An Approach for Truth Discovery by Resolving the Conflicts on Categorical Data

P Bastin Thiyagaraj¹, A Aloysius²

1, ²Assistant Professor
1Department of Information Technology, St. Joseph’s College, Trichy, India.
2Department of Computer Science, St. Joseph’s College, Trichy, India.
¹bastinstar@gmail.com.

Abstract. Data is playing an important role in the world for competitive advantages among nations, organizations and decision making for business development. It is indeed that every day quintillion bytes of different types of data from variety of sources are created and complex in processing (from Medical data, Business transactions, Data captured by sensors, Social media/networks, Banking, Marketing, Government data, etc.). The different sources produce large amount of structured, semi structured and unstructured data with varieties of data. Normally, the unstructured content collected by an organization is in the form of textual format, corporate documents to web pages and social media content. When number of sources produces various descriptions for a particular object, it leads to data conflict. It can be processed, when numerical data and measurement data are produced for the objects using big data tools and statistical tools for having high achievements. But, it is the challenging one to extract and analyse conflicted text from corpus data, collected from various sources. The objective of our work is to produce the true information even when distance between observed values and mean values are closest to each other.

Keywords: Big Data analytics, Levenshtein Distance, Accuracy, Naïve Bayse classifier.

1. Introduction

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 capture by sensors, Social media/networks, Banking, Marketing, Government data, etc.). 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]. So that a buzzword is used for describing huge amount of data both unstructured and structured that is big data. The big data has produced a variety of datasets from various sources in different domains. These kind of datasets consist of a different Representation and density. Data analytics [2] is the science if examining the raw data to draw conclusion about the information.

The Data Analytics [3] is playing an important role to analyze the conflicted data using statistical approach to find highest achievable performance by maximizing desired factors and minimizing undesired ones, under the given constraints. In comparison, maximization means that to attain the highest result or outcome without regard to expense. It is very complex task to analyze huge amount of different varieties of conflicted data from various conflicted sources. Here the heterogeneous data are involved such as both numerical data (measurement data) and categorical data (classified data). In the complicated world, it is a tedious process to estimate source reliability to find out the correct information from conflicted text data, especially when sources provide low quality information, such as faulty sensors provide wrong data, and spam users who spread false information on the Internet[18]. Text analytics techniques have been actively involved in several emerging areas, for example
information extraction, topic models, question answering (Q/A), and opinion mining [2]. The rest of this paper is organized as follows: in the next section related works of my work is presented. In section 3 the methodology is developed. Accuracy and Truth values are identified based on the methodology in section 4 and 5. In section 6 the result is discussed and in section 7 conclusion is presented.

2. Related Works

Text analytics [4] plays an important role in information retrieval and computational linguistics. Document representation and query processing are the foundations for information retrieval, to develop the vector-space model, Boolean retrieval model, and probabilistic retrieval model, and these became the basis for the modern digital libraries, search engines, and enterprise search systems [5]. Statistical natural language processing (NLP) techniques for lexical acquisition, word sense disambiguation, part-of-speech-tagging (POST), and probabilistic context-free grammars have also become important for representing text in computational linguistics, [6]. In addition to document and query representations, user models and relevance feedback are also important in enhancing search performance.

In [7] this paper they focused on the Levenshtein distance Algorithm which is used between two strings of character to convert one string to another string, which are equal to the minimum number of insertions, deletions and substitutions. The Levenshtein distance used for many areas, e.g. text analysis (detection of plagiarism) [8, 9], spell-checking in text processors [10], web mining (search engine robots) [11, 12], bioinformatics (Levenshtein-Damerau distance for DNA structure analysis [13, 14]), etc.

The method [15] “Edit distance” is playing an important role for quantifying how two dissimilar words to transform to one string into the other string by counting the minimum number of operation. Normally, Edit distances are used in natural language processing, where automatic spelling correction can be determined for a misspelled word. The words are selected from a dictionary that has a low distance to the word in question.

In [16] a string metric (also known as a string similarity metric or string distance function) is a metric that measures distance (“inverse similarity”) between two text strings for approximate string matching. These methods are not suitable for categorical data to select which data produced is correct from the conflicted data given by various sources when scalability is increased.

