International Journal for Innovative Engineering and Management Research

A Peer Revieved Open Access International Journal

www.ijiemr.org

IJIEMR Transactions, online available on 18th Feb 2022. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 02

DOI: 10.48047/IJIEMR/V12/ISSUE 02/66

Title **Data Reduction Based Truth Discovery Analysis by Resolving the Conflicts in Big Data using Continuous Data**

Volume 12, ISSUE 02, Pages: 423-427

Paper Authors

Dr. P. Bastin Thiyagaraj, Dr. A. Aloysius

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Data Reduction Based Truth Discovery Analysis by Resolving the Conflicts in Big Data using Continuous Data

Dr. P. Bastin Thiyagaraj[1], Dr. A. Aloysius[2]
[1]Department of Information Technology, St. Joseph's College (Autonomous), Affiliated to Bharathidasan University, Trichy – 620002, TamilNadu, India.
bastinstar@gmail.com
[2]Department of Computer Science, St. Joseph's College (Autonomous), Affiliated to Bharathidasan University, Trichy – 620002, TamilNadu, India.
aloysius1972@gmail.com

Abstract

Data are normally generated from various sources and valuable insight for business success. Though, it is complex to process and analyse the data to derive valuable information to strengthen business strategy, performance and efficiency. Big Data provides solutions for companies to make sense of random information. Big data is a term given to describe the volume of data (petabytes (1,024 terabytes) generated from websites, portal, and online applications), unstructured (include emails, voicemails, hand-written text, ECG reading, audio recordings), and complex in processing (from Medical data, Business transactions, Data capture by sensors, Social media/networks, Banking, Marketing, Government data, etc.). The problem is to analyse the data collected from the heterogeneity of sources to identify the truth from the conflicting information. But it is difficult to retrieve the true information when conflicted data consists of the outliers and it affects the performance of truth discovery. The main purpose of this paper is to remove the outlier to enhance the performance and identify the true information from the conflicting data. Since a reliable source can be identified by computing the source weight by removing the outliers.

Keywords: Big data analytics, Loss distance, Source Reliability, Mean Absolute Error.

Introduction

In recent years, due to the development of technology, data is growing exponentially as it is generated and recorded from everywhere. *For example, sensor devices, online social networks, weather online sources, health records, human genome sequencing, phone logs, government records, and professionals such as scientists, journalists, and writers [1].* The amount of useful information *from multiple sources with high volume, velocity, and variety by different digital devices* available on the web is growing at an exponent rate, *giving birth to the term* big data [2]. *But, it is a question whether the information produced by the sources is quality or not for the web searchers. However, many of the* sources produce different types of information for the same objects, which leads to conflicts. It is the moving research to distinguish which source produces a quality of information and which information is fit for an object. The main objective of this paper is to identify a trustworthy source by resolving the conflicts and removing outliers. Here, the heterogeneous information numerical information (Continuous data) and categorical data are involved for the experimental purpose.

Data Analytics serves an important role in analysing conflicted data to produce the true values and quality of sources. To attain the maximum performance with minimum error, the distance-based data reduction approach with the outlier reduction method is used [11]. In Big data evolution, truth discovery has served as an important technique to solve conflicts collected from the huge amount of data provided from heterogeneous sources. The most challenging task is to estimate source reliability and identify the true values supported by the high quality of sources [3]. It is of great interest to identify and remove the outliers out from conflicting data to the communities of machine learning and data mining. Indeed, identifying or eliminating outliers becomes an essential pre-processing stage in data analysis [4].

The main objective of this paper is i) to reduce the outliers and retain reliable values ii) to reduce the maximum deviated values from each source and maximize the weight for each source. This paper is organized as the related works are presented in the following section, and the methodology of the work is delivered in section three. The fourth section illustrates the experimental results. In the fifth section, the comparison of results is given and finally, the future enhancements with summary are presented in section six.

Related Works

In several types of research, a number of outlier reduction and truth discovery methods have been used to identify the quality of source and truth information. Here the outlier and truth discovery is reviewed from the good papers.

