Distance Based Measurement Approach for Truth Discovery by Resolving the Conflicts in Big Data

P BastinThiyagaraj, A Aloysius

Abstract. Big data is a term that describes volume of data (terabytes to Exabyte’s), unstructured (include text and multimedia content), and complex in processing (from Medical data, Business transactions, Data captured by sensors, Social media/networks, Banking, Marketing, Government data, etc.). The traditional technologies are not sufficient to store, process and analyze the data. The unique technologies should be needed to analyze, manage the huge amount and unprocessed data. The number of sources produce huge amount of various descriptions for same object. This leads to data conflict and source conflict, when various sources generate various descriptions for same objects. Here it is the challenging one to identify which source produces quality information and which data is truly fit for an object. Here the heterogeneous data involved such as both numerical data (measurement data) and string data (classified data). The Data Analytics plays an important role to analyze the conflicted data. Distance based approach is used to find highest achievable performance, by minimizing the distance and maximizing the reliability of the sources. The main objective of this work is to resolve the conflicts from the heterogeneous data and identify the true information among the conflicted data from the various sources. Here, the continuous data only taken into an account to identify the true information.

Keywords: Big data, big data analytics, Reliability, Accuracy, Square loss distance.

1. Introduction

Big data is a term that implies huge amount of data from various sources. According to the definition of Gartner “Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making”[1]. The huge amount of structured, semi structured and unstructured data are generated from the various sources with different types of data. For example, every day Face book generates 500 terabytes of data (including uploaded photos, likes, and users’ posts) [2]. According to International Data Corporation’s “Digital Universe” forecasts 40 ZB of data will be generated by 2020[3]. Similarly number of sources produce amount of various descriptions for the same object. This situation leads to the data conflicts. For example Google for the query like “the height of Mount Everest” includes “29,035 feet”, “29,002 feet” and “29,029feet”[4]. It is very tedious task to analyze large volumes of data with varieties of conflicted data from various sources. Data variety increases with various branches of science and societal systems [5].

Here, in this paper we focus on identifying quality of source with reliability and true information by resolving the conflicts by using Square loss distance approach which is a statistical method. It is crucial to estimate source reliability to find out the correct information from conflicting data when the sources providing low quality information, such as faulty sensors that produce wrong data, and spam [6]. The rest of this paper is organized as follows: In the next section related works of my work is
2. Related Works
The World Wide Web has an important role to provide the quality of information for the most people. But we could not confirm that all the web sites are providing Quality information. Some of the websites provide conflict information. This is due to recording or transmission errors, device malfunction, or malicious intent to manipulate the data. Data sources usually contain noisy, out-dated, missing or erroneous records. Thus multiple sources may provide conflicting information [6].

In [8], the truth information identified by using “TRUTHFINDER” method. They found to improve truth discovery by detecting dependence between sources and analyzing accuracy of sources. The Bayesian model is used to discover copiers by analyzing values shared between sources.

The Statistical machine learning based on mathematical models and powerful algorithms, techniques such as Bayesian networks, Hidden Markov models, support vector machine, reinforcement learning, and ensemble models, has been applied to data, text, and web analytics applications [7].

In [9] Single truth identification method is explored. The algorithms Cosine, 2-Estimates and 3-Estimates are experimentally used in synthetic and real world data to identify truth information. 3 – Estimates predicted better results compare with other two algorithms. In [10] the truth information could be identified by experimentally over three domains which are city population, basic biographies, and American vs. British spelling with four datasets. All domains are using the methods Vote Distance cost function and Vote Loss vote redistribution.

In (6) Maximum Likelihood method is to find out the truth discovery from “social Sensing data”. Regular EM (Expectation Maximization) is proposed for crowd/social sensing applications. They found that “non-trivial estimation accuracy improvements can be achieved by the proposed maximum likelihood estimation approach compared to other state of the art solutions”. In [11] they explained that human as a “sensor network” that was identified twitters resources were the human resources. Here humans provided the information voluntarily or re-tweet the information.

In [12], heterogeneity of data has been used to identify the truth by resolving the conflict. Normally truth identification involves only on categorical or continuous types of data. However, the previous approaches are designed only for single-type of data and they do not take advantage with heterogeneous types. Different in loss function used along with CRH (Conflict Resolution on Heterogeneous Data) to identify the characteristics of data.

3. Methodology
Here the conflict resolution problem is solved by identifying the reliability (source weight) of the sources from the various conflicted sources. In existing system they resolved conflict resolution problem by identifying source weight of each sources using optimization techniques. If source weight is high, it produces most reliable trustworthy observation from the conflicted multi sources of data. Here, square loss distance method is used to identify the distance between observed values and truth values in our proposed method.

3.1 Problem identification:
Object: An object is a person or thing or an entity.
Property: property is used to describe the object.
Source: information provider. For example, “Bob” is an object; “height” is a property; and a database that provides the information is a source.

Observed value: is the data describing a property of an object from a source. For example the observation on Bob’s height from Source 1 is 1.74m.

True value: The truth of an entry is defined as its accurate value. For example, the real height of Bob is the truth of the entry.

Input: Suppose there are K sources (X=X₁, X₂, ..., Xᵦ), N objects (O=o₁, o₂, ..., oₙ) and M properties (P=p₁, p₂, ..., pₘ) whose data types can be different, and these objects are observed by K sources. The observation of the mᵗʰ property for the nᵗʰ object made by the kᵗʰ source is  vₘₙ(k).

3.2 Framework
In our system the highest weight of a source produces most reliable trustworthy information using distance based methods. Based on this principle our system working better to resolve the conflict from various sources and identify the trustworthy information from the conflicted data. Let us consider the framework for identifying truth discovery by resolving conflict using Square loss distance method.

Figure 1: Distance based approach

An algorithm given below, explains to identify the true information by identifying the reliability based on the distance approach.
Algorithm : Distance based Approach

**Input:** Data from K sources (X₁, X², X³... Xᵏ)

**Output:** Truth valus X*

Initialize the truth value (x*) // Mean values from the observed values

for n=1 to N do
  for m=1 to M do
    Update the distance of the n-th object on the m-th property
    \[ v_{nm}^{(\ast)} \]
    Using \[ d_m(v_{nm}^{(\ast)}, v_{nm}^{(k)}) = \frac{(v_{nm}^{(\ast)} - v_{nm}^{(k)})^2}{\text{std}(v_{nm}^{(1)}, ..., v_{nm}^{(k)})} \]

If X₁ is minimum distance then
  MdX₁=MdX₁+1
else if X₂ is minimum distance then
  MdX₂=MdX₂+1
else if X₃ is minimum distance then
  MdX₃=MdX₃+1
else
  MdXₙ=MdXₙ+1
end for
end for

Calculate Reliability of sources (from total minimum distance) and Accuracy.

Return truth value X*.

Here, MdXₙ indicates that the variable to store the total number of minimum distance values among the sources of the properties.

### 3.3 Reliability Estimation (Source Weight)

The k-th source X^{(k)} is the collection of observations made on all the objects by the k-th source. It is denoted as a matrix whose nm-th entry is \( v_{nm}^{(k)} \). \( X^{(1)}, X^{(2)}, ..., X^{(k)} \) are the K source observation tables. Actually the sample data about the citizenship given below taken from [6], project the conflicted information from various sources. The methodology can be developed by using this sample data.

| Object | X1   | X2   | X3   |
|--------|------|------|------|
| Bob    | 1.72 | NYC  | 1.9  |
| Mary   | 1.62 | LA   | 1.85 |
| Kate   | 1.74 | NYC  | 1.65 |
| Mike   | 1.72 | LA   | 1.85 |
| Joe    | 1.72 | NYC  | 1.85 |

| Object | Voting/Average Height |
|--------|------------------------|
| Bob    | 1.77                   |
| Mary   | 1.69                   |
| Kate   | 1.7                    |
| Mike   | 1.76                   |
| Joe    | 1.76                   |

The above tables are the Observation table with observed values from the various sources and truth table. Truth values can be identified by calculating the mean of the conflicted data from continuous data. Tables have produced conflicted values both Categorical values and Measurement values (continuous data). Here the continuous data are taken to identify the reliability of every source. The source weight is calculated by identifying reliability. In order to find the reliability of the sources, the
distance can be found between observed values and Voting/average values. There are statistical methods which are available to find the distance between two numerical values.

