Shemon's meeting note

Seokki Lee

December 8, 2021

Contents

1 *<2021-12-08 Wed>*

1.1 Approximate computation for IG

- R = { (1,a), (1,a'), (2,b), ... }
  - Fine-grained (FG): 1 -> 000, a -> 001, a' -> 010, 2 -> 011, b -> 100, ...
  - Coarse-grained (CG): 1 -> 000, a -> 001, a' -> 001, 2 -> 011, b -> 100, ...
- Annotate a group of values that are similar enough with same annotation
- r_1^o = { (1,a), (1,a'), (1,b) }
  - FG: r_1^o = { (000,001), (000,010), (000,100) } -> 000·001 + 000·010 + 000·100
  - CG: r_1^o = { (000,001), (000,100) } -> 000·001 + 000·100
- r_1^i = { (1,a), (1,a'), (2,b) }
  - FG: r_1^i = { (000,001), (000,010), (011,100) } -> 000·001 + 000·010 + 011·100
  - CG: r_1^i = { (000,001), (011,100) } -> 000·001 + 011·100
- Above can also reduce the width of records.
- Another example
  - r_1^o = { (000,001) }
  - r_1^i = { (000,001), (011,100) }
- TODO
  - What are the "correct" similarity metrics?
  - What is the bottom line of our accuracy?
- Optimizations
  - Grouping based on similarity
  - Pruning out non-candidates for distance measure

1.2 TODO

- Why do we need provenance?
  - Accuracy (example in the proposal)
- Why should we use provenance for measuring IG?
- 2 ways to apply the approximation above
  - Similarity over actual values and then assign provenance annotations for each group
  - Annotate each value first and then apply similarity measure over provenance annotation

1.3 Challenges on hashed value size

- Any metric that bounds the hash size?
  - LSH: locality sensitivity hash
  - Different hash size for each distinct value
    - Then, how to correctly compute the distance among different sized hash values?

2 *<2021-11-24 Wed>*

2.1 Provenance annotation extension

- Hashing as a function for generating provenance annotation
- Extend the provenance notation to encode the query operation using the annotation
- How to express range values using the extended notation
- Confirm if the distance can be measured by just looking at the provenance annotation
  - Generalization is needed to represent IG as a unit interval
  - But, no normalization is needed.

2.2 Fuzzy provenance annotation!

- LSH: locality sensitivity hash
- Using LSH, we can group similar valued into same buckets

3 *<2021-11-19 Fri>*

3.1 Value based annotation

3.2 Where + how provenance needed!

4 *<2021-11-10 Wed>*

4.1 Research on finding papers for information gain metrics

- Existing works from information retrieval
- Definition of information gain
- What would be the correct measure among strings?
  - Directly applying entropy and SED doesn't work.

4.2 Main contributions: how to use provenance to improve

- Develop provenance model
  - TODO: find the cell-level provenance model, aka. where provenance
- How to efficiently capture the new provenance
- Based on the new provenance model,
  - how to efficiently compute IG (improvement on performance computing IG)
  - how to accurately compute IG (improvement on quality of IG)

4.3 Set-up

- latex for writing papers
  - You can use some tools, e.g., emacs, MacTex, Sublime Text, etc, whichever convenient for you.

- Repository for the paper

5 *<2021-11-03 Wed>*

5.1 To reconfirm

- Similarity measure over numbers and strings
- Extension of SED or ED based on Shemon's observation
- Normalization matric that resovle the issue

5.2 TODO for Seokki

- Meaning of new information such that what is considered as a new information
  - For example, changing alice → bob is fully new information.
  - What about peter → tom? Although one charactor is overlaped, it is a complete change. Is it true?

6 *<2021-10-27 Wed>*

6.1 Observations from the example

- Normalizing data first before measuring the distance introduces bias
  - e.g., the salary 190 in the original dataset is treated same as the salary 200 in the integrated dataset.
- Think of how to obtain the meaning of IG over normalized data
  - e.g., if the normalized value is 1, can we confirm that completely new value is introduced?
- Applying other normalization metrics
- How do we apply the ED if the integrated data contains range values
  - e.g., age and salary may have range values.
  - Naive way: increase the rows by the number of combinations of min and max values in the range
  - Optimal way?

7 *<2021-10-22 Fri>*

7.1 Create an example

- Using the example data we have (e.g., D_o(name, age, salary) and D_i(name, age, salary))
- Algorithm to compute IG
  - Separate string and number columns
  - Apply SED (for string columns) and ED (number columns with considering as multidimensional) accordingly
  - Normalize the distance seprately
  - TODO: think of how to combine the normalized values
  - Note: if the values are range, compute the distance for lower and upper bound for IG

7.2 "where" provenance

- Any techniques for capturing cell-level provenance over data integration
- Any notion over where provenance to encode the alternative of values
  - e.g., $p_A \cdot p_B \cdot x_C$ or $p_A \cdot x_B \cdot x_C$
- Reference
  - Provenance in databases: Why, how, and where
  - You Say What, I Hear Where and Why? (Mis-)Interpreting SQL to Derive Fine-Grained Provenance

8 *<2021-10-13 Wed>*

8.1 Normalization over strings

- How to translate strings to, e.g., numbers?
  - Any existing methods?
  - If possible, then apply the same normalization metric used over the numbers.

8.2 "where" provenance

- Any techniques for capturing cell-level provenance over data integration
- Any notion over where provenance to encode the alternative of values
  - e.g., $p_A \cdot p_B \cdot x_C$ or $p_A \cdot x_B \cdot x_C$

9 *<2021-10-06 Wed>*

9.1 Observation

- New provenance notions that represent the cell-level provenance information like where provenance
  - Correctness of the IG measure
  - Performance improvement: looking at the provenance annotation, we can filter out unnecessary evaluation.

9.2 TODO

- Normalization over different metrics for IG computation
  - Can we have $0 < \text{IG} < 1$ overall?
- Any existing approaches capturing where provenance over integration process?

10 *<2021-09-30 Thu>*

10.1 What can we do with provenance?

- Provide correct mapping between intergrated data and owned data
- Improve the computation by avoiding arbitrary combination of comparisons
- Need to think
  - Possible that provenance can play a role like weight function? (example over multiple datasets)
  - ...

10.2 IG over range values

- Use ED for numbers by considering as 2 dimensions
  - x is a particular cell
  - owned dataset: x $<$ 80
  - integrated dataset: 40 $<$ x $<$ 80
  - 0 $<$ IG $<$ 40
    - all values in x between 0 and 40 are not possible or correct
    - the probability of having correct age is improved
    - TODO: any other meaning coming out from this?
- For string, 2 cases
  - SED over cell in the case that one value is chosen or concatenated
  - using SED, we can also compute the range of IG, e.g., $O(s_1)$, $I(s_1, s_2) \rightarrow 0 <$ IG < 2

10.3 TODO for next

- normalization is needed over the IGs computed from ED, SED, and over range values
- related work: existing information gain metrics

11 *<2021-09-23 Thu>*

11.1 Think about case 6

- Any case that the integration returns null value?
  - Meaningful to confirm incorrect values
  - However, the owner for example may feel upset when the null value is revealed.

11.2 Findings

- IG needs to be measured for both parties (differs depending on the position).
- IG over multiple values

  - Computing IG as a range
  - number: Euclidean distance in 2-dimension
  - string: Edit distance

- IG over range values
  - number: TODO: find a metric to compute
  - string: NA
  - Check if the data integration returns a particluar value over the range.

12 *<2021-09-16 Thu>*

12.1 Levenshtein edit distance

- Confirm whether this metric works over different integration cases
- No need to invert → higher number means more steps
- Think of the case in the excel sheet

12.2 Computing distance between range values

13 *<2021-09-09 Thu>*

13.1 Compute distance over strings

- using string edit distance, e.g., jaro winkler distance

13.2 Other cases

- Considering the distance of completely new values (currently, the distance differs but it should be considered as same, i.e., 1)
- Computing distance over range values
- How to compute the smaller bounds as higher IG
  - "invert"

14 *<2021-08-26 Thu>*

14.1 Extend the new method for measuring the distance of IG over your own datasets and integrated dataset

14.2 Confirm the extended method working over the datasets that are integrated in different ways

- Case0: no records in the own datasets match with the datasets in the shared datasets (Union applied)
  1. No new values/records
  2. All new values/records
  3. Partially new values/records
- Case1: all the records in your own datasets may match all in the shared datasets (Join applied)
- Case2: The records in your own datasets partly match those in the shared datasets (Similarity Join applied)
  - Partly 1 (only rows grow): subset of the records in your own datasets match with those in the shared dataset
  - Partly 2 (both rows and attributes grow)

14.3 Think of how to extend the method to record the contribution of each cell value

- If the current method doesn't work over the cases above

15 *<2021-08-11 Wed>*

15.1 Information gain model

- Use entropy or similiarity measure to have more concrete IG for each cell
- Try to apply it to few other example datasets to verify
- Think of how to generalize the IG (computed on each cell) over the integrated data

15.2 Data integration

- Understand how the record linkage works
  - Any existing solution can be used for our case

16 *<2021-08-03 Tue>*

16.1 Build background knowledge for the components of data integration

- Schema alignment
- Record linkage
- Data fusion
- Additional:
  - Any systems that implement the data integration
  - Any techniques that use SQL only

16.2 Measuring information gain (IG)

- Using the example datasets, can we define a model that computes IG?
- Can we generalize the developed model to other datasets?

17 *<2021-07-27 Tue>*

17.1 TA opportunity

- Fall 2021: email Prof. Wen-Ben your latest CV
- Spring 2022: I will inform you when announced

17.2 Measuring information gain (IG) using provenance

- Running example ready?
- Sample dataset + simple model until next week!
  - Create a sample normalized datasets (one for shared and owned) similar ones in the paper

- Consider multiple ways of integration (like T^{union} and T^{join} in the paper)
- Measure information gain, i.e., think of the model that we can use to express the amount of new values as a number
  - For example, IG from T^{union} in Figure 1 can be 3/10, because there are 3 new values out of 10 values.
  - Similarly, for T^{join} the IG can be 3/5.
- How do we know which one is considered as a new value?
  - We capture provenance for each of the values.
  - How to capture it efficiently is another challenge to address.