Krushak Odisha

Truth Finder Confidence Algorithm

1. Introduction:

The Truth Finder Algorithm [1] is a type of Truth discovery method[2] which is the process of extracting the true value from a set of data sources that provide conflicting information. These methods often calculate a confidence i.e. probability of value being true for each fact provided by the data sources and pick the fact with the highest confidence as the 'true value'. We can leverage these methods by calculating the confidence of the Krushak Odisha values in the algorithm.

2. Algorithm Premise:

The algorithm is based on these 4 premises:

Premise 1: Usually there is only one true fact (Value) for a property (Data Field) of an object (Farmer Record).
We assume that there is only one true fact for a property *(data field)* of an object *(farmer)*.
For example: There can only be 1 true value for name of a farmer.

Premise 2: This true fact (Value) appears to be the same or similar on different sources. Different sources that provide this true fact may present it in either the same or slightly different ways.
For example: if 'Aman' is the true value, then there are likely to be multiple data sources saying 'Aman' and some sources providing similar names like 'Amin'.

Premise 3: The false facts (Values) on different sources are less likely to be the same or similar: Amongst a set of facts, the sub-set facts that doesn't match with any of the others are unlikely to be true.
For example: For one farmer record, 3 data sources with different values saying 'Bob', 'Chandru', and 'Dave' are likely to be all false values

Premise 4: A source that provides mostly true facts for many objects will likely provide true facts for other objects. There are trustworthy sources such as Aadhaar and untrustworthy sources. A source that is correct for many objects is more likely to be correct for other objects
For example: if 'Aadhaar' is giving truest value of Name for 80% of farmers, it is more likely than other sources to also give true values for 20% of the remaining farmers.

3. Confidence Calculation:

Defining some terms in the formulae:

Object: An object is whatever the informations is being record about - *in our case each farmer record existing in KO*

Property: Property of an object is the characteristic that we have values for
Ex: Farmer Name, Primary mobile number, Landholding size, etc. are properties in KO of a given object 'Farmer1'

f: Fact: Values provided by a source for the property of an object.
Ex.: Aman, Bob, Chandru etc are facts from the 'Farmer Name' property of object 'Farmer1'

f': Other Facts: Facts apart from the Krushak Odisha value for which the confidence level is being calculated
Ex.: Land record area value in PMFBY is the other fact for KO's 'Farmer's Area Under Cultivation' property

t(w) : Trustworthiness of the source: This determines the reliability of one property of a particular source.
Ex: t(w) of Aadhar for names is 0.99, t(w) of source A for names is 0.8, t(w) for Aadhar for age is 0.97

w: Source: Each attestation source from which fact f is derived
Ex.: PMFBY is a source for fact 'Land record area value'

W: List of all sources: List of all attestation sources available for a property
Ex.: PMFBY,Seed Supply, P-PAS, M-PAS are sources W for fact KO's Kharif Crops property

s(f) : Unadjusted confidence: Confidence of a fact f considering only those sources which have the fact

s*(f) : Adjusted confidence: Confidence of a fact f adjusted for other facts f'

s0*(f): Scaled Adjusted Confidence: Adjusted confidence of fact that has been scaled logistically (from 0-1). This is the final score that will be used as the confidence level

τ(w) : Trustworthiness score of the source: This is a log transformed version of trustworthiness t(w) of the source w to account for underflow. Underflow is when extremely low values are created by the multiplication of the low(1-t(w)) values with each other. The extremely low values are often rounded off to zero by programs while calculating leading to unexpected errors

E.g, If we have 10 matching sources and their t(w) is 0.99, then the product of (1-t(w)) will become 1^{-10}

$$\tau(w) = -\ln(1 - t(w)) \quad (1)$$

σ(f): Unadjusted confidence score for a fact: This is a logarithmic transformed version of unadjusted confidence s(f) again to prevent underflow

$$\sigma(f) = -\ln(1 - s(f)) \quad (2)$$

σ*(f): Adjusted confidence score for a fact: This is a logarithmic transformed version of adjusted confidence s*(f) to prevent underflow

$$\sigma^*(f) = -\ln(1 - s^*(f)) \quad (3)$$

ɣ : Damping factor: This is a parameter to be provided to the model to account for lack of independence amongst sources. We will be considering its value as 1 for now, assuming that the attestation sources are completely independent of each other. In case of any interdependence between two or more attestation sources, we can check for various parameters of ɣ (from 0-1)

imp(f'→f): Impact of f' on f: This is the effect of other facts f' on fact f. In our case, we can define it as -1 all the time, assuming that we require all facts to exactly match with each other. This measure can be improved to use as a similarity score scaled from 1 to -1.
Ex: if we have Krushak Odisha value as Aman and Source A fact 'Amana', impact of fact 'Amana' on 'Aman' can be -0.2 and impact of 'Bob' on 'Aman' can be -1. Hence Krushak Odisha value of 'Aman' will have a higher confidence score if source A says 'Amana' rather 'Bob'

The confidence score can be calculated from the trustworthiness of each data sources as:

$$s_0^*(f) = \frac{1}{1 + e^{\sigma(f) - \sum_{f'} \sigma(f')}} \tag{4}$$

where,

$$\sigma(f) = \sum_{w \in W(f)} \tau(w) \tag{5}$$

and,

$$\tau(w) = -\ln(1 - t(w)) \tag{1}$$

Illustrative Example for calculating confidence:
We want to calculate the confidence for a farmer's name and we have corresponding values available from 4 different attestation sources A, B, C, and D as Bob, Aman, Charan, and Aman as shown below. Let us assume we have trustworthiness values for each data source as shown below. We can calculate the trustworthiness score from the above equations:

	Krushak Odisha	Source A	Source B	Source C	Souce D
	Aman	Bob	Aman	Charan	Aman
t(w)	-	0.90	0.82	0.74	0.78
1-t(w)	-	0.1	0.18	0.26	0.22
τ(w) = -ln(1-t(w))	-	2.30	1.72	1.36	1.53

Image - 3.1

We have 3 facts here - 'Bob', 'Aman' and 'Charan'
From the equations above, we can calculate the unadjusted confidence scores as:

σ(Bob) = τ(Source A) = 2.3 *(from equation 5)*
σ(Aman)= τ(Source B) + τ(Source D) = 1.72 + 1.53 = 3.25 *(from equation 5)*
σ(Charan)= τ(Source C) = 1.36 *(from equation 5)*

Here, as Krushak Odisha value is 'Aman' for which we need to calculate the confidence,
f= 'Aman' and f' = {'Bob','Charan'}

Hence, to find confidence from equation 4, we can calculate $\{\sigma(f) - \Sigma\sigma(f')\}$ first which is within the denominator (needs to exponentiated)

$\sigma(f) - \Sigma\sigma(f') = \sigma(Aman) - \sigma(Bob) - \sigma(Charan) = 3.25 - 2.3 - 1.36 = -0.41$

Final confidence = $1/(1+e^{-(-0.41)}) = 0.39$ *(from equation 4)*

4. Algorithm iteration:

The algorithm is iterative, gradually improving the estimate of the confidence score and the trustworthiness of the sources. It considers trustworthiness of a source to be the average confidence of all the facts provided by it.

$$t(w) = average(s_0 * (f))$$

Steps to run the algorithm:

Step I: Assume t(w) for all sources = 0.5
Step II: Update the confidence values for all the objects using the above calculations
Step III: Update the t(w) for all the sources by taking the average of confidence for each source
Step IV: Repeat the process until t(w) doesn't change from the last iteration
Step V: Calculate the final confidence values from t(w)

Sno	Krushak_Odisha	Source_A	Source_B	Source_C	Source_D
1	Dhoni	MS	MS	MS	MS
2	Sehwag	Sehwag	Viru	Sehwag	Sehwag
3	Gambhir	Gautam	Gautam	Gautam	Gautam
4	Sachin	Sachin	Sachin	Sachin	Sachin
5	Yuvaraj	Yuvaraj	Yuvaraj	Yuvaraj	Yuvaraj
6	Raina	Raina	Suresh	Raina	Raina
7	Kohli	Virat	Kohli	Kohli	Virat
8	Nehra	Ashish	Ashish	Ashish	Ashish
9	Yusuf	Yusuf	Yusuf	Yusuf	Yusuf
10	Munaf	Munaf	Patel	Patel	Patel
11	Sreeshant	Sreeshant	Sreeshant	Sreeshant	Sreeshant
12	Chawla	Chawla	Piyush	Chawla	Piyush
13	Ashwin	Ashwin	Ashwin	Ravi	Ashwin
14	Dravid	Dravid	Dravid	Rahul	Dravid
15	Ganguly	Ganguly	Ganguly	Ganguly	Saurav
16	Agarkar	Agarkar	Agarkar	Agarkar	Ajit
17	Karthik	Karthik	Karthik	Karthik	Karthik
18	Kumble	Kumble	Kumble	Anil	Anil
19	Irfan	Irfan	Irfan	Pathan	Irfan
20	Uthappa	Uthappa	Uthappa	Robin	Uthappa
21	Kaif	Kaif	Kaif	Kaif	Kaif
22	Laxman	Laxman	VVS	VVS	Laxman
23	Mongia	Mongia	Naman	Dinesh	Mongia

Image - 4.1: Sample Dataset for property 'Farmer Name'

Step I:
We assume that t(w) of all sources = 0.5

Iteration 1 :
t(w):

	Source_A	Source_B	Source_C	Source_D
0	0.5	0.5	0.5	0.5

Image - 4.3

Step II: Update the confidence values for all the objects using the equation 4

Image 4.3 shows a set of rows from the data considered and *Image 4.4* shows the calculated confidence scores

Source_A	Source_B	Source_C	Source_D
Sachin	Sachin	Sachin	Sachin
Raina	Suresh	Raina	Raina
Munaf	Patel	Patel	Patel
Kumble	Kumble	Anil	Anil
Laxman	VVS	VVS	Laxman
Mongia	Naman	Dinesh	Mongia

Image - 4.3

	Source_A	Source_B	Source_C	Source_D
3	0.9412	0.9412	0.9412	0.9412
5	0.8000	0.2000	0.8000	0.8000
9	0.2000	0.8000	0.8000	0.8000
17	0.5000	0.5000	0.5000	0.5000
21	0.5000	0.5000	0.5000	0.5000
22	0.5000	0.2000	0.2000	0.5000

Image - 4.4

Step III: Update the t(w)for all the sources by taking the average of confidence for each source:

The new t(w) values can be calculated by taking the average of the confidence calculated for each of the sources:

	Source_A	Source_B	Source_C	Source_D
1	0.763939	0.724808	0.672634	0.737852

Image - 4.5

Step IV: We need to keep iterating i.e. repeat the steps starting from *Step II* but considering the t(w) values that were calculated in *Step 3*:

Repeating *Step 2* with the new t(w) values, we get :

Source_A	Source_B	Source_C	Source_D
Sachin	Sachin	Sachin	Sachin
Raina	Suresh	Raina	Raina
Munaf	Patel	Patel	Patel
Kumble	Kumble	Anil	Anil
Laxman	VVS	VVS	Laxman
Mongia	Naman	Dinesh	Mongia

Image - 4.6

	Source_A	Source_B	Source_C	Source_D
3	0.9945	0.9945	0.9945	0.9945
5	0.9314	0.0686	0.9314	0.9314
9	0.0909	0.9091	0.9091	0.9091
17	0.5692	0.5692	0.4308	0.4308
21	0.5928	0.4072	0.4072	0.5928
22	0.5928	0.0686	0.0494	0.5928

Image - 4.7

And then calculating the new t(w) values , we get :

	Source_A	Source_B	Source_C	Source_D
2	0.842887	0.765989	0.675967	0.800032

Image - 4.8

We must keep iterating this process until the t(w) value stops changing. *Image - 4.8* is showing the t(w) values for 10 iterations:

	Source_A	Source_B	Source_C	Source_D
0	0.500000	0.500000	0.500000	0.500000
1	0.763939	0.724808	0.672634	0.737852
2	0.842887	0.765989	0.675967	0.800032
3	0.870441	0.763453	0.662284	0.818919
4	0.883730	0.757941	0.655355	0.826934
5	0.890666	0.754483	0.652289	0.830245
6	0.894357	0.752660	0.650900	0.831473
7	0.896343	0.751735	0.650247	0.831874
8	0.897422	0.751263	0.649930	0.831981
9	0.898011	0.751019	0.649771	0.831995

Image - 4.9

As we can see above, it starts with 0.5 for all sources and then keeps changing with every iteration. However the degree of change keeps reducing with each iteration and by the tenth iteration, the degree of change is negligible and the algorithm is stopped.

We've put the condition that the algorithm stops when the difference between t(w) of 2 consecutive iterations is less than 0.001

Step V: We calculate the final confidence values based on the last iterations' t(w) values using Equation 4:

Source_A	Source_B	Source_C	Source_D
Sachin	Sachin	Sachin	Sachin
Raina	Suresh	Raina	Raina
Munaf	Patel	Patel	Patel
Kumble	Kumble	Anil	Anil
Laxman	VVS	VVS	Laxman
Mongia	Naman	Dinesh	Mongia

Image - 4.10

	Source_A	Source_B	Source_C	Source_D
3	0.9985	0.9985	0.9985	0.9985
5	0.9765	0.0235	0.9765	0.9765
9	0.1256	0.8744	0.8744	0.8744
17	0.6985	0.6985	0.3015	0.3015
21	0.8358	0.1642	0.1642	0.8358
22	0.8358	0.0235	0.0120	0.8358

Image - 4.11

We can the calculate the KO values confidence by matching the KO values to the source:

	Krushak_Odisha	Source_A	Source_B	Source_C	Source_D	KO_confidence_score
3	Sachin	Sachin	Sachin	Sachin	Sachin	0.998508
5	Raina	Raina	Suresh	Raina	Raina	0.976465
9	Munaf	Munaf	Patel	Patel	Patel	0.125601
17	Kumble	Kumble	Kumble	Anil	Anil	0.698536
21	Laxman	Laxman	VVS	VVS	Laxman	0.835773
22	Mongia	Mongia	Naman	Dinesh	Mongia	0.835773

Image - 4.12

Annexure:

1. **Why are we following the Truth Finder Algorithm to calculate the confidence level of our data points? Is there any precedence establishing that this is the best/ideal way to go about it?**

- The Truth Finder algorithm is a well-known research paper cited extensively in the data science world [1] including by Google for it's Knowledge-based-Trust Algorithm [2] which is Google's patented method of carrying out truth-discovery[8] to improve its search results. It's based on the same iterative principles as Page-Rank but considering the reliability of websites instead of number/quality[3].

- Model is completely data driven and does not require estimating any accuracy parameters

- Model is relatively simple to implement, runs quickly and does not require expensive infra

- Model works well with low number of data sources (most truth discovery models are built for websites and often scrape from 1000's of websites to check)

- The model has been implemented previously as open source code in Java/R [4] [5]

2. Where has this model been implemented previously? Can we get some test results or evidence of success for this model?

- The Truth finder consistently displays good results for truth discovery on real world data sets: In this work[6], it was tested on the following datasets:

    - **The AbeBooks data set** : It's a comparison of author details for computer science books extracted from AbeBooks websites in 2007 It consisted of 33,235 claims on the author names of 1,263 books by 877 book seller sources.
      The 'true value' was available for 100 randomly sampled books for which the book covers were manually verified by the authors.

      The Truth finder algorithm had an accuracy of 94% which was 2nd best amongst algorithms compared with the least computation time

    - **Weather data set :** The Weather data set consists of 426,360 claims from 18 sources on the Web for 5 properties (temperature, humidity etc) on hourly weather for 49 US cities between January and February 2010. 'True value', was deemed to be the AccuWeather website values which were available in 75% of the cases

      Truth Finder had an accuracy of 86% which was the best amongst all algorithms compared with the least computation time

    - **Biography data set :** The Weather data set consists of 9 biography details (father name, mother name, age etc) extracted from Wikipedia with 10,862,648 claims over 19,606 people and 9 attributes from 771,132 online sources

      Truth Finder had an accuracy of 90% which was the 2nd best amongst all algorithms compared with the least computation time

    - **Biography data set :** The Population data set consists of 49,955 claims on city population extracted from Wikipedia edits from 4,264 sources. The 'true value' was considered to be the official US census data.

      Truth Finder had an accuracy of 87% which was the 2nd best amongst all the algorithms