In [6], the quality of big data sources has focused. The factor quality includes pattern conflict, identity conflict, and data conflict. In most of the research work, the field of data content is involved to reduce conflicts. In [12], the outlier detection method is proposed to handle high dimensional data to project the high-dimensional into low dimensional space. The distance is the observation that is considered as an outlier deviated far from the truth values [13]. In [7], they observed inconsistency and low accuracy on data, collected from the different sources. Semi super wised truth discovery method is used to identify the confidence score on each observed value. If a confidence score is high the observed values are closest to 1 otherwise to -1. Value 0 indicates the observed values are either true or false [8]. In [9], an algorithm has been designed to find the truth discovery from heterogeneous sources. The objective of the work is to connect the sources (URL) and reliable information. Various web sources provide conflicted objects for the same entity, which leads to object conflict problems [5]. In [10] an optimized framework was proposed to resolve the conflicts on continuous data to compute the source reliability estimation. The main objective of this work is to compute the source weight by identifying the distance between the observed values collected from the various sources and ground truth values. In [14], a method has been proposed to find the most trustworthy sources and the true information by considering the accuracy and coverage of each source at the same time. Weight assignment methods used for continuous data to minimize the distance.

Methodology

Figure 1 shows the overall methodology diagram for CRCO (Conflicts Resolving on Continuous data). CRCO is a data reduction based truth discovery analysis, which consists of two different phases, data reduction, and computing source weight. The weather report data set can be used for the experiments, has the properties, temperature (continuous properties) and weather type (categorical properties), etc. Raw data pre-processed to structure and various distance methods processed the structured data (nine sources each have pair of 425 data sets both continuous and categorical type that are labelled with ground truth values) to analyse and remove the outlier values[11]. The reliable values identified by the Absolute loss distance can be processed by the CRH method to compute the source weights by resolving the conflicts on Continuous data.

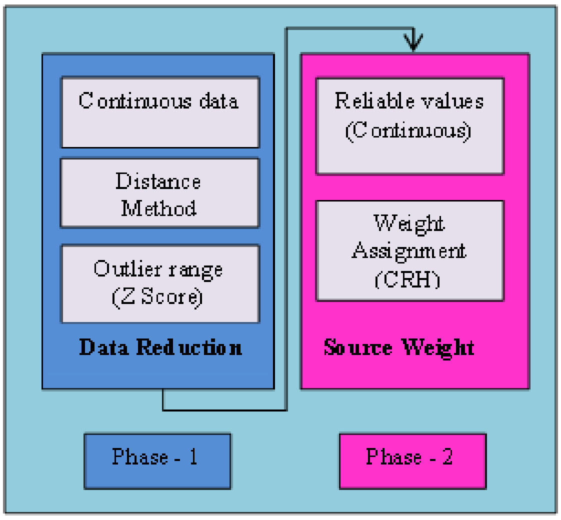

Fig.1. Methodology diagram for truth discovery

Source weights can be computed by measuring the distance of each value of the objects provided by each source. CRH method is an optimization method used to compute the source weights with multi-source heterogeneous data. The basic idea of the CRH framework is “the reliable sources provide trustworthy observations”. Therefore, the observations should be very close to the ground truths from reliable sources. The main objective of the CRH is to maximise the weight of each source by minimizing the distance of each observed value [10]. Therefore, the maximum deviated distance can be removed by removing the outlier values with the help of the Z score numerical outlier reduction method. In general, outlier affects the performance of the methods, so it is to be removed to improve the performance.

If a source is reliable, it satisfies the following constraints i) a maximum number of reliable values after reduction and ii) the highest weight compared to the other sources.

To know the conflicts, there are some terms to be known in this work.

- *Object:* It is a person or thing of interest.
- *Property:* It is a feature or claim used to describe the objects.
- *Source:* It is a place where information about objects’ properties can be collected.

Therefore, to compute the source weights, the solution is given by CRH [10] framework is,

$$w_s^{(co)} = -log\left(\frac{\sum_{i=1}^{N}\sum_{j=1}^{M} d_{ij}\left(v_{ij}^{(t)}, v_{ij}^{(o)}\right)}{\sum_{s=1}^{k}\sum_{i=1}^{N}\sum_{j=1}^{M} d_{ij}\left(v_{ij}^{(t)}, v_{ij}^{(o)}\right)}\right) \qquad (1)$$

Where,
S = Sources ($S^{(1)}, S^{(2)}, .., S^{(K)}$)
W = Source weights ($w_1, w_2, ..., w_K$)
co = Continuous data
i = Object from 1, 2……N
j = Property of the objects from 1,2, ……M
d_{ij} = Distance between the true value and observed value
$v_{ij}^{(t)}$ = Ground-truth value
$v_{ij}^{(o)}$ = Observed values of the objects provided by the various sources

$$d_{ij}\left(v_{ij}^{(t)}, v_{ij}^{(o)}\right) = \frac{\left(v_{ij}^{(t)} - v_{ij}^{(o)}\right)^2}{std\,(v_{ij}^{(1)}, ..., v_{ij}^{(k)})} \qquad (2)$$

Equation (2) is used to maximize the difference between the observed and true information. Z score numerical outlier detection performs on the squared loss distance to remove the maximum deviated distance. Finally, the source weight is computed by taking a negative log on the distance of each source.

Algorithm:
Input: Observations made by K sources: *{$S^{(1)}$, $S^{(2)}$, . . . , $S^{(K)}$}*
Output: *Weight for each source:* W = {w_1, w_2 , ... , w_n}

Step1: Computation of distance of each value from each source
Step2: Generation of range ‘r’ from the values of distance by Z score.
Step3: if distance > ‘r’, then
Step4: Remove outlier values
Step5: end if
Step6: if distance < ‘r’, then

Step7: Apply CRH to compute the weight for each source.
Step8: Apply negative logarithm (weight should be between 0 to 1other wise to be normalised)
Step9: end if
Step10: Return the weight of the sources $w_S^{(co)}$ (S=1,2, 3…k)

Experimental Result

The table1 shows the experimental result for source weight by resolving the conflicts on continuous data. The main objective of this paper is to identify the reliable source from the following constraints: the source has to retain maximum reliable values after reduction and the highest source weight. Since S4 has maximum reliable values and the highest source weight as shown in the table. In [10] which said that, if a source weight is high, the information provided by that source is trustable and reliable. So that the information provided by the source S4 is reliable.

Table.1. Source Weight

Sources	Outlier Reduced Values	Retained Values (Reliable Values)	Weight *(W)*
S1	153	272	0.95711
S2	162	263	0.974909
S3	175	250	0.96465
S4	81	344	0.991708
S5	92	333	0.964957
S6	93	332	0.938385
S7	100	325	0.922185
S8	134	291	0.919457
S9	189	236	0.960116

Results and Discussion

The table.2 shows the performance analysis of deviation rate with existing work. The performance of the CRCO method can be evaluated by Mean of Absolute Deviation (MAD) which works on continuous data.

Table.2. Mean Deviation Rate

S.No	Author	Method	Deviation
1	Q.Li et al.	CRH	4.6947
2	**Bastin et al.**	**CRCO**	**3.9482**

To evaluate the performance, the observed values can be taken from the source that has the highest weight and compared with ground truth values to identify the deviation rate. Here, CRCO method reduces its deviation rate compared to the existing method.

Conclusion and Future Direction

To extract the knowledge from the enormous amount of information generated by heterogeneous sources, it is crucial to identify the true information from the multiple conflicting data sources. In this model, a reliable source is defined as the smallest weighted from multi-source input in which weights represent source reliability degrees. Here, the continuous type of data only can be used to estimate the source weight and identify the reliable source. But, it is not enough to compute the source weight by using continuous data only. Normally, the object consists of more than one property. For example, weather data consists of continuous data and categorical data. Therefore, future work is to compute the source weight by categorical information. In truth finding problems, reliable source and reliable information can be identified.

