
# Data Fusion through Truth Discovery

Nassim Habbash (808292) *University of Milano-Bicocca*
*Final Project for the Data Management course*
n.habbash@campus.unimib.it

***Abstract*—Data Fusion is a set of techniques aiming to resolve conflicts from a collection of multiple conflicting sources. This work sets to use one of such techniques - truth discovery - through the implementation of the TruthFinder model [1], and analysis of its performance on a dataset of scraped books from the web.**

***Index Terms*—Data Management, Data Fusion, Truth Discovery**

## I. Introduction

The four big Vs of Big Data represent important properties of data in modern times. Volume, Variety, Velocity and Veracity: of these four, one of the most tricky to handle is Veracity. With so much data available, ensuring data quality, source trustworthiness and processing accuracy becomes crucial for its effective employment, especially considering how easy it is nowadays to publish, copy and spread false information on the web. As such, multiple pipelines, algorithms and complex heuristics have been researched in the past decades to tackle the complex problem of reconciling possibly untrustworthy data. One of these is TruthFinder [1], a model that tries to solve the Veracity problem by using the relationship between sources and their information in a probabilistic framework.

## II. Dataset

The dataset used for this work is the Books dataset [2]. The dataset was obtained through scraping from AbeBooks.com, a book aggregator website, and is comprised of a set of Computer Science books. Every record is composed of 4 attributes:

- **Source**: The source bookstore
- **ISBN**: The book's numeric identifier
- **Title**: The book's title
- **Authors**: The authors of the book

The dataset is composed of 33971 records, 1263 unique books and from 894 sources.

The dataset is coupled with a Gold Standard, which has been obtained by sampling 100 unique random books and manually checking and inputting the full list of their authors from the covers of the books. The Gold Standard is composed of 2 attributes:

- **ISBN**: The book's numeric identifier
- **Authors**: The authors of the book

### A. Preprocessing

Obtained through web scraping, the dataset presents multiple incongruences and heterogeneities that make it hard to process as it is. The attribute presenting most heterogeneities is the **Authors** field: different sources spell the authors names differently, employ different listing characters or present HTML unescaped tags in them. For example, for the book with the ISBN code 0131869000, the dataset presents, between many others:

- Deitel, H. M./ Deitel, P. J.
- Deitel & Associates, (Harvey & Paul)

Preprocessing has been performed on the **Authors** and **Title** fields, and the process has been defined as follows:

1) Escaping leftover HTML characters (e.g. & $\rightarrow$ &)
2) Escape characters removal (e.g. \r, \n)
3) Lowercasing
4) Parenthesis removal
5) Separators replaced by spaces (e.g. pipe character)
6) Special characters removal
7) Digits removal
8) Trailing and double whitespaces removal
9) Missing values uniformation (e.g. NA, Not Available and empty strings replaced by NaN)

The operations have been performed in such order to maximize the number of successfully cleaned fields.

Many sources, for example, present a list in the format ”Stuart J. Russell—Peter Norvig”, which through a straight special characters removal fuses the last name and the first name of the authors.

The cleaning procedure has been also applied to the Gold Standard.

### B. Data Exploration

The dataset presented after the processing 7557 duplicated records, that have been dropped to make easier further processing. The Data Quality dimensions analyzed are Attribute, Table Completeness and Source Coverage, as effective Accuracy will be analyzed further down the line through the TruthFinder model.

TABLE I: Data Completeness

| | Source | ISBN | Title | Authors |
|---|---|---|---|---|
| Null Count | 0 | 0 | 0 | 649 |
| Attribute Completeness | 1 | 1 | 1 | 0.97 |
| Table Completeness | 0.99 | | | |

The dataset presents few null values in the Author attribute, amounting to just the 0.03% of the total record, with an overall table completeness of 99%, as shown in Table I.

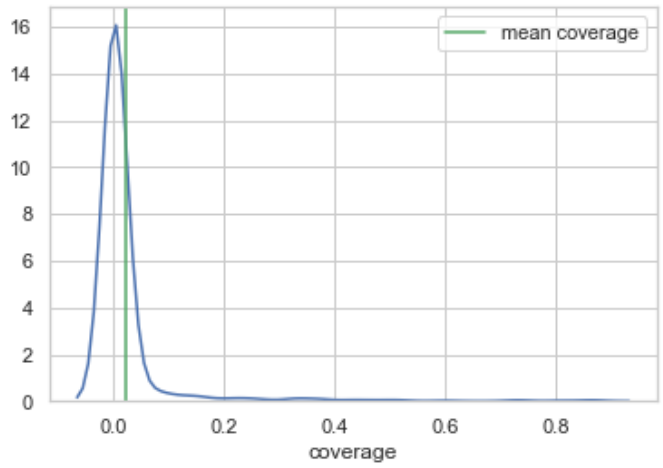


Fig. 1: Source coverage distribution

Figures 1 shows the source coverage distribution. With a mean of 0.02, most sources cover just a tiny fraction of the unique books in the dataset. This is

Figure 2 shows the top and bottom 10 sources for coverage. Coverage has been computed as the percentage of unique ISBN in the dataset covered by each source. We can see how only few sources cover most of the dataset, namely Revaluation Books and A1Books, offering each both 86% of the book catalogue.

TABLE II: Data statistics before preprocessing

| | source | isbn | title | authors |
|---|---|---|---|---|
| count | 33971 | 33971 | 33968 | 33971 |
| unique | 894 | 1265 | 11095 | 9627 |
| top | A1Books | 0321263588 | (...) | \r |
| freq | 2403 | 159 | 90 | 713 |

TABLE III: Data statistics after preprocessing

| | source | isbn | title | authors |
|---|---|---|---|---|
| count | 33971 | 33971 | 33971 | 33172 |
| unique | 894 | 1265 | 7195 | 6901 |
| top | a1books | 0321263588 | (...) | meyers scott |
| freq | 2403 | 159 | 108 | 136 |

Tables II and III show the data statistics before and after preprocessing. We can see that A1Books has the most entries in the dataset, with 2403 records as a source, but probably not as many unique books as Revaluation Books, as seen above. The most common ISBN is 0321263588, with 159 records, and the most common title is ”Computer Networking and the Internet”, with 108 records, and not surprisingly corresponds with the previously mentioned ISBN, although less times probably due to conflicting titling from different sources. Before preprocessing, the most common title was ”Modern Database Management”, with 90 records. At last, the most frequent author is ”Meyers Scott”, with 136 records, while before preprocessing, at least 713 Authors fields contained nothing more than the escape symbol \r.

The dataset presents 894 unique sources, 1265 unique ISBNs, 7195 unique titles and 6901 unique Authors. The next section will delve into how, between the many conflicting informations the attributes have for an ISBN (here representing a unique object), it’s possible to extract the most probable true value.

## III. Truth Discovery

TruthFinder [1] is a computational model based on the following four heuristics:

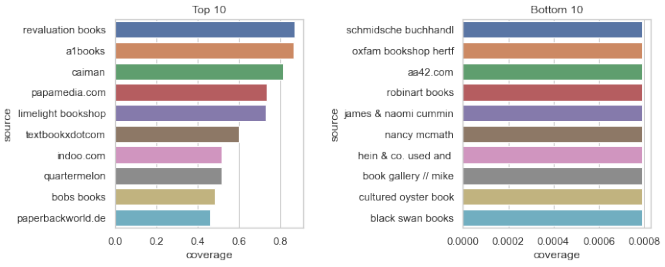


Fig. 2: Top 10 and bottom 10 sources by coverage

1) Usually there is only one true fact for a property of an object
2) This true fact appears to be the same or similar on different sources
3) The false facts on different sources are less likely to be the same or similar
4) In a certain domain, a source that provides mostly true facts for many objects will likely provide true facts for other objects

With object, we refer in this case to books, and with fact to their attributes, or more generally, something claimed to be a fact by some source, which can be either true or false - in this case, Authors or Title are facts. The model takes into account the trustworthiness of sources and confidence of facts; the **confidence** of a fact $f$ is the probability of $f$ being correct. The **trustworthiness** of a source $w$ is the expected confidence of the facts provided by $w$. Figure 3 shows how the data is structured.

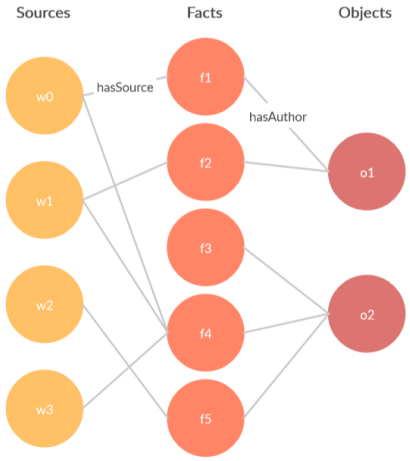


Fig. 3: TruthFinder conceptual schema

Confidence and trustworthiness are dependant on each other, so the model takes an iterative approach to make an estimation. The model consists of two major steps:

1) Fact confidence score computation
2) Source trustworthiness score computation

The model iterates both steps until it comes to stability and there are no more major changes in both estimations.

### A. Fact confidence estimation

Fact confidence is estimated through three progressive steps.

Given a fact $f \in F(w)$, $F(w)$ being the set of facts of source $w$, we call fact confidence $\sigma(f)$ the following:

$$\sigma(f) = \sum_{w \in W(f)} \tau(w) \tag{1}$$

Equation 1 computes the initial fact confidence by summing all source confidence score $\tau(w)$.

$$\sigma^*(f) = (1-\rho)\cdot\sigma(f) + \rho\cdot \sum_{o(f')=o(f)} \sigma(f')\cdot imp(f' \rightarrow f) \tag{2}$$

Equation 2 adjusts the confidence score of a fact using the similarity it has to other facts (represented by the implication). To the initial confidence $\sigma(f)$ an added factor (modulated by $\rho$) controls the influence other similar facts have on the fact in consideration: similar names do probably implicate each other (George Luger implicates George F Luger).

$$s(f) = \frac{1}{1 + e^{-\gamma\sigma^*(f)}} \tag{3}$$

Equation 3 adjusts again the confidence, squishing the function through a sigmoid and adding a dampening factor $\sigma$. The dampening factor is to account for source dependance, because it's possible that sources do copy from each other. The sigmoid function allows to transpose negative probabilities in the range of $[0, 1]$. Note that $s(f)$ is the confidence of $f$, while $\sigma(f)$ is the confidence score of $f$. The quantities are tied to each other and to the source trustworthiness and trustworthiness scores $t(w)$ and $\tau(w)$ by other equations explained in depth in [1]. Briefly, we have that the scores are the negative log likelihood of the values, so $\sigma(f) = -ln(1 - s(f))$ and $\tau(w) = -ln(1 - t(w))$

### B. Source trustworthiness estimation

Source trustworthiness is computed as the average confidence of all the facts from a specific source, as follows in 4:

$$t(w) = \frac{\sum_{f \in F(W)} s(f)}{|F(w)|} \tag{4}$$

### C. Process convergence

The process iterates, updating facts confidence and source trust scores at each loop. Given the vector $t_i$ of trust scores at time $i$, and the vector $t_j$ of trust scores at time $j$, $j = i + 1$, if the change measured as error between the two vector is low enough, the process is considered terminated. More formally:

$$error = 1 - \frac{t_i \cdot t_j}{||t_i|| \cdot ||t_j||} \tag{5}$$

## IV. Implementation

### A. Tools

To implement the model, the following tools and frameworks have been used:

- **Python** 3.8
- **Pandas** for data management
- **Strsimpy** and **FuzzyWuzzy** for string similarity
- **Numpy** for scientific computation

To aid the analysis and workflow, the data has been analyzed and the model tested on Jupyter Notebook.

### B. Code

The model has been implemented as a Python class. Listing 1 reports the main parameters of the model and their default initialization. The standard parameters initializations are:

- **Dampening factor** ($\sigma$): 0.3
- **Relatedness factor** ($\gamma$): 0.5
- **Base similarity**: 0.5
- **Max iterations**: 10
- **Tolerance**: 0.001
- **Initial source trustworthiness**: 0.9
- **Implication function**: cosine similarity

These values have been taken by the original paper.

Listing 1: TruthFinder initialization

```
class TruthFinder(object):
    '''
    TruthFinder model implementation, finds true
        values about objects from conflicting
        sources (Veracity problem).

    Attributes:
        df (DataFrame): DataFrame containing the
            data,
        fact (string): Fact/Attribute column name
             in the DataFrame
        obj (string): Object/Identifier column
            name in the DataFrame
        implication (function): Similarity
            function between strings
        initial_trust (float): Initial sources
            trustworthiness
        dampening_factor (float): Dampening
            factor (gamma) to account for source
             dependance
        relatedness_factor (float): Relatedness
            factor (rho) to account for the
            influence of related facts
        base_sim (float): Threshold for positive
            implication
    '''
    def __init__(self, df, fact, obj, implication
         = None, initial_trust = 0.9,
        dampening_factor = 0.3,
        relatedness_factor = 0.5, base_sim =
        0.5):
        self.df = df
        self.fact = fact
        self.object = obj

        if implication==None:
            self.implication = cosine_sim
        else:
            self.implication = implication

        self.initial_trust = initial_trust
        self.dampening_factor = dampening_factor
        self.relatedness_factor =
            relatedness_factor
        self.base_sim = base_sim
```

Listing 2 is the main iteration of the model with another set of parameters. The model iterates the functions `self.compute_fact_confidence()` and `self.compute_source_trust()` a fixed number of times or until convergence, and returns the set of unique objects with their associated facts with the most confidence.

Function `self.compute_fact_confidence()` implements and computes Equations 1, 2 and 3 sequentially, while function `self.compute_source_trust()` implements and computes Equation 4.

Listing 2: TruthFinder compute function

```
    def compute(self, max_it = 10, tolerance = 0.001)
        :
        '''
        Iterative computation of fact confidences and
             source trustworthiness until stability
        '''

        self.df['trust'] = self.initial_trust
        self.df['confidence'] = 0.0

        for i in range(max_it):
            t1 = self.df.drop_duplicates("source")["
                trust"]

            self.compute_fact_confidence()
            self.compute_source_trust()

            t2 = self.df.drop_duplicates("source")["
                trust"]

            # Convergence of the process is measured
                by the change in trustworthiness of
                sources
            error = (t1 @ t2.T) / (np.linalg.norm(t1)
                *np.linalg.norm(t2))
            error = 1 - error

            if error > tolerance:
                break

        return self.extract_truth()
```

The work implemented two kinds of similarity functions, `cosine_sim(f1, f2` and `character_token_sim(f1, f2)`. The first function implements cosine similarity between the two strings converted into vectors, while the second function computes similarity through a ngram intersection between the two, following particular heuristics. Similarity is offset by the parameter base similarity, as such strings with low similarity ($<$ base similarity) end

up with a negative relatedness factor in the confidence computation.

## V. Data Fusion and results

A first experiment was run with the standard parameters taken from the original paper. Data Fusion accuracy has been measured as the average accuracy of facts selected by TruthFinder for books present in the Golden Standard. For each object (book) in the GS, cosine similarity has been applied to check how much the fact with the highest confidence matches with the fact as reported on the GS.

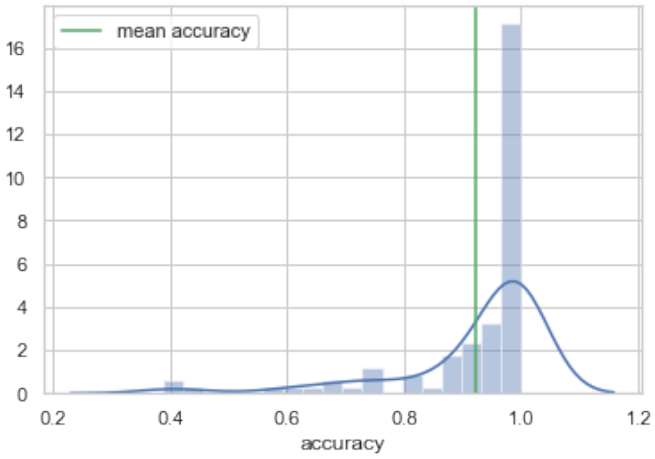


Fig. 4: Fact accuracy distribution, mean=0.92

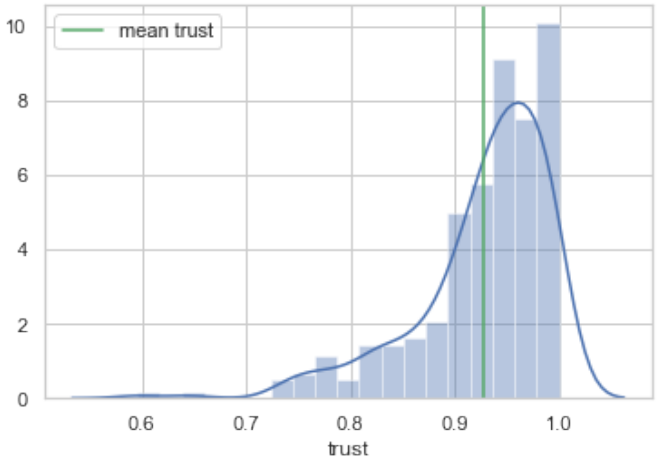


Fig. 5: Source trustworthiness distribution, mean=0.93

The model achieved on the dataset an accuracy of 92%, with an average source trustworthiness of 94%. Figure 4 reports the fact accuracy distribution: the curve is greatly skewed towards the higher end of the distribution, showing how the model managed to retrieve most author lists correctly. Figure 5 shows the source trustworthiness distribution. With an initial trustworthiness of 0.9, after process stability has been reached, most sources have converged towards the higher half of the distribution, meaning that, while presenting a long left tail, many sources have been deemed trustworthy in the dataset.

TABLE IV: Top 5 sources by trustworthiness

| source | trust |
|---|---|
| spine and crown | 1.0 |
| er books | 1.0 |
| a novel idea bookstore | 1.0 |
| strand book store, abaa | 1.0 |
| gail p. kennon, book-comber | 1.0 |

TABLE V: Bottom 5 sources by trustworthiness

| source | trust |
|---|---|
| reliable enterprises, inc. | 0.52 |
| hyannisport books | 0.64 |
| opoe-abe books | 0.67 |
| textbooksnow | 0.68 |
| technischer overseas pvt. ltd. | 0.71 |

Tables IV and V show the top and bottom 5 sources for trust after process stability. While there are plenty of sources thast converged on 100% trustworthiness, in the long tail of the distribution there are few sources that have gone below 70%.

### A. Parameter search

The initial tolerance of 0.001 (as in the original paper) in this work's implementation brought to convergence after just one iteration, as the error easily trespassed the tolerance threshold. As such, for the following experiment, the tolerance has been raised to 0.01, and the accuracy measured for 1, 10, 20 and 30 iterations of the algorithm.

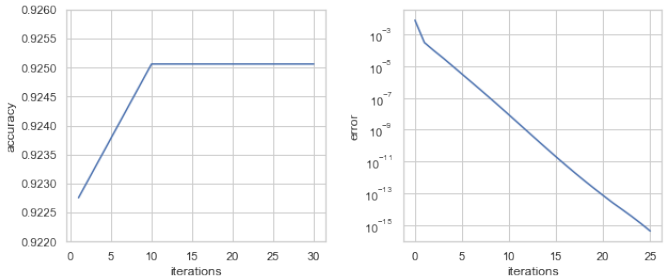


Fig. 6: Error and Data Fusion accuracy at varying iterations

Figure 6 shows how the error and accuracy change in function of the maximum number of iterations. The error exponentially converges to smaller values close to zero (in the graph the y axis is on a logarithmic scale), while the accuracy converges to 92.5% after 10 iterations, making further computation virtually useless, as even with the tolerance non trespassed additional accuracy hasn't been gained.

A grid search has been applied for the dampening and relatedness parameters ($\sigma$ and $\rho$). The other parameters have been kept as in the former experiment, except the initial trustworthiness, set at 80%.

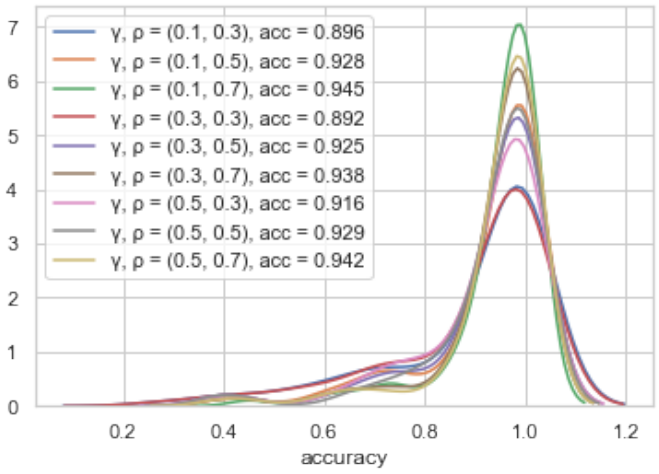


Fig. 7: Data Fusion accuracy distribution for different parameters

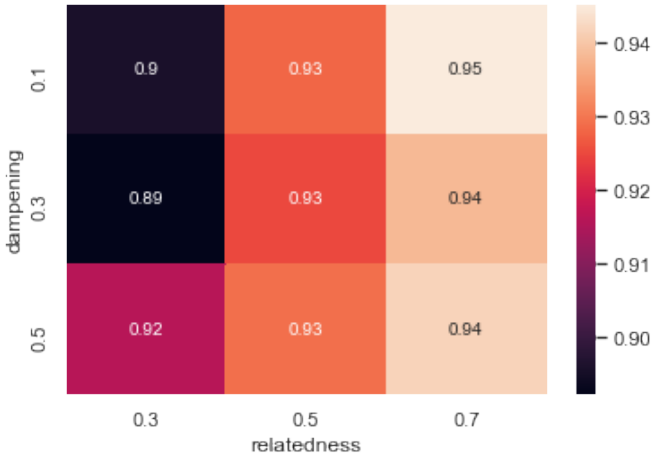


Fig. 8: Grid search heatmap, accuracy displayed by color

The model achieved maximum accuracy with the parameters $\sigma = 0.1$ and $\rho = 0.7$, meaning giving little account to source dependency, but much more to fact similarity. This way, the model achieved a marginally better accuracy of 94.5% on the Golden Standard, as Figure 7 shows.

The heatmap presented in Figure 8 shows how the parameter having the most influence on accuracy is the relatedness factor. This can be interpred in different ways. One hypothesis could be that in the dataset, collected in 2007, sources copying each other wasn't as diffused as today, making source dependancy a lesser problem in the dataset. Another hypothesis may be that the dataset presents a bias towards similar authors, making it easier for the parameter to prevail on the other in the confidence computation.

Both implemented similarity functions have been experimented, but cosine similarity outperformed the ngram-based similarity by a margin of 3-4% on Data Fusion accuracy, while also being faster.

## VI. Conclusions

TruthFinder presented great performances on the dataset, reaching a 95% accuracy in the Data Fusion process after an additional parameter search step. The results show how, given consideration towards the attributes to reconcile through appropriate preprocessing, it's possible to find truthful values even for structured data (in this case, a list of authors). Future works may account: a more extensive parameters search including other parameters, such as base similarity; analysis of different similarity functions and how they impact different types of attributes; extension of the algorithm to its more modern evolutions (Source Selection through Marginalism, Source Dependancy Analysis).

## References

[1] P. S. Y. Xiaoxin Yin, Jiawei Han, "Truth discovery with multiple conflicting information providers on the web," 2007.

[2] L. D. Xiaoxin Yin. Books dataset, http://lunadong.com/fusionDataSets.htm.