Data Fusion through Truth Discovery

Master Project for the Data Management Course @UnimiB • Nassim Habbash (808292)

Task

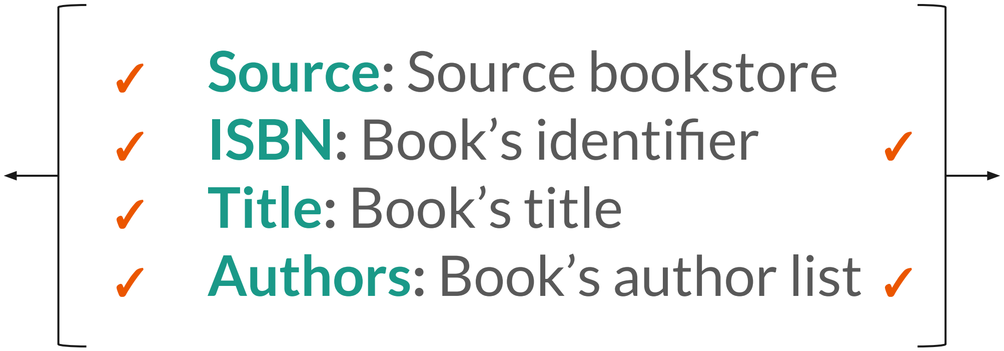
- **Design** of a Data Fusion pipeline
- **Implementation** of a Data Fusion model for conflict resolution and truth discovery
- **Analysis** of its performances on a benchmark dataset

Dataset

Composed by a **Main Corpus** and a **Golden Standard**

Main Corpus:

- 33971 records
- 1263 unique books
- 894 sources



✓ **Source:** Source bookstore

✓ **ISBN:** Book’s identifier ✓

✓ **Title:** Book’s title

✓ **Authors:** Book’s author list ✓

Golden Standard

Precise author lists manually on 100 randomly selected books

Preprocessing

Big, dirty data

The main corpus contains many heterogeneities in both the **Title** and **Authors** fields.

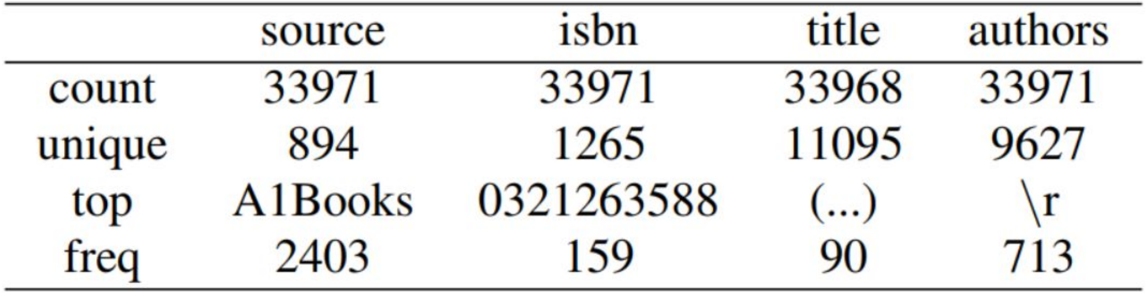
- Different naming conventions
- Different listing styles
- Unescaped HTML symbols
- (Others...)

Cleaning procedure

1. Escaping HTML characters (e.g. &amp; → &)
2. Return characters removal
3. Lowercasing
4. Parenthesis removal
5. Separators replacement
6. Special characters removal
7. Digits removal (Only on Authors)
8. Trailing whitespace removal
9. Missing values uniformation

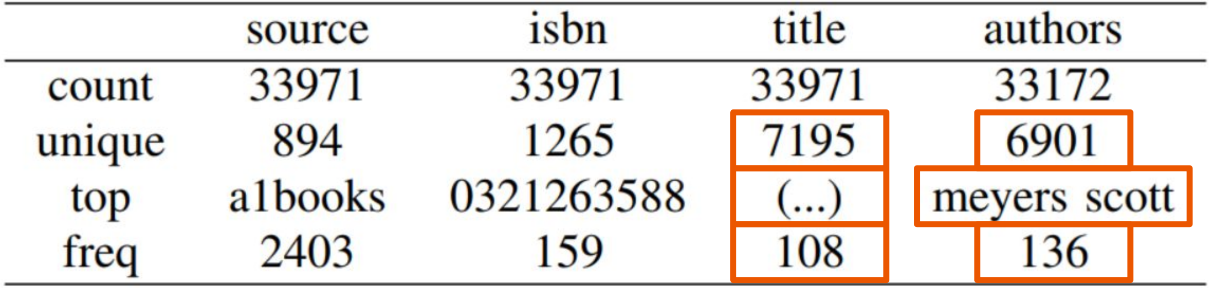
Data Exploration

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

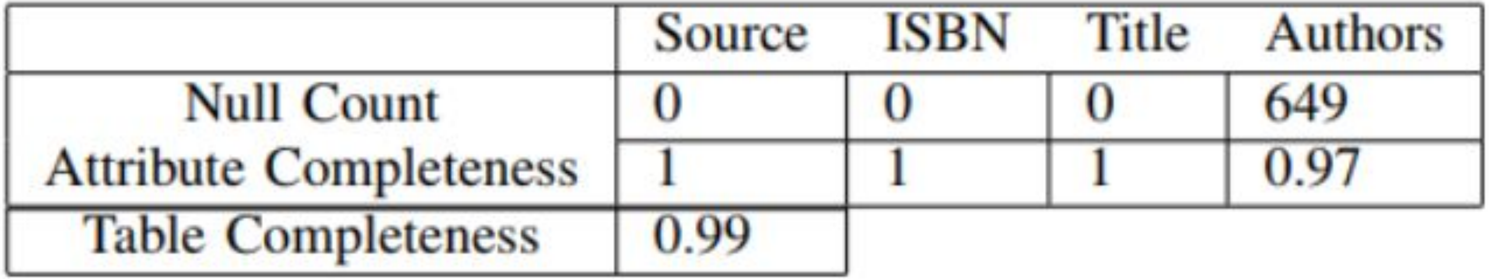
After preprocessing



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

Data Quality Dimensions

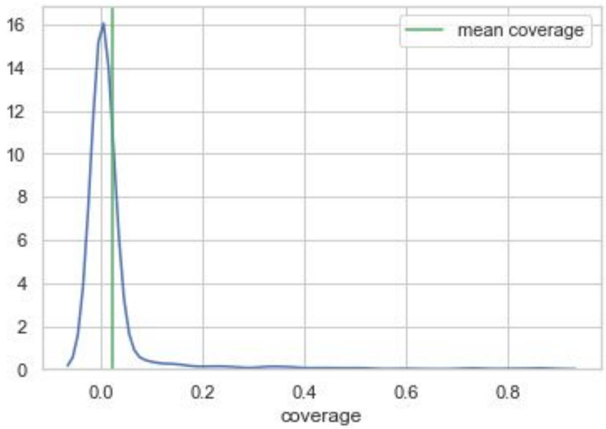
Completeness

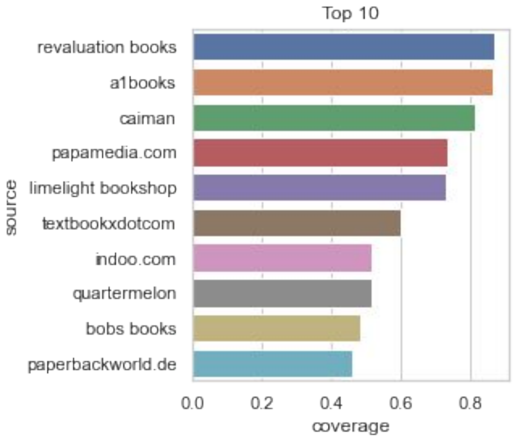


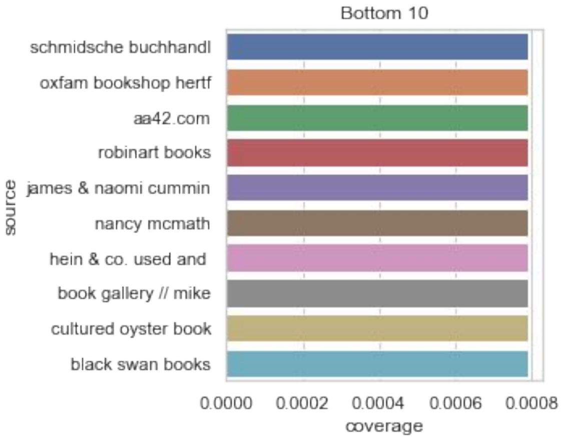
	Source	ISBN	Title	Authors
Null Count	0	0	0	649
Attribute Completeness	1	1	1	0.97
Table Completeness	0.99			

Data Quality Dimensions

Coverage: how many **unique** books (ISBN) does each source **cover**







Underlines one of the issues with big data

Truth Discovery

TruthFinder

Veracity of data

It's **hard** to ensure quality, accuracy and trustworthiness of big data.

