Robust Identification of Fuzzy Duplicates

Authors:

Surajit Chaudhuri (Microsoft Research)
Venkatesh Ganti (Microsoft Research)
Rajeev Motwani (Stanford University)

Publication:

21st International Conference on Data Engineering (ICDE 2005)

Presented By:

David Xu

Agenda

1. Introduction to Fuzzy Duplicates
2. Overview of Machine Learning
3. Duplicate Elimination Strategy
4. Duplicate Elimination Algorithm
5. Evaluation
6. Results

Introduction - Definition

- “Fuzzy Duplicates are multiple seemingly distinct tuples which represent the same real-world entity” [1]
- Database: Distinct
- Reality: The same

ID	ArtistName	TrackName
1	The Doors	LA Woman
2	Doors	LA Woman

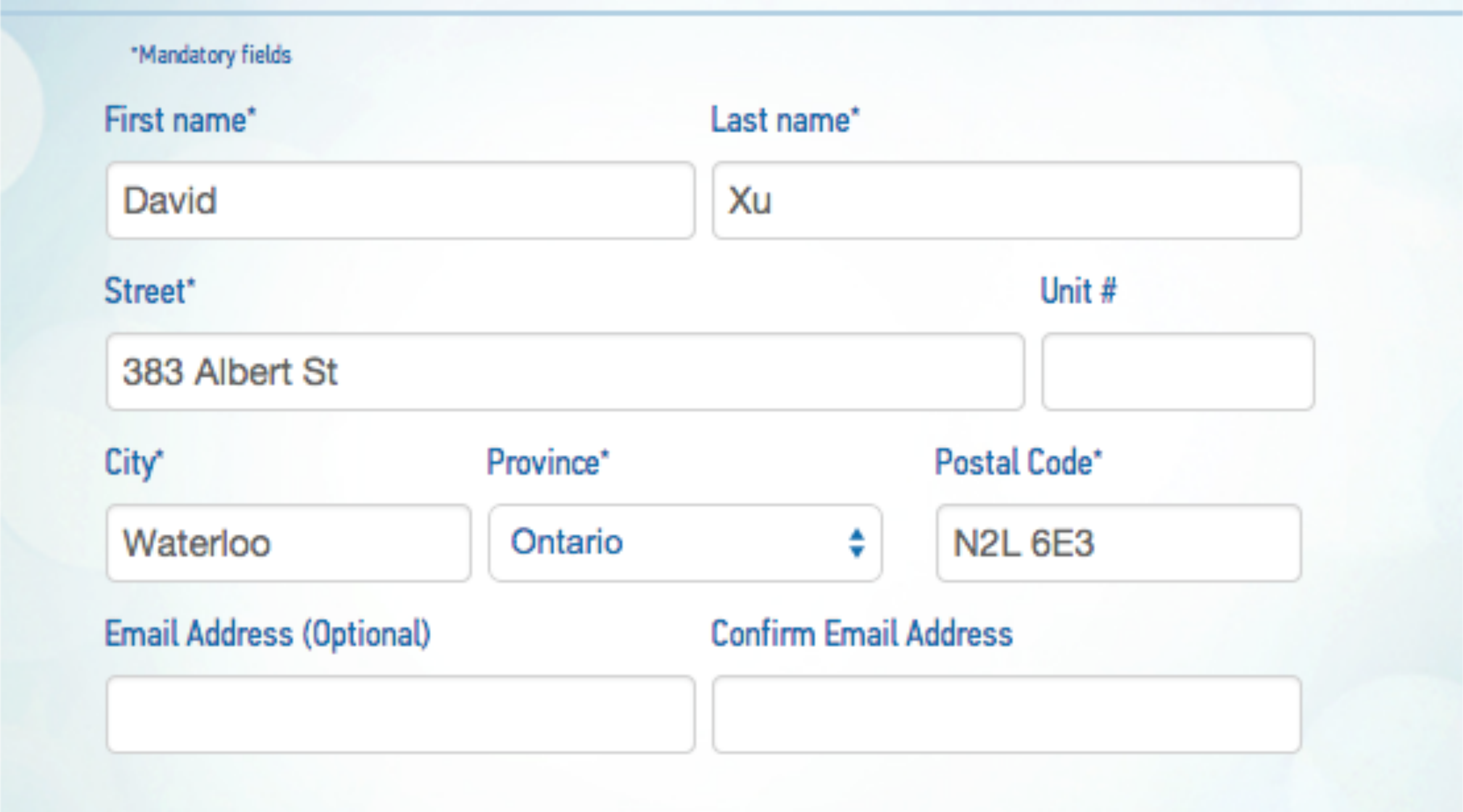
[1] S. Chaudhuri, V. Ganti, and R. Motwani. Robust Identification of Fuzzy Duplicates . In Proc. Int'l Conf. on Data Engineering (ICDE'05), 2005, pp. 865-876.

Example – Real World Example

SHIELD & REPAIR
SENSITIVE TEETH§
PLUS THE FRESHNESS YOU TRUST FROM COLGATE*.
Request a FREE sample ▸
Colgate*
Sensitive Pro-Relief™
Instant & Lasting Relief†
clinically proven PRO-ARGIN

http://www.colgatesensitiveprorelief.ca/

Example – Real World Example



*Mandatory fields

First name*
David

Last name*
Xu

Street*
383 Albert St

Unit #

City*
Waterloo

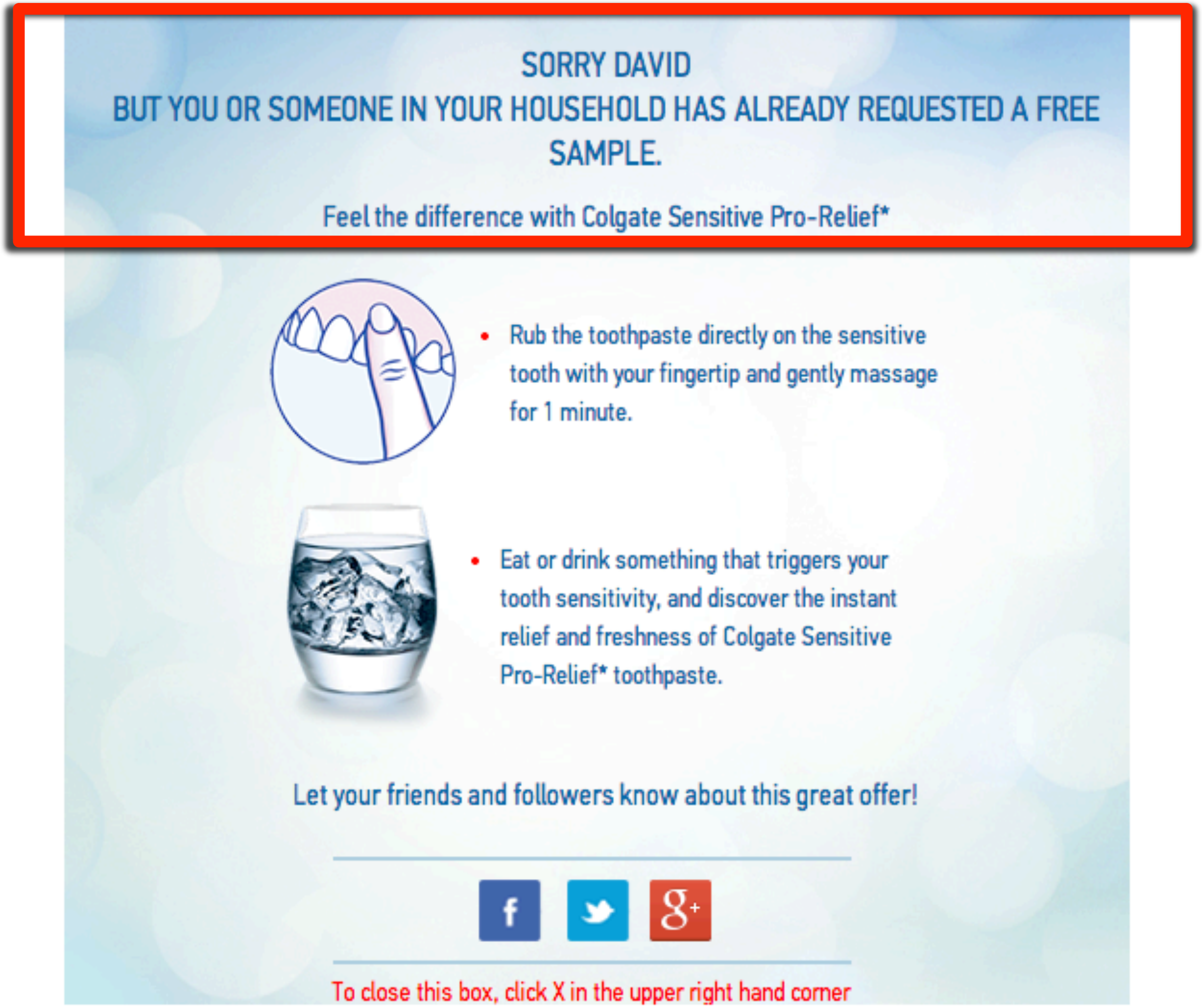
Province*
Ontario

Postal Code*
N2L 6E3

Email Address (Optional)

Confirm Email Address

Example – Real World Example



SORRY DAVID
BUT YOU OR SOMEONE IN YOUR HOUSEHOLD HAