Enhancing Data Quality using Human Computation and Crowd Sourcing

Vikram Kumar Kirpalani[1], Muhammad Ejaz Tayab[2]

[1,2]*Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST) Karachi, Pakistan*

[1]vikram.kirpalani@gmail.com

[2]ejazshaikh@yahoo.com

Abstract—This paper is aimed at addressing the issues that are present in the data dumps available at DBpedia by using the concept of associations i.e. concept hierarchy to enhance the quality of those data dumps. These data dumps are extracted from Wikipedia and the issues that prevail in these data dumps is because of either the data extraction frameworks or the human error during crowd-sourcing efforts made on Wikipedia. By using Human Computation techniques and employing Crowd sourcing together with query morphing, diving deeper into this subject would become easier in a better way. One of the key issues with the datasets is the presence of multiple values in a single attribute and vice versa especially in the "Place of Birth" field of important personalities. This paper highlights the implementation process in order to solve these issues and adds a survey conducted on Crowd Sourcing to highlight its impact.

I. INTRODUCTION

Zaveri et al [1] identified various issues in the DBpedia data dumps which are a result of using automation approaches of data extraction. Moreover, the transformation of unstructured data to structured form also adds these issues in the data dumps. Among those issues, the author had chosen to evaluate and solve the issue of the presence of multiple values in a single attribute and vice versa in the previous paper [2] using Human Computation techniques. This paper is an extension to previous paper and here, author have aimed to solve the same issue in a better way by using additional techniques namely Crowd Sourcing and Query Morphing in addition to the previous technique.

In addition to solving data quality issues, author has highlighted the importance of the method of Crowd Sourcing to solve these issues through a survey. This would help to identify why Crowd Sourcing is an important solution and how it can help to solve these issues. By using Crowd Sourcing, a crowd or a group of people will have access to the unimproved data dumps and their combined effort and input would improve the data quality. In addition to that, the element of Query Morphing is used which would help people to identify and solve the issues in such data dumps regarding which they are knowledgeable. In the survey, author have also tried to find the impact of freelancing on crowdsourcing i.e. whether providing incentive to the crowd would result in better quality of work or not. Additionally, author have also tried to find out whether crowd sourcing is a cost effective in solving small computation tasks as compared to hiring programmers to code sophisticated programs to solve those issues. Examples of tasks that can be solved using crowd are translation, image labeling, object detection, brain storming, and etc. The focus of his research is on the attribute of Place of Birth (POB). At DBpedia, data dumps are available in the form of RDF (Resource Description Framework). So, if a person's POB is Town A, City B, Province C and Country D, such information is not mentioned on Wikipedia together with Town, City, Province and Country prefix. It directly mentions the POB as A, B, C, D and people can read this information to identify which country does a person belongs to. In DBpedia, this four or more detail of the POB field is mentioned as four or more separate attributes of POB. For a particular person mentioned above, it would list four POB attributes as A, POB as B, POB as C and POB as D. When developers use DBpedia data sets in their applications, they would face this trouble of identifying the actual country amongst all those POB attributes. In previous paper, author solved this problem using human computation in which people could spot these various POB attributes and identify the parent and child relationship. For example, country D is the parent of province C which is the parent of city B, and which is the parent of area A. In this paper, author attempts to solve the same issue using crowd sourcing and query morphing in addition to previous techniques. [1, 3, 4, 5] have looked at similar issues from bird's eye perception while other researchers [6] have identified the same sort of issues that are addressed in this paper. Query morphing and the aspect of contextualizing the user input on every step as adding variation to the base query has also been identified as one of the solution techniques [7]. An important aspect of developing a crowd sourcing system is to add to its learning mechanism by using confidence parameter. In this way, if a user marks a parameter X as country name and any other user identifies parameter X as city name, then the system needs to confirm for the correct answer by asking the same question to several other users to increase the confidence of a particular answer. Then, if a machine feels that the answer provided by the user is out of context, it can notify the user in an automated manner [6].