The loss function is Bregman divergence includes a variety of loss functions such as squared loss, logistic loss, ItakuraSaito distance, squared Euclidean distance, Mahalanobis distance, KL-divergence and generalized I-divergence [6]. The Square loss distance function is used to find out distance between true value and observed values from various sources. Any loss function can be used to identify the distance between the observed values and true values. Now let us consider to mitigate the effect of outliers, we can use Square loss distance is on continuous data:

\[
d_m(v_{nm}^{(*)}, v_{nm}^{(k)}) = \frac{(v_{nm}^{(*)} - v_{nm}^{(k)})^2}{\text{std}(v_{nm}^{(*)}, \ldots, v_{nm}^{(k)})}
\]

Where,

\(d_m\) is the distance function, measures the difference between the information of the sources \(v_{nm}^{(k)}\) and identified truths \(v_{nm}^{(*)}\) depends on data type of \(m\)-th property. \(v_{nm}^{(k)}\) is the observed values of the objects provided by the various sources. \(v_{nm}^{(*)}\) is an identified truths. From the above table the values, \(v_{nm}^{(*)}=1.77\) from truth table and \(v_{nm}^{(k)}=1.72\) from observation table can be applied in the equation 1 to find the minimum distance. Similarly we can find the distance for all the above observed value with truth value.

### Table 3: Distance between observed value and truth value

| Object | X1  | X2  | X3  |
|--------|-----|-----|-----|
|        | Distance | Distance | Distance |
| Bob    | 0.022696 | 0.044484 | 0.153425 |
| Mary   | 0.036091 | 0.047139 | 0.188555 |
| Kate   | 0.033857 | 0.008464 | 0.052901 |
| Mike   | 0.019645 | 0.044201 | 0.099453 |
| Joe    | 0.020486 | 0.032009 | 0.10371 |

Here it is the notable one that the smallest distance of the objects of the sources produces approximate values and Maximum weight, which are close to the truth values. From the table 3, the source X1 produces the reliable information compare with sources X2 and X3. The reliability degrees of the sources are,

\[X1 = 4/5*100=80\%;\]
\[X2=1/5*100=20\%;\]
\[X3=0\%;\]

The source X1 produces 80% of reliable information, X2 produces 20% of reliable information, and Source X3 0% of reliable information. Here in source X3 no value is close to the truth value.

### 3.4 Truth Identification

It is important to identify the true value of the objects from the conflicted data using reliability of the sources. Let us consider an example,
Example 1: Consider Kate’s height from Table 1. The observations from Sources 1, 2, and 3 are 1.74, 1.72, and 1.65 respectively. Suppose the weights are 80%, 20%, and 0% for Sources 1, 2, and 3 respectively. Then Kate’s height can be calculated by truth vector as:

\[
\frac{1.74 \times 80 + 1.72 \times 20 + 1.65 \times 0}{80 + 20 + 0} = 1.736
\]

Example 2: Consider Bob’s height from Table 1. The observations from Sources 1, 2, and 3 are 1.72, 1.7, and 1.9 respectively. Then Bob’s height can be calculated as:

\[
\frac{1.72 \times 80 + 1.7 \times 20 + 1.9 \times 0}{80 + 20 + 0} = 1.716
\]

Similarly, Joe, Bob and Mary have 1.718, 1.716, 1.618 which are closest to the observed values of the source X1. It is obvious that the results are closer towards the observation from Source 1 because it is more reliable. Since, the source X1 produces approximate results than X2, and X3. Similarly the other objects Mary, Mike, Joe have the heights are closest to the height of the source X1 due to its reliability is high. So that the source X1 produces reliable values.

