## Polygon Inequality Filters for Identifying Duplicates in XML Databases

**Newton Campbell – Sponsor: Dr. Junping Sun**

### INTRODUCTION

Data cleaning describes the process of identifying, removing, and fixing flaws in a given dataset. Particularly in data fusion and integration, multiple groups of records representing the same entity are commonly inserted into a common database as duplicates. While the allowability of duplicate records in a database depends on the domain, performing knowledge discovery in a database with duplicate objects can seriously disturb many forms of pattern recognition and machine learning.

Finding duplicate objects is a problem of identifying similarity amongst a group of objects. In a well-designed relational database, objects are typically defined as a row of a relational table, with the structure of the table clearly defining the object. In XML, while a tree structure supported by a schema exists, the semantics within a dataset are mutable.

### METHODOLOGIES

- **General Approach**
  - Object Extraction: Extract object document contain dissimilar data (XQuery)
  - Graph Generation: Represent XML in a graph structure
  - Similar Data Detection
    - **Apply edit distance filters**
      - Object Filter: Reduce object comparisons
      - Duplicate Detection: Pairwise Object Comparison
      - Duplicate Clustering: Repeat at the next level of the XML tree

- **Length Distance Filter**
  - \(|l(x)| - |l(y)| \leq \tau \)
  - \(d_{edit} x, y - d_{edit} y, z \leq d_{edit} x, z \leq d_{edit} x, y + d_{edit} y, z\)

- **Triangle Inequality Filter**
  - \(\tau = 2\)
  - \(|d_{edit} x, y - d_{edit} y, z| \leq d_{edit} x, z \leq d_{edit} x, y + d_{edit} y, z\)
  - \(\text{max} < \tau \rightarrow x \text{ and } z \text{ are similar}\)
  - \(\text{min} \geq \tau \rightarrow x \text{ and } z \text{ are not similar}\)

- **Generalized Polygon Inequality Filter**
  - \(\tau = 2\)
  - \(\text{Distance of Polygon Inequality Filter} = n - 2\)

### EXPERIMENTAL RESULTS

1. **Increased Information Experiment (w/length distance filtering)**
   - Initially use length distance filter to narrow the space
   - After the filter is used, identify how many additional strings are filtered for \(n=4, 5, 6, 7, \ldots\) separately until an \(n\) is reached that filters all strings
   - Attempt for increasing thresholds

2. **Increased Information Experiment (w/o length distance filtering)**
   - Attempt Experiment 1 without initial use of distance filter

3. **Filter Comparison Experiment**
   - Directly compare the number of pairs removed after the length distance filter for the triangle inequality filter, polygon inequality filter for \((n=4)\) and \((n=5)\), and the q-gram filter

\[
\text{Length Distance} = \text{poly}(n=4) = \text{poly}(n=5) = \text{q-gram}(n=2)
\]

To filter all results, let \(k\) equal to the number of sides of the polygon required and \(x\) equal the number of distinct pairs needed to be compared. Then

\[
\text{worst case} = \Theta(\sum_{x} x - 1)
\]

### CONCLUSIONS

In this research effort, a new class of filters for evaluating string pairs for duplicated detection was theoretically and experimentally characterized. Results were collected, identifying that this new generalized polygon inequality filter can sufficiently categorize a set of token strings as being similar or dissimilar. These results provide the capability for one filter to completely perform this duplicate detection task, with minimal need for edit distance calculations. The polygon inequality is sufficient for performing such filtering and, with increasing \(n\), can adequately prune the space, even to the point of completing the filtering. An analysis study of the detriments of each of these filters should take place, such that the optimal scenarios for generalized polygon inequality filtering can be identified.

### REFERENCES


### ACKNOWLEDGEMENTS

Special thanks to Dr. Junping Sun and Nova Southeastern University GSCIS for his support in the CISD 794 course in the Fall 2013 semester. Also, special thanks to Raytheon BBN Technologies for mentorship, guidance, and funding throughout the course of the CISD program.