A. Data Quality

The term quality has several attributes and definitions so it is not suitable to give a single line definition to define the term data quality. Every person would define data quality in a different way according to the parameters of quality which feels important to them. For instance, some people compare data with time and believe that if a data arrives in an untimely manner, it would be of low quality. A good definition that the author came across defines data quality as data is of high or good quality if it is conforming to requirements at a given time. So, according to the definition, if time is of importance and if someone wishes to get the data on time, it must arrive by that time. If it arrives late, even though it would be of high quality but it is useless. Similarly, if quality is of importance, then the data must be of high quality no matter how late it arrives. If it is of low quality then no matter how early it arrives, it would be useless. The table 1 below shows data quality attributes and characteristics of all the factors that are linked to the quality of data.

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

B. Human Computation

Human Computation can also be termed as human based computation because this process involves input from human beings for the purpose of computation. In computation, the need for computer is still required for the application to run but the input from human beings saves tons of lines of code. This code would have to be written in order to identify data quality attributes which human beings can easily identify by using their cognitive and mental power and save both cost and time. For example, consider a case where a computer is presented with an image of a pyramid. Computer might initially recognize that object as a triangle but if a human is asked to recognize that image, he would immediately recognize it as a pyramid. Making the computer detect the same image would take a huge amount of code and several machine learning principles for training and testing. Another example can be of a situation, if a computer is presented with an image of a cat and a dog running then a computer may be able to recognize these two animals as a cat and a dog but the action that they are performing would not be recognized by it. In this scenario, if a human is presented to identify what activity is being performed by those two animals then a human can easily identify and mark that they are running. Similar to the examples above, other tasks that can be performed with human computation are brain storming, translation, image labelling, and etc. A good example where the human computation is being used online is reCAPTCHA. This method was used to differentiate bots from humans while filling an online form or other sorts of registration processes because this method did not allowed automatic scripts to make multiple accounts on websites by signing up dozens of forms automatically. ReCAPTCHA solved two issues simultaneously; that are, preventing bots from signing up and to convert hand written or old book manuscripts to digital form, which the standard OCR procedures failed to perform. So using reCAPTCHA, this task was performed by using small portion of those manuscripts at the end of a signup form which humans could easily identify and input the characters presents in that small manuscript portion. Confidence rating of a particular word could be tested using same image multiple times on different forms which would help assure that the word has been transformed correctly.

