as Q and 3 as R, after this the user is asked to tell out of P, Q, and R, who is the parent of who say for example P is parent of Q and Q is parent of R, so wherever in the data exist the Q it will get to know Q is parent of R and Child of P, even if an entity say XYZ does only have information of Q alone. It will also show you the summary of the no. of votes given to some Parent-Child Relationships, based on which the majority will decide it to be true, out of which the trust and confidence of that particular choice can be calculated, Fig. 1. Below shows the roadmap of the process

The maintenance of quality of data is undoubtedly time taking, labor-critical and domain specific job , so far available techniques focuses on automated solutions of improving data quality, which are uncertain and/or risky and without the response or feedback from the user, existing solutions have proposed a narrative framework for improving data quality via direct user interaction in the cleaning process to optimize available automatic technique whilst also ranking the repairs

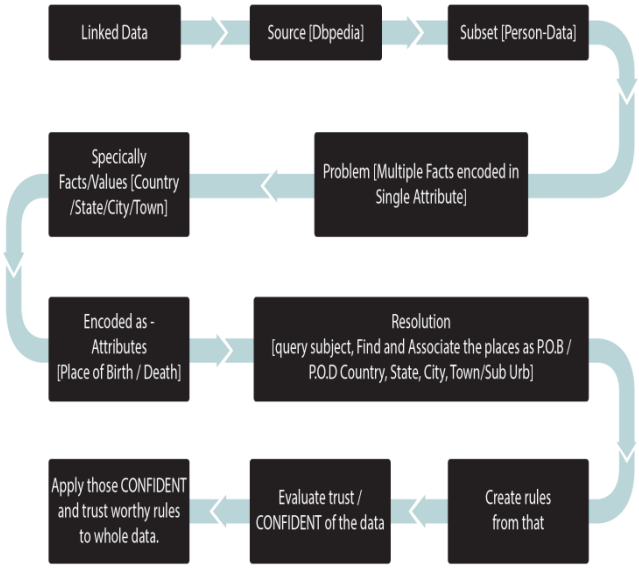


Fig. 1. Roadmap of this research.

or fixes having known the value of information, furthermore reducing user contribution as user response can be learnt by the machine through active learning by having a refined training set. [1, 2, 3]. The first phase to improve data quality is to set and define the Data Quality rules which could be in the form of database the DQR Db, the rules are defined as Constraints to data [4] Several tasks in Data interlinking with the help of human intelligence requires motivation and incentives for the sake of unbiased and interest full participation of the crowd [group of humans], for which the Games with purpose are famous for. For a data to be linked and published, the main steps are to assign consistent URIs, generating links and publishing Meta data [8] So far it has not been made to quantify the knowledge base of DBpedia in terms of quality, nor the improvement from one release to another, the approach proposed in here is comparison between the self-made best case scenario and the ontology based extraction from DBpedia, so that the precision of the framework which is used for the extraction purpose and the completeness of DBpedia knowledge base which is actually a subset of Wikipedia Knowledge base could be estimated in a quality assessment framework proposed herein [9] In [5] Zaveri and group has analyzed the quality issues in dbpedia that comprises of approximately 500 resources of linked data sets, they chose to analyze it by Manual i.e., a resource is selected based on User's choice completely random or from a particular class, later after this process the second phase is to select the mode of evaluation of the selected entity which are again of three types, the first one is the Manual in which the selected resource is assigned to an individual or a group of people to evaluate the selected class, if the user chose this to be in a semi-automated way an evaluation is done through a tool which later is validated by the user, this technique can also trigger machine learning if required, the third option is fully automatic in which the tool validates itself and no user is involved in this process, The process is mapped in the Figure 2.

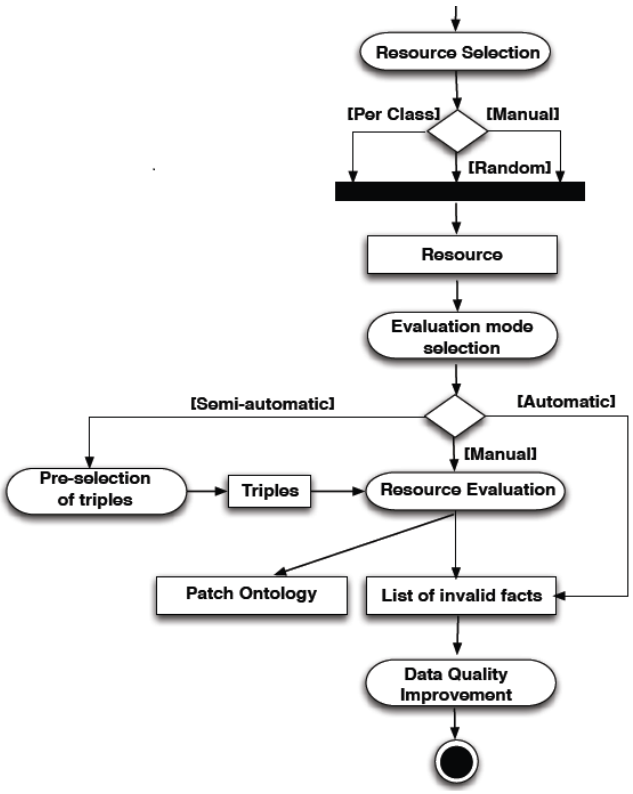


Fig. 2. Data quality assessment framework.

After this step the final step is the Quality Improvement phase which as per the [5] can be achieved in at least two ways either direct or indirect the direct here means that edit the triple right away with the correct value if confident or either by indirect in which the user feedback is gathered and based on the numerous feedbacks the data is made correct. Similar to this is my work; I have included the indirect mechanism by inducting the Confidence of the data by calculating the number of votes. The Figure 2 shows the workflow of the Quality Assessment of Linked data on dbpedia discussed in [5], The tool they are referring to is known as Triple Check Mate, that forces you to sign in using Google, you are given a random class and asked about that if you think it has some issues or is not correct or is correct, or need variation from the listed options or something you think is missing so can add there, later you are known with your Google ID. The motivation used for this was prizes / competitions. The Figure 2 shows the work flow of the work done by Zaveri. A, in [5].

## III. FUTURE WORK

### *3.1 Query Morphing*

As in [10], while the user is inputting the data he/she knows, meanwhile in the backyard let the processor contextualized the query in a manner that it may more be alike the keyword user is inputting rather than waiting for the user to complete the whole query and then hit the point user is pointing to if it exists or satisfies the conditions mentioned in the query,

leaving apart enough processing that could be working on the continuity of the user input, if the pre-computing thing is working at one end, let the other part work on what user is coming or going towards, but showing the user available options that could perhaps be the one he/she is looking for. Famous examples include Google Instant, that is, it shows up result while you are still writing it also works if you do not want to go with the term related to the words you have typed, you can always go with the term you want to search, write it up completely and then hit the go.

*3.2 Advancements in current approach*

Currently the scope of my work is limited to the Person Data in English language and the attribute of birth place of some entity present in the Person data. The RDF format of Linked data comprises of the 3 attributes, the subject the object and the predicate, the subject is the OBJECT in real world, the Predicate is the Name of the Property of the value, and object is the value of that particular property it corresponds with, In the next iteration I would enhance the features of my current model, it will have dynamicity in terms of Subject, Predicate and object, based on the knowledge of the user, it will contextualize the query by Query morphing and then deep dive into the class user is familiar with, be it Birth place, death place, or date of birth, or the person itself. Or the relationship itself. Currently my model has an static data set, which is given as an input, but could be altered as needed, but in future I will be working with it to be as dynamic as possible, another aspect that is assumed to be added is the Machine learning component, which could infer the learning part through the decisions being taken to improve Data Quality.

## CONCLUSIONS

*4.1 Roadmap of the process*

The high level architecture of the proposed approach works in this flow, We queried at DBpedia's Online SPARQL Query Editor, if the output is as we desired to have, we exported that data to the CSV format, we imported that CSV data to the PHP and in PHP, within an array, Cleaned it in such a way that it gets presentable to the user emphasized more towards cleaning of the non-utf8 characters, URL -decoded the entries than wherever the place of birth occurred for a single entity they were stored separately against corresponding person, they were gathered and shown to the user in such a way that a user can identify if he/she knows about that particular place that if the two shown entries are parent or child of the other i.e. they were asked to tell the relationship between them, if Pakistan and Karachi has occurred so user is to tell whether or not the Pakistan is parent of, or Karachi is parent of Karachi, Pakistan

Respectively. When the user selects the answer its count is taken as one, when a single choice has higher no. of votes for likely option, the greater no. of votes will decide if the user is right or wrong, as majority is authority we will bring up the results for those who voted more for a particular choice, and calculation of the trust and / or confidence of the result is calculated based on no. of votes for yes or no, divided by total, higher the votes, higher the chances of the data of being authentic. Finally it would also show up the summary of the total. Based on this we can identify if Pakistan is Parent of Karachi, and Clifton is child of Karachi, so we can show up the hierarchy of Clifton, that it is child of Karachi which is child of Sind which is child of Pakistan.

*4.2 Results*

Based on the above roadmap, the PHP application gathered unbiased results on random entities, shown the total no. of votes given to the particular entity for being Parent/Child of the other in the summary. It also Shows the hierarchy of a particular place to and from of that particular node i.e., all nodes preceding that and all nodes following that (off course if available), like Asia is PARENT OF Pakistan, and Pakistan is PARENT OF Sindh, and Sindh is PARENT OF Karachi, likewise save it in the form of triples i.e. with subject, predicate and the object, all having a common predicate of PARENT OF, i.e. the relationship of the subject (City/Country/etc) with object (State/City/etc.). one entity can be mapped with the other like ABC born in Pakistan would be same as PQR born in Pakistan as long as the Pakistan remains same both semantically and syntactically, this was made with the help of the concept hierarchy, i.e., every node is basically a child and/or parent of the other or some other node, if they are associated with each other, now if within the data anywhere if exists Sindh, one can figure out what level of the hierarchy does it exists on, and what is the preceding or following levels.