References

[1] Kaisler, S., Armour, F., Espinosa, J.A., Money, W., “Big data: issues and challenges moving forward”, 46th Hawaii International Conference on System Sciences (HICSS), IEEE (2013), pp. 995–1004.

[2] Mudasir Ahmad Wani and Suraiya Jabin, “Big Data: Issues, Challenges, and Techniques in Business Intelligence”, Springer Nature Singapore Pte Ltd. 2018, pp: 613-28.

[3] Xueling Lin, Lei CHEN, “Domain-Aware Multi-Truth Discovery from Conflicting Sources”, PVLDB, 11(5), DOI: https://doi.org/10.1145/3177732.3177739, 2018, Pp: 635- 647.

[4] J. Ha, S. Seok, and J.S. Lee, “Robust outlier detection using the instability factor”, Knowledge Based on System, volume 63, issue 6, 2014, pp: 15–23.

[5] Wenqiang Liu, Jun Liu, Bifan Wei, Haimeng Duan, Wei Hu, “A new truth discovery method for resolving object conflicts over Linked Data with scale-free property”, Springer Nature 2018, pp:01-31.

[6] Wei Jiang, Xiuli Ning, Yingcheng Xu., “Review on Big Data Fusion Methods of Quality Inspection for Consumer Goods”, 5th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud), 2018, pp:95-102.

[7] Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, Divesh Srivastava, “Truth Finding on the Deep Web: Is the Problem Solved?”, Proceedings of the VLDB Endowment, Volume 6, issue 2, 2015.

[8] D. Yu, H. Huang, T. Cassidy, H. Ji, C. Wang, S. Zhi, J. Han, C. Voss, and M. Magdon Ismail, “The wisdom of minority: Unsupervised slot filling validation based on multi-dimensional truth-finding”, In Proc. Of the International Conference on Computational Linguistics (COLING’14), 2014.

[9] Palaiyah Solainayagi a, Ramalingam Ponnusamy, “Trustworthy media news content retrieval from the web using truth content discovery algorithm”, 1389-0417, Elsevier, 2019, pp: 26-35.

[10] Li, Q., Li, Y., Gao, J., Zhao, B., Fan, W., & Han, J. “Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation”, ACM, 2014, pp: 1187–1198.

[11] P. Bastin Thiyagaraj, Dr. A. Aloysius, “Empirical Study on Distance Based Data Reduction for Truth Discovery using Conflicted Data”, Mukt Shabd Journal, 2347-3150, Volume 9, Issue 7, 2020, pp:2204-08.

[12] Huawen Liu Xuelong Li, Jiuyong Li, and Shichao Zhang, “Efficient Outlier Detection for High-Dimensional Data”, IEEE Transactions On Systems, Man, And Cybernetics Systems, ISSN: 2168-2216, 2017, pp:01-11.

[13] H. Huang, K. Mehrotra, and C. K. Mohan, “Rank-based outlier detection”, J. Stat. Comput. Simulat., volume 83, issue 3, 2013, pp. 518–531.

[14] 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, pp:01-04.

[15] A.Angelpreethi, Dr.S.Britto Ramesh Kumar,” Dictionary based approach to improve the accuracy of opinion mining on big data”, International Journal of scientific research in computer science and management studies, volume 7, issue 5 ,sep 2018.

Authors Profile

P. BASTIN THIYAGARAJ is working as an Assistant Professor in the Department of Information Technology, St.Joseph’s College(Autonomous), Tiruchirappalli, Tamil Nadu, India. I have 12 years of experience in teaching and 7 years in research.

Dr A. ALOYSIUS is working as an Assistant Professor in the Department of Computer Science, St. Joseph’s College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He has 21 years of experience in teaching and research. He has published many research articles in the National / International conferences and journals. He has acted as a chairperson for many national and international conferences. Currently, eight candidates are pursuing Doctor of Philosophy Programmed under his guidance.