In [17] they proposed a method to find the most trustable source and identify the true information. They found to minimize the distance between the true information and the overall observed descriptions through considering the accuracy and the coverage of all the data sources at the same time. They used to find the distance for categorical data which is 0-1 loss function. In this paper each observation will be either correct or wrong from the true fact for categorical data. It is very tedious process to identify truth discovery when the objects have continuous data [18].

3. Methodology

It is very tedious process to find truth information from conflicted categorical data (with continuous data) produced from the various sources for the objects. The previous system used single type of data to resolve conflict produced from the various sources. For text data there are number of tools and algorithms are available to replace the letters of two strings. In [18] they produced better result in order to find the truth information from the conflicted categorical data from the various sources. But, this system is not suitable for scalability issues. Our system uses the heterogeneous data both continuous data and categorical data given in [18] in order to overcome the scalable issues. In order to overcome this issue, we use Naive Bayes approach to identify the frequency rate of conflicted categorical data from the various sources. This approach is highly scalable.
3.1 Source Observation
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. The data of citizenship values are taken from [18] given below the table1. Here the property “city” is considered into an account with the property height.

| Object | X1 | X2 | X3 |
|--------|----|----|----|
|        | City | Height | City | Height | City | Height |
| Bob    | NYC | 1.72 | NYC | 1.7 | NYC | 1.9 |
| Mary   | LA  | 1.62 | LA   | 1.61 | LA  | 1.85 |
| Kate   | NYC | 1.74 | LA   | 1.72 | LA  | 1.65 |
| Mike   | NYC | 1.72 | LA   | 1.7 | DC  | 1.85 |
| Joe    | DC  | 1.72 | NYC  | 1.71 | NYC | 1.85 |

The table2 explains that the truth value for Height (Continuous data) is taken as the number of occurrences from the conflicted observed values from various sources.

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

The methodology diagram and algorithm are shown below to expose our work to identify the true information (categorical data) from the conflicted information generated by various sources. The Naïve Bayse approach is used to identify the frequency rate of the objects.

**Methodology Diagram**

![Methodology Diagram](image)

*Figure 1. Methodology Diagram* The figure shows that the identification of true values among conflicted categorical data from conflicted sources.
Algorithm:
Input: Data from K sources (X₁, X₂, X₃... Xₙ)
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
        Calculate the Accuracy of the n th object on the m th property 𝑣ₙₘ using
        
        \[ \text{Accuracy} = 100 - \left( \frac{\left( \frac{(v_{nm}^{(c)} - v_{nm}^{(k)})^2}{\text{std}(v_{nm}^{(c)} - v_{nm}^{(k)})^2} \right)}{\text{Truth value}} \right) \times 100 \]
        
        Calculate the frequency rate of each properties of object using Naive Bayes Statistical classifier
        
        \[
        P\left( \frac{\text{NYC}}{X1(Mike)} \right) = \left( \frac{P(X1\text{Mike})}{P(X1)} \right) 
        \]
    end for
end for

Return truths from maximum frequency rate.

4. Accuracy Estimation
It is the challenging one to identify the truth information from the conflicted categorical data with continuous data. Existing system provided better result to find truth information of categorical data, and the scalability is not satisfied in that system. The voting method is not suitable for selecting the truth value from the conflicted data. First the source accuracy is to be found in order to find the truth value of categorical data. The distance can be found from the formula is

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

The average percentage of accuracy of every property of an object is given below,

| Object | X1     | X2     | X3     |
|--------|--------|--------|--------|
|        | City   | Height (%) | City | Height (%) | City | Height (%) |
| Bob    | NYC    | 98.72  | NY1   | 97.49  | NY1  | 91.33  |
| Mary   | LA     | 97.86  | LA    | 97.21  | LA   | 88.84  |
| Kate   | NY1    | 98.01  | NY1   | 99.5   | LA   | 96.89  |
| Mike   | NY1    | 98.88  | LA    | 97.49  | DC   | 94.35  |
| Joe    | DC     | 98.84  | NY1   | 98.18  | NY1  | 94.11  |

In order to find the total accuracy, the equation (1) is rearranged as,

\[ \text{Total Accuracy} = \sum_{n=1}^{N} \sum_{m=1}^{M} 100 - \left( \frac{\left( \frac{(v_{nm}^{(c)} - v_{nm}^{(k)})^2}{\text{std}(v_{nm}^{(c)} - v_{nm}^{(k)})^2} \right)}{\text{Truth value}} \right) \times 100 \]
The calculated values of continuous data are given by,

Table 4: Total Accuracy percentage of Source and Object

| Object | X1 City | X2 City | X3 City | Total |
|--------|---------|---------|---------|-------|
|        | Height  | Height  | Height  |       |
| Bob    | NYC     | NYC     | NYC     | 287.54|
| Mary   | LA      | LA      | LA      | 283.91|
| Kate   | NYC     | NYC     | LA      | 294.4 |
| Mike   | NYC     | LA      | DC      | 290.72|
| Joe    | DC      | NYC     | NYC     | 291.13|
| Total  | 492.31  | 489.87  | 465.52  | 1447.7|

From the above table 4, it is clear that the source X1 produces better accuracy of information that means source weight is high compared with X2 and X3. So that, the source X1 produces truth information from the conflicted data. In [18] they used 0-1 loss function in order to find the true categorical data of objects from the conflicted data. But, it is not suitable for scalable issues.

Assumption 1: Normally, the highest number of occurrences is considered as mean value for categorical data. For the object Mike, three sources produced different city name. So, here randomly chosen value considered as mean value. Suppose the observed values (continuous data) of all the sources are closest to the truth value, the weights of the three sources are closest to each other. At this situation the 0-1 loss function [18] fails to predict the true value (from the source weight is high).

Assumption 2: Consider the object “Joe”, which has the city name as “DC”, “NYC”, and “NYC” from the three sources X1, X2, and X3. Here, “NYC” considered as truth value (Highest number of occurrence). But, the value “DC” from source X1(highest source weight) taken as final result by using 0-1 loss function. At this situation, suppose the source increases and has “NYC” as truth value, the 0-1 loss function will fail to predict the value from the source X1 (highest source weight). So that, the individual probability is taken by using Naive Bayes classifier approach for conflicted categorical data in order to solve the above problems.

5. Truth Estimation

Naive Bayes classifier is a “probabilistic classifiers” approach based on applying Bayes’ with strong (naive) independence assumptions between the features, and this method is mainly used for text categorization with word frequencies. The equation for the probabilistic model of Naive Bayes Classifier is given for the object Mike to get which property (city name) from the sources,

\[ P\left(\frac{NYC}{X1(Mike)}\right) = \left(\frac{P(X1_{Mike} \text{NYC}) \times P(NYC)}{P(X1_{Mike})}\right) \]  

(3)

Example 1: Let us consider an object Mike which has the city name from the sources X1, X2, X3 are NYC, LA, DC. Here which one is true information for the object Mike? From the table 3 the values are applied to the equation (3) to identify truth information.

The final result, produced the value “NYC” from the source X1, which has the highest frequency (but the mean value is DC) as shown in the table 5.

Example 2: Let us consider an object Joe which has the city name form the sources X1, X2, X3 are DC, NYC, NYC (The mean value is “NYC”). Here which one is true information for the object Joe?
The final result produced the value “DC” from the source X1, which has the highest frequency as shown in the table 6.

### Table:5 Frequency rate for Mike

| Object | Source | City | Frequency | Final Result |
|--------|--------|------|-----------|--------------|
| Mike   | X1     | NYC  | 0.340121  | 0.340121 (NYC) |
|        | X2     | LA   | 0.336715  |              |
|        | X3     | DC   | 0.324539  |              |

### Table:6 Frequency rate for Joe

| Object | Source | City | Frequency | Final Result |
|--------|--------|------|-----------|--------------|
| Joe    | X1     | DC   | 0.339505  | 0.339505 (DC) |
|        | X2     | NYC  | 0.337238  |              |
|        | X3     | NYC  | 0.323498  |              |

6. Results and Discussion

The accuracy is identified using continuous data and true value of categorical data can be identified using Naive bayes probabilistic classifier. The 0-1 loss approach is not suitable for scalability. Whereas, our method is highly scalable to identify the truth data from the conflicted data of the sources. The final result is given as in the table 7.

### Table 7: Result

| Object | Voting/Average | Our method |
|--------|----------------|------------|
|        | City           | City       |
| Bob    | NYC            | NYC        |
| Mary   | LA             | LA         |
| Kate   | NYC            | NYC        |
| Mike   | DC             | NYC        |
| Joe    | NYC            | DC         |

So that, the source X1 produces better information than the sources X2, and X3 from our conclusion.

7. Conclusion and Future Direction

Big data is a term given for the volume of data. Big data analytics is the process of analysis the data to predict the conclusion for future. But, here the heterogeneous types of data are taken to identify the true information from the conflicts. It is the challenging process to identify the true data (categorical data) for the objects from the conflicted data even when distance between observed values and mean values are closest to each other in all the sources. First, the source accuracy is identified to find the truth categorical data of the objects. In future this system will be implemented by using Big data hadoop environment system with real data set.

References

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

[2] Zhihan Lv, Houbing Song, Pablo Basanta-Val, Anthony Steed , Minho Jo, “Next-Generation Big Data Analytics: State of the Art, Challenges, and Future Research Topics”, *IEEE*, DOI 10.1109/TII.2017.2650204, pp: 01-09.

[3] https://www.simplilearn.com/data-science-vs-big-data-vs-data-analytics-article.
[4] 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.

[5] Salton, G.. Automatic Text Processing, Reading, MA: Addison Wesley 1989.

[6] Manning, C. D., and Schütze, H., “Design Science in the Information Systems Discipline,” Foundations of Statistical Natural Language Processing, Cambridge, MA: The MIT Press. March, S. T., and Storey, V. C. 2008. MIS Quarterly (32:4), pp. 725-730.

[7] Artur Niewiarowski, Marek Stanuszek, “PARALLELIZATION OF THE LEVENSHTEIN DISTANCE ALGORITHM”, Technical Transactions Fundamental Sciences, 2014.

[8] Niewiarowski A., “Term frequency optimization for the vector space model”, Czasopismo Techniczne, 9-M/2012, 155-165.

[9] Ramos J., “Using TF-IDF to determine word relevance in document queries”, Proceedings of the First Instructional Conference on Machine Learning, 2003.

[10] Wypych M., “Stochastic Spelling Correction of Texts in Polish”, Institute of Linguistics, Adam Mickiewicz University, Poznań, Poland; Speech and Language Technology. Volume 6, Poznań 2002.

[11] Runkler T.A., Bezdek J.C., “Web mining with relational clustering”, International Journal of Approximate Reasoning, Vol. 32, Issues 2–3, February 2003, 217-236.

[12] Niewiarowski A., Działanie parsera, ”Part-of-Speech Tagging” wjęciu mechanizmu Web Content Mining, Wydawnictwo VI Ogólnopolskiej Konferencji Naukowej Nauka i Przemysł, Politechnika Krakowska im. Tadeusza Kościuszki, Kraków 2011, 93-100

[13] Damerau F.J., “A technique for computer detection and correction of spelling errors”, Communications of the ACM, 7 (3), 1964, 171-176.

[14] Niewiarowski A., Stanuszek M., “Parallelize edit distance algorithm”, Proceedings. Seventh ACC Cyfronet AGH Users’ Conference, Academic Computer Centre Cyfronet AGH, Zakopane 2014, 31-32.

[15] https://en.wikipedia.org/wiki/Edit_distance

[16] https://en.wikipedia.org/wiki/String_metric

[17] Fan Zhang, Li Yu, Xiangrui Cai, Ying Zhang, Haiwei Zhang,”Truth Finding from Multiple Data Sources by Source Confidence Estimation”, 978-1-4673-9372-0/15 $31.00 © 2015 IEEE DOI 10.1109/WISA.2015.45, pp:153-56.

[18] 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.