## GLOSSARY

*5.1 Amazon Mechanical Turk*

Turk was named after a famous 18th century's automatic machine which could play chess. This robot type thing was powered by clock's machinery and was a sensation of its time. Amazon's mechanical Turk on the contrary doesn't hide a human's mind inside a machine but makes a large number of humans a part of its machine namely the mechanical Turk. AMT is a portal, where one sees a list of TASKS, Search and browses those HIT (Human Intelligence tasks) which could be things like

1. Selecting suitable category for a product line perhaps from a list of Categories.
2. Telling if particular 10 products are the same Translation of the text into some language be it English/Arabic/Urdu and so on, Accept that challenge, work on it, by following the given instruction, when done, and submit your work, when the requester (One who have kept it as task) approves it your money is deposited into your Amazon Payment account. The sign up process takes up to 48

hours, and sometimes it is also denied which the AMT do not mention the reason as per their confidential review criteria. This website is growing at a nice rate and the work quality is good because people get paid to do the task which increases the chances of getting better results than working for free. Moreover it's under amazon's brand name which assures people that whether the task pays a couple of pennies or for a coffee, they will get the payment because Amazon is a famous brand which most likely won't cheat people by not paying them for their hard work. Unfortunately international workers can't signup to work for mechanical Turk because of violating amazon's terms of service. There is also an online forum http://goo.gl/sq6aNH with thousands of posts about how to break and violate Turk's terms of service and make multiple accounts etc. due to which international accounts were banned.

*5.2 Data Quality*

Data Quality cannot be defined explicitly, strictly or specifically by a single definition, quality can differ in different contexts, but a generic definition to data quality could be data that conforms to requirements, are of same meaning as referred, complete, totality of features and characteristics of factors relating the given data and consistent. There exists several other key points but to a generalized level this definition could be fulfilled. Following are the data quality attributes.

Table 1: Data quality attributes.

| | |
|---|---|
| Degree of excellence | Totality of features |
| State of completeness | Conformance to requirements |
| Validity | Conformance to acceptable criteria |
| Consistency | Completeness |
| Timeliness | Standardized |
| Accuracy | Time stamped |

## ACKNOWLEDGMENT

I wish to express my special thanks first of all to God, my family (because of whom I have been here), my colleague and my supervisor, Syed Saif ur Rahman who has been the ideal supervisor, with His sage advice, insightful criticisms, and patient encouragement aided the writing of this report and conduction this research work in innumerable ways. Specially, I would also like to thank my friends, and Mr. Zulfiqar. A. Memon IBM Digital Marketing and Analytics Consultant at Royal Cyber Inc. whose direct and indirect support in the development, creation and implementation of the demonstration used in this independent study, all became a cause of the creation of this project. Lastly, I offer my regards and blessings to all of those who supported me in any and every respect during the completion of this report.

## REFERENCES

[1] *Mohamed Yakout. Guided data quality improvement through direct/ indirect interaction. VLDB PhD Workshop, pages 24 – 29, 2010.*

[2] *Jennifer Neville, Mohamed Yakout, Ahmed K. Elmagarmid and Mourad Ouzzani. Gdr: A system for guided data repair. SIGMOD, pages 1223 – 1226, 2010.*

[3] *Mohamed Yakout, Jennifer Neville, Ahmed K. Elmagarmid, Mourad Ouzzani and Ihab F.Ilyas Guided data repair. PVLDB, pages 279 – 289, 2011.*

[4] *Jennifer Neville, Mohamed Yakout, Ahmed K. Elmagarmid. Ranking for data repairs. DB Rank workshop of ICDE, pages 23 – 28, 2010.*

[5] *Mohamed A. Sherif, Lorenz Buhmann, Mohamed Morsey, Soren Auer, Amrapali Zaveri, Dimitris Kontokostas and Jens Lehmann. User-driven quality evaluation of dbpedia. I-SEMANTICS, pages 97 – 104, 2013.*

[6] *Using crowdsourcing to search for flight mh 370 has both pluses and minuses. Available online http://qz.com/188270/using-crowdsourcingto-search-for-flight-mh-370-has-both-pluses-and-minuses/(accessed 31 March 2014).*

[7] *Application to search for missing aeroplane mh-370 using sattelite images. Available online http://www.tomnod.com/nod/challenge/mh370 -indian ocean/(accessed 31 March 2014).*

[8] *Katharina Siorpaes and Elena Simperl. Incentives, motivation, participation,games: Human computation for linked data. FIA, 2010.*

[9] *Paul Kreis. Design of a Quality Assessment Framework for the DBpedia Knowledge Base. PhD thesis, Free University of Berlin, 2011.*

[10] *Stefan Manegold, Erietta Liarou, Martin L. Kersten, Stratos Idreos. The researchers guide to the data deluge: Querying a scientific database in just a few seconds. PVLDB, pages 1474 – 1477, 2011.*