C. Person Data and the Data Quality Taxonomy

In this paper, the main focus is on improving the place of birth attribute of person's data set present on Wikipedia data dumps i.e. Birth place, data of birth, etc. and if a person is deceased then date and place of death. To narrow down the scope and proving the concept specifically, person's data set has been selected for improvement in this paper. Data dumps from DBpedia were used in English language and were present in TTL (terse turtle language) which is an extension of rdf. Due the memory issues faced on implementation of the system in previous paper, author decided to use an online query editor for the implementation used in this paper. This online query editor assists in querying the data which is present online on DBpedia. In this way, the complete data dump does not have to be downloaded. This would save downloading time and system restrictions which would slow down the querying process on those large data dumps. This service is free of charge and can be accessed at http://dbpedia.org/sparql. To confine the results to place of birth attribute, the query used in the query editor was:

```
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 concerned issue here is related to the dimensions of accuracy and implicit relationship. This was one of the most redundant issues as recognized in previous paper because the place of birth attribute is one attribute that keep on reoccurring in a single entity. So if someone queries place of birth of a particular person, they will see multiple place of birth attributes for that same person which would make it harder for a person to identify the actual place of birth.

D. Crowd Sourcing

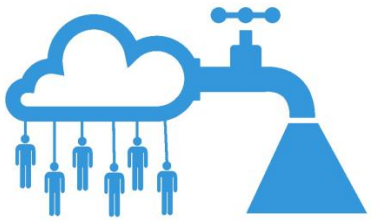

Fig. (1). Crowd Sourcing

The above figure 1 depicts the term crowd sourcing in the best possible way i.e. minds of people working together as a crowd to provide input into the system with the goal that the output produced by the system would of use to common people. This term was coined back in 2011 with the aim to solve several micro tasks which were sub tasks of a bigger task and to involve people in solving those tasks such that the parallel input from people would solve it in an efficient way. This process is being used in a wide variety of different ways. Crowd sourcing was used to find the Malaysia flight MH-370 which was lost during its flight course and could not be located by the radar. Satellite images of several places were shared online to people over the internet for all possible routes that the plane could have taken. This approach helped because people were able to find minor clues in those images and report if they appeared to be related to the plane [8]. Another similar system is Tomnod [9], a portal that helps people in solving real world problems such as searching for a lost plane or a planet. Some queries require human input for better manipulations and that input is provided in terms of crowd sourcing [10]. Similarly, another important innovation which involves similar principle is Amazon mechanical Turk AMT which provides an infrastructure and an open platform for people to perform such small tasks which are too expensive for a computer to solve and human input has better chances of success in those tasks. Tasks like games with purpose, image labelling, segmentation of images, or tagging random images are one of the most important areas in which crowd sourcing is being used. Motivation and incentives would be required so a large quantity of people could make such crowd sourcing tasks possible.

E. Query Morphing

Query morphing was one of the proposed future works in previous paper [2] which was planned to be used in this paper. The term query morphing is used in the same context as used in terms of image processing. It is a relatively new term given to morph-age of a query at a given time. The best example which author can use to explain this term is to consider google search engine. Whenever a person visits the search engine to find some information and as soon as the person begins typing a few letters, the engine starts giving relevant predictions to help user and attempts to guess what the user wills to search at that time. So, if someone type the first letter for example "S" then it would give suggestions based on that letter and as soon as someone begin typing further letters to complete word, the search engine starts predicting and attempts to complete the search term by suggesting multiple suggestions to assist a user. So as soon as a character gets input, it tries to contextualize in every respect and keeps on suggesting as the user enters the input letters or characters. This process is a real example of query morphing. When the user gives a starting query X, then the query is used syntactically to create some variations in result R as in R1. . . .RN [7].

II. STUDY WORK FLOW

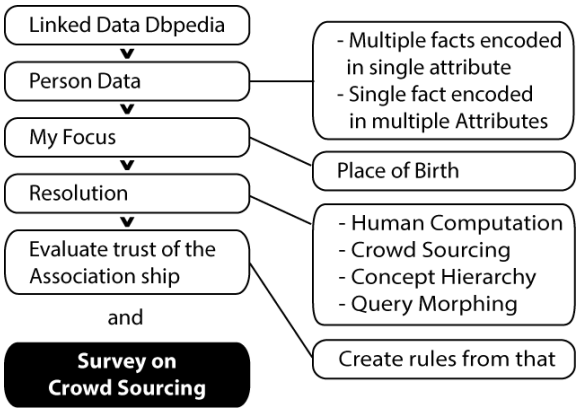

Fig. (2). Flowchart

The figure 2 highlights the steps followed in this paper. The idea is to experiment with the Linked Data that is available at DBpedia and to use and analyze it specifically as far as the person's data is concerned in English language. The main reason to choose this particular attribute apart from the fact that it is present in redundant form on DBpedia is that this is easily understandable to common man. A person would visit Wikipedia regularly to search information regarding various subjects available there and if; have ever searched for a particular person, would be well aware of the date of birth and place of birth field and how that data is represented on the website. So using concept hierarchy and adding crowd sourcing and human computation together with query morphing, author have tried to solve the issue of the presence of multiple attributes in a single field specifically the place of birth field. Once the crowd has contributed their knowledge to identify the correct attribute, then as an enhancement process, evaluation of trust of that particular answer can be performed by rechecking the answer to the same question and comparing it with different results which have been provided by the crowd. In addition to that, author have also performed a survey on crowd sourcing to highlight its importance and impact on people which helped to arrive at the conclusion that it is preferred by the industry as an excellent method to solve the issues that have been addressed.

A. Implementation

For the implementation in this paper, author chose php as the primary language for demonstrating the system mainly because querying with Jena tool consumed too much memory and respond late. Since, it was not feasible to do it on a single machine, author decided to use online query editor sparql that is available online and generated comma separated values (csv) file. The file in MySQL database was imported and then php was used on server side to create the platform that will be available to the crowd to take their input. The crowd would first get registered on the portal and then they can use their input to create associations between the data sets which has to be improved whether by searching (Contextualize search based on query morphing) or by choosing a value from the drop down field or assigning a random class data to them [1].

B. Survey

The survey conducted mainly includes questions related to crowd sourcing and how it accomplishes better results, its impact on the industry and how important it is according to the people. Following section shows the questions that were asked in the survey and the responses that were obtained in the survey are highlighted in the figure 3 to 14. The survey questionnaire was distributed to friends and colleagues working in different fields in IT industry, freelancing or working on some extents of crowd sourcing. Explanation of various terms was made clear to the people so that they could answer properly.

Crowd Sourcing and its Impact: The questions asked, the choices and their responses are presented below statistically. Around 150 questionnaires were distributed out of which response from 52 questionnaires were received including one of author's response as well.

1) Gender

a) Male (47)
b) Female (05)

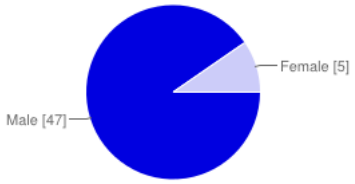

Fig. (3). Gender

2) Age

a) 18-24(12)
b) 25-30(28)
c) 30-35(07)
d) Greater than 35(05)

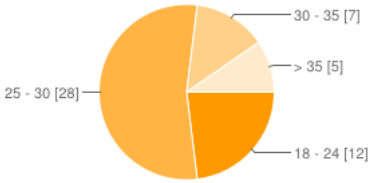

Fig. (4). Age

3) What is your Highest or Ongoing Qualification?

a) Undergraduate (7)
b) Graduate (15)
c) Masters/M.Phil./PhD (26)
d) Industry professional (04)

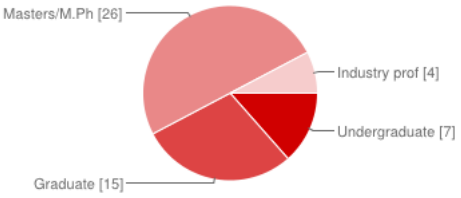

Fig. (5). Qualification

4) Exposure to Freelancing world

a) Yes I know a bit (28)
b) I am strongly familiar with it (21)
c) No, I don't know (3)

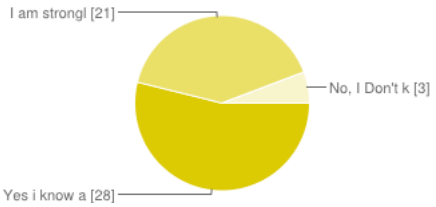

Fig. (6). Exposure

5) How do you think, Crowdsourcing (Freelancing) can help in tasks like Image Labelling, Brainstorming and Translation?

a) Really Helpful (23)
b) Useless (1)
c) Could bring value to save computation.(7)
d) These tasks could be done better by a computer (4)
e) A group of people can do it in a far easier and efficient manner (17)

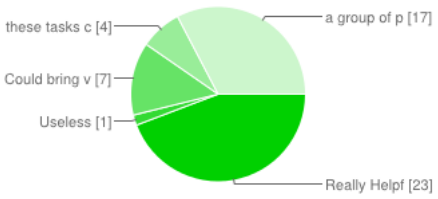

Fig. (7). Opinion

6) What impact could Crowdsourcing have on the Software Industry?

a) Positive (48)
b) Negative (2)
c) No impact at all (2)

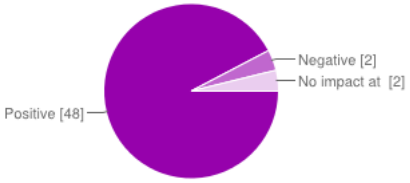

Fig. (8). Impact

7) Do you think Crowd sourcing can help a complex and tedious problem to be solved quickly with the help of crowd (Group of people) or do you think too many cooks spoil the food?

a) It benefits (44)
b) It is a disadvantage of crowdsourcing (08)

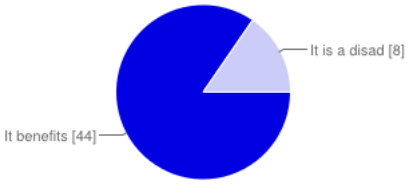

Fig. (9). Benefits

8) Main drawback of crowdsourcing is lack of quality control. Especially with microwork, crowdsourcing tends to be of a very low quality (it's mostly suited towards work that does not have to be done accurately but simply needs to get done).

a) Agree (40)
b) Disagree (12)

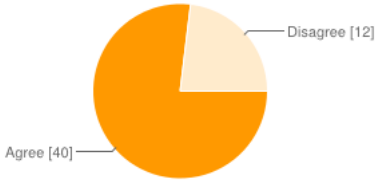

Fig. (10). Drawbacks

9) Schedules of your interests, Incentives and Motivation in crowdsourcing is relatively lower than of Freelancing, Do you think if this is taken care of can lead to better results?

a) Absolutely, it should (41)
b) The quality of work has nothing to do with motivation and incentives (07)
c) No, it won't affect the quality of work (04)

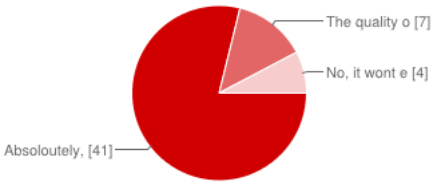

Fig. (11). Interests, Incentives and Motivation

10) Did you notice this survey filled crowd sourced data. Many people can input online to this survey. How do you found crowd sourcing? (Rate it on a scale of 5, where 5 being the best)

a) 1. (04)
b) 2. (04)
c) 3. (15)
d) 4. (13)
e) 5. (16)

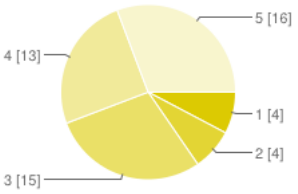

Fig. (12). Rating

11) Do you think Crowd Sourcing divide a Complex task into multiple sub-tasks and it can be Time and Cost Savvy?

a) Yes
b) No
c) Other

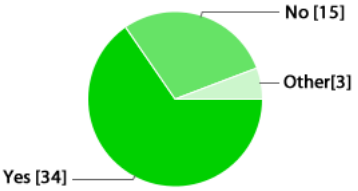

Fig. (13). Time and Cost saving

Following were the answers in the other category

a) Depends on the kind of people, because if they are really into solving something they are going to probably put into their best effort, but if they are doing it for money then there are various people issues that need to be taken into account.

b) It's more pain, especially when it can be done by a single person, (skill full resource)

c) Depend upon nature of work.

12) Do you think Human-Based Computation i.e. Human Computation is a better way to get complex problems solved rather than writing a code specifically for something that a computer finds complex to do. (Tasks like, Image Labeling, translation etc.)

a) One of the preferred mechanisms (43)
b) No, not preferred (09)

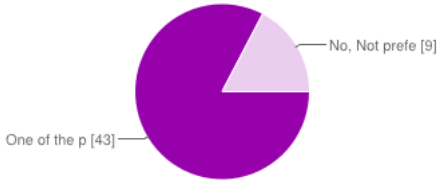

Fig. (14). Time and Cost saving

III. FUTURE WORK

For future work, author would consider the issues that were faced during the implementation process of the current system and would try to address areas where the lacking was found. Regarding the major concerns that were identified in this paper relating to the redundancy of information present in place of birth attribute, author suggest that the redundant information should be erased. So if there is a country X and it is being repeated for more than one attribute then it should be only considered once rather than considering it every time that it is repeated. This can be achieved by assigning it a unique id throughout the system. Apart from this query morphing and contextualizing of the user, input could be made more interesting and broad in terms of diving deep into the context i.e., it should produce random results during implementation rather than the recommendations that are made in the current system. Crowd sourcing should be accompanied by motivation and incentives [11] so it should be presented as an opportunity to work for the user. This can be done by either introducing a payment mechanics as a reward or by introducing the phenomena of games with purpose so the user keeps feeling the motivation with the rewards being awarded in the game. In addition to this, the user won't feel the boredom and demotivation by doing repeated tasks

IV. RESULTS AND DISCUSSION

The conducted survey helped author in realizing the importance and advantages of crowd sourcing and the likelihood of people's interest in being a part of the crowd sourcing system. 90% of the people who filled the survey responded in favor of crowd sourcing and 10% selected other options but none of the people responded against the use of crowd sourcing. Apart from the survey the demonstration was made available on the link at http://digitalme.pk/projects/vikramnew/

The implementation system clearly managed to enhance the data quality by using triplets in the form of structured questions. The crowd was given the option to either answer the questions of interest of individual users by assigning a random class. To narrow the scope, only the attribute of birth place was selected. This helped in implement human computation, query morphing and crowd sourcing in a short period of time duration for this research. The platform implemented was not very vast in scope but author manage to accomplish the main purpose. In addition to that, author also included the implementation for the element of trust i.e. checking the confidence or trust of a particular answer for a question by asking the same question to different users and observing their response.

V. CONCLUSIONS

In this paper, author was able to improve and enhance the data quality using an implementation which was designed to cater the issues that were explained in previous portion of the paper. The scope was limited to a subset of data i.e. the place of birth attribute from DBpedia data set of renowned personalities was considered. In addition to that, the survey results clearly identify crowd sourcing and human computation as the key players in solving problem regarding the quality of data. Moreover, the survey highlights the main reason for lacking of such systems are factors like motivation and incentives and if such systems could be made similar to freelancing then they would reach a wide array of audience. Very few people have access to such crowd sourcing platform so if such issues are solved then these technologies would grow at an enormous rate.

VI. ACKNOWLEDGMENTS

I would like to thank God, My Family, Friends and my Supervisor Mr. Muhammad Ejaz Tayab. Their constant support & valuable feedback led the base of this paper. Lastly, Mr. Zulfiqar Memon, Digital Marketing Consultant at Royal Cyber, who helped implement the demonstration for this paper.

REFERENCES

[1] Zaveri, D. Kontokostas, M. A. Sherif, L. Buhmann M. Morsey, S. Auer and J. Lehmann, "User-driven quality evaluation of dbpedia". In *Proceedings of the 9th International Conference on Semantic Systems (I-SEMANTICS '13)*, 2013, pp: 97-104.

[2] V. Kirpalani and S. Saif ur Rehman, "Improving data quality using techniques from human computation," *Journal of Independent Study and Research*, vol. 12, no. 2, May 2014.

[3] Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann and S. Auer, "Quality assessment methodologies for linked open data. A survey," unpublished.

[4] P. Kreis. "Design of a Quality Assessment Framework for the DBpedia Knowledge Base". Ph.D. dissertation, Free Univ. Berlin, Germany, 2011.

[5] M. Yakout, A. K. Elmagarmid, J. Neville and M. Ouzzani, "Gdr: A system for guided data repair". In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (SIGMOD '10)*, 2010, pp: 1223-1226.

[6] M. Yakout, A. K. Elmagarmid, J. Neville, M. Ouzzani and I. F. Ilyas, "Guided data repair," In *Proceedings of the VLDB Endowment*, 2011, vol. 4, no. 5, pp: 279-289.

[7] M. L. Kersten, S. Idreos, S. Manegold and E. Liarou, "The researcher's guide to the data deluge: Querying a scientific database in just a few seconds". In *Proceedings of the Very Large Databases Endowment (PVLDB)*, 2011. vol. 4, no. 12, pp: 1474-1477.

[8] Merelli (2014, March 15). *Using crowdsourcing to search for flight MH 370 has both pluses and minuses* [Online]. Available: http://qz.com/188270/using-crowdsourcing-to-search-for-flight-mh-370-has-both-pluses-and-minuses/

[9] Tomnod, [Online], Available: http://www.tomnod.com/nod/challenge/mh370indianocean

[10] M. J. Franklin, D. Kossman, T. Kraska, S. Ramesh and R. Xin, "Crowddb: Answering queries with crowdsourcing," In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data (SIGMOD '11),* 2011.

[11] K. Siorpaes and E. Simperl. "Incentives, motivation, participation, games: Human computation for linked data," In *CEUR Proceedings of the Workshop on Linked Data in the Future Internet at the Future Internet Assembly*, 2010.