4. Accuracy Estimation

It is notable that to estimate the Accuracy of each property (observed values and Truth values) of an object produced by every sources. The percentage of the accuracy of an attribute of an object is identified by estimating the error percentages between the observed values and Truth values.

\[
Accuracy = 100 - \left( \frac{\left( v_{nm}^{(c)} - v_{nm}^{(k)} \right)^2}{\text{std}(v_{nm}^{(1)}, \ldots, v_{nm}^{(k)})} \right) \times 100
\]

The above equation 2 is used to find the error percentage and accuracy of the observed values from various sources. Also the average accuracy of the source X1 is high compare with other sources.

| Object | X1       | X2       | X3       |
|--------|----------|----------|----------|
| Bob    | 98.72    | 97.49    | 91.33    |
| Mary   | 97.86    | 97.21    | 88.84    |
| Kate   | 98.01    | 99.5     | 96.89    |
| Mike   | 98.88    | 97.49    | 94.35    |
| Joe    | 98.84    | 98.18    | 94.11    |
| Average| 98.46    | 97.974   | 93.104   |

From the table 4, it is clear that those observed values have minimum distance with truth values, are having maximum accuracy or weight, and the above chart, clears that a source X1 has the highest accuracy when the reliability is high.
5. Result and Discussion

In this paper we focused on the distance based measurement to identify the true information from the conflicted data from the various sources X1, X2, and X3. From the above discussion Source X1 produce reliable information and produced better accuracy compare with other sources. Source X1 produces 80 percentage of reliable information and Source X2 produces only 20 percentage of reliable information. Whereas the source X3 produce zero percentage of reliable information. The source 1 produces minimum distance with true value as well as maximum weight (accuracy) compare with other values.

6. Conclusion and Future Direction

The paper is about the distance based measurements to identify the conflicts from the continuous data only produced from the various sources. The quality element “reliability” involved in identifying weight of the sources in order to identify the truth information from the conflicted data. It leads to enhance the quality in resolving conflicts. This concept is to be applied in big data hadoop system with real data set provided by the different sources to identify the true information from the conflicted data.

References

[1] Amir Gandomi and MurtazaHaider “Beyond the hype: Big data Concepts, Methods and analytics”, *International Journal of Information Management (IJIIM) ELSEVIER*, 2015, pp: 137-144.

[2] Provost, F., & Fawcett, T. “Data science and its relationship to big data and data driven decision making”, *Big Data, 1*(1), (2013), 51–59.

[3] Cai. L and Zhu.Y,“The Challenges of Data Quality and Data Quality Assessment in the Big Data Era”. *Data Science Journal*, 14: 2, 2015, pp.1-10, DOI: http://dx.doi.org/10.5334/dsj-2015-002.

[4] AruminaKumari, Dr. Dinesh Singh, “Reviewing Truth Discovery Approaches And Methods For Big Data Integration”. *International Journal of Science, Engineering and Technology Research (IJSETR)*, Volume 5, Issue 8, August 2016, pp: 2766-2774.

[5] KushalPatel ,“Big Data, its Issues and Challenges” , *JEDR*, Volume 5, Issue 3, ISSN: 2321-9939, 2017, pp:123-27

[6] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han, “Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation,” in *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. ACM, 2014, pp. 1187–1198.

[7] Hsinchun Chen, Roger H. L. Chiang, Veda C. Storey, “Business Intelligence and Analytics: From big data to big impact”, *MIS Quarterly* Vol. 36 No. 4, December 2012, pp. 1165-1188.

[8] D. Wang, M. T. Amin, S. Li, T. Abdelzaher, L. Kaplan, S. Gu, C. Pan, H. Liu, C. C. Aggarwal, R. Ganti, et al. Using humans as sensors: An estimation-theoretic perspective. In Proc. of the *International Conference on Information Processing in Sensor Networks (IPSN’14)*, 2014, pp:35-46.

[9] Li, Yaliang, Jing Gao, ChuishiMeng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. "A Survey on Truth Discovery." arXivpreprint arXiv:1505.02463, 2015.
[10] Fan Zhang, LiYu, Xiangrui Cai, Ying Zhang, Haiwei Zhang, “Truth Finding from Multiple Data Sources by Source Confidence Estimation”, 12th Web Information System and Application Conference, 2015

[11] R. C. Luo, C.-C. Yih, and K. L. Su. Multisensor fusion and integration: approaches, applications, and future research directions. IEEE Sensors Journal, 2(2), 2002, pp: 107-119.

[12] F. Ma, Y. Li, Q. Li, M. Qiu, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji, and J. Han. Faitcrowd: Fine grained truth discovery for crowd sourced data aggregation. In Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’15), 2015, pp: 745-754.