- Which source is most trustworthy?
- Which value is the true value?
- Are sources copying each other?
- And so on...

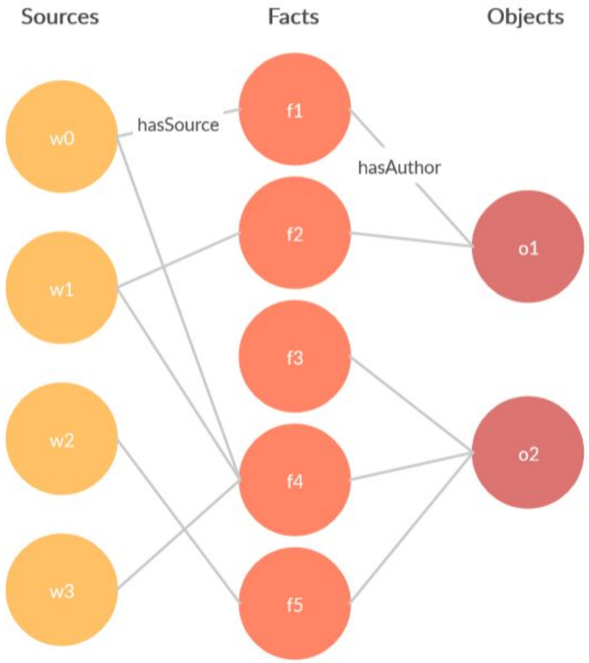
Possible solution

Make use of the **relationships** between **sources** and the **facts** they claim in a **probabilistic framework** to find the most probable **true facts**.

TruthFinder

Based on the following intuitions:

1. There is **one true fact** for a property of an object
2. This true fact appears to be the **same or similar** between **different** sources
3. The **false facts** in **different** sources are **less** likely to be **similar**
4. In a certain domain, a source that **provides mostly true facts** for many objects will likely **provide more true facts** for other objects



TruthFinder

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

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

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

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

$$\tau(w) = -ln(1 - t(w)) \qquad \sigma(f) = -ln(1 - s(f))$$

The model is based on the computation of **fact confidence** and **source trustworthiness**

s(f)	Confidence probability of fact f
σ(f)	Confidence score of fact f
t(w)	Source trust. probability of source w
т(w)	Source trust. score of source w
γ	Damping factor
ρ	Relatedness factor
imp(f’, f)	String similarity between facts f’ and f

TruthFinder

There’s a **dependency** between fact confidence and source trustworthiness - i.e. we can’t compute one without the other

Solution: iterative computation of both until stability

Initialization source trustworthiness at some value *initial_trust*.

Given the source trust scores at time ***i***, and the source trust scores at time ***j=i+1***, the process has converged if the error, defined as:

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

is **lower** than a set **tolerance** threshold

TruthFinder - Implementation

Tools

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

Data Fusion and Results

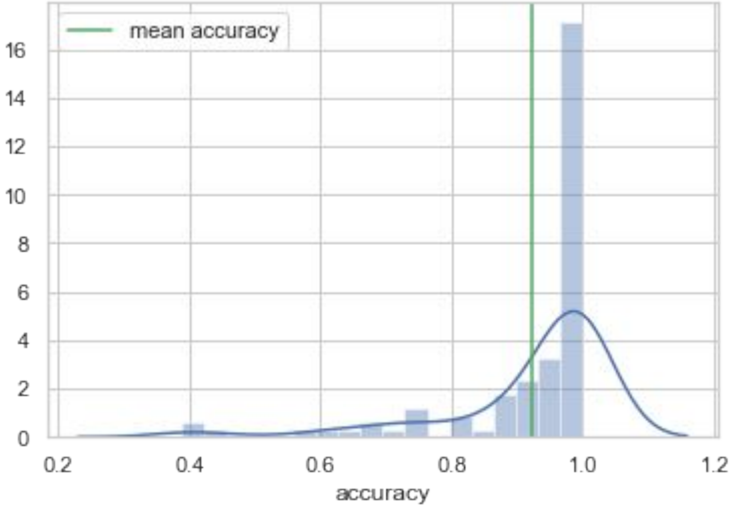
Data Fusion

The model has been first run with the parameters given from the original paper:

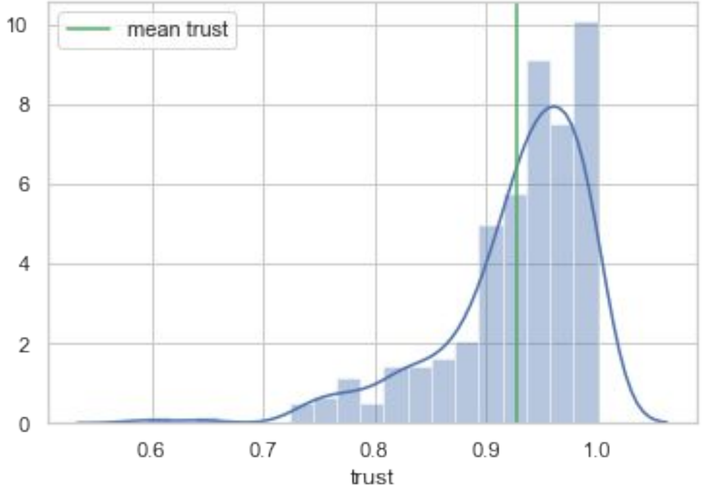
- Damping factor: **0.3**
- Relatedness factor: **0.5**
- Base similarity: **0.5**
- Max iterations: **10**
- Tolerance: **0.001**
- Initial trust: **0.9**
- Implication function: **cosine similarity**

Data Fusion accuracy is measured as the **average** of the accuracy of facts returned by TruthFinder for object o to the true fact for the same object in the Golden Standard (acting as a groundtruth)

Data Fusion - Results

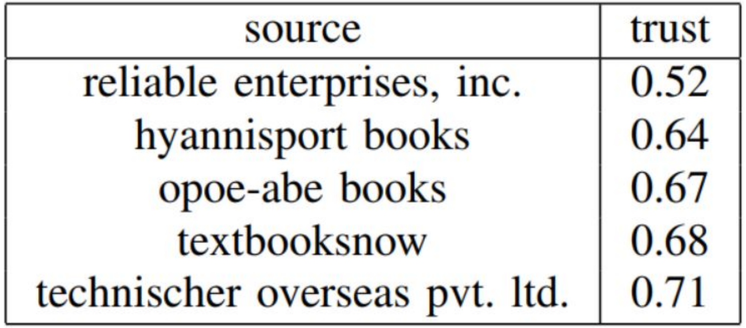


Accuracy distribution, mean accuracy of 92%



Source trust distribution, mean trust of 93%

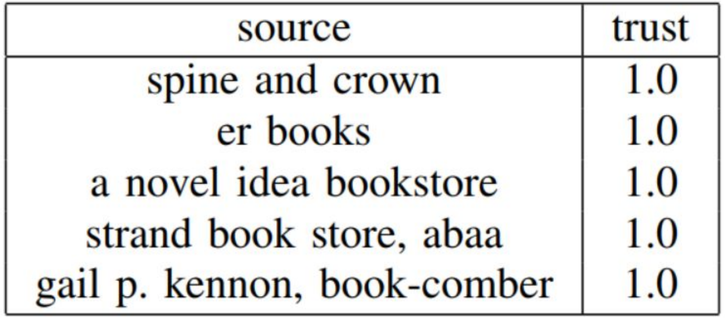
Data Fusion - Results



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

Bottom 5 sources by trust:

Most facts reported by these sources are incorrect



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

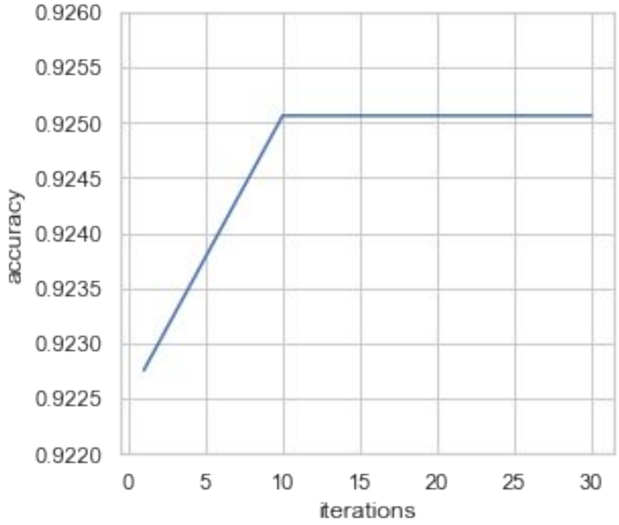
Top 5 sources by trust:

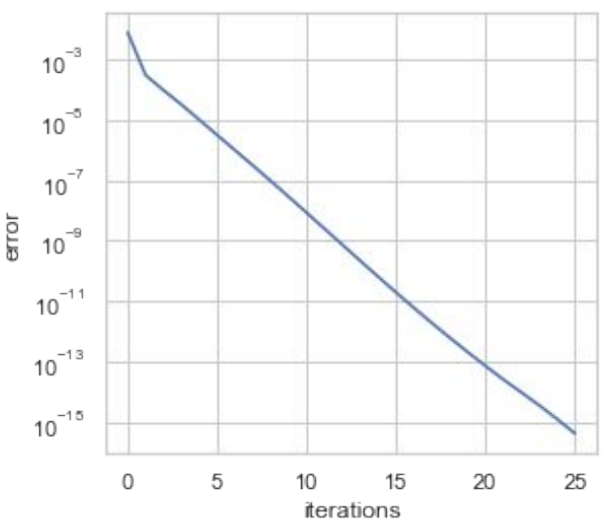
Most facts reported by these sources are correct

Data Fusion - Results

The initial tolerance (0.001) brought to convergence after 1 iteration

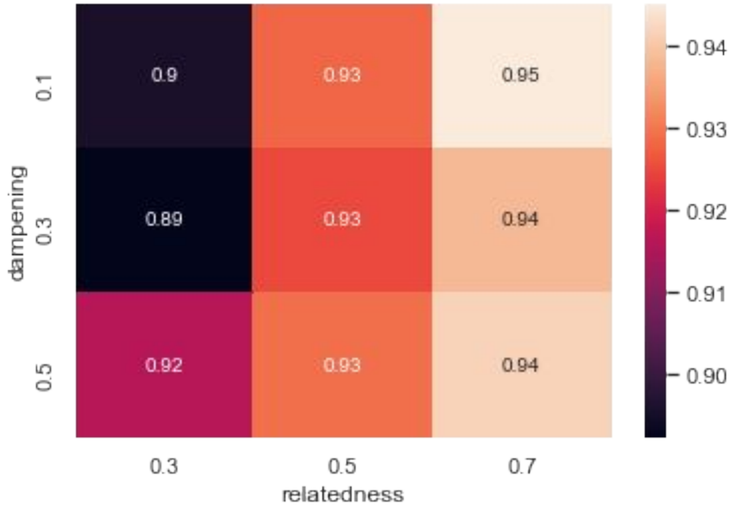
Analysis of number of iterations towards error and accuracy for tolerance=0.01 (other parameters are the same)



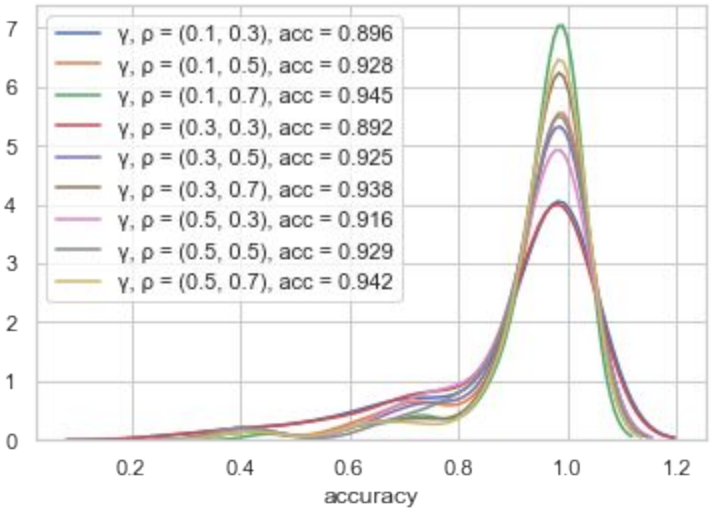


Data Fusion - Parameters search

Grid search for damping and relatedness factor, initial_trust = 0.8



Accuracy change in function of damping and relatedness factors



Accuracy distributions for different models

Conclusions

1. After a grid search, the model achieved a **Data Fusion Accuracy of 95%**.

2. The **relatedness factor** in the dataset is **more influential** than the damping factor

3. Different similarity functions might work differently, as the original paper applied a **weighting** towards **Authors names parts**

4. Possible future works might include: **extension** and **comparison** of TruthFinder to more modern applications, such as Source Selection through Marginalism, Source Dependency with Bayesian nets

Repository: *https://github.com/nhabbash/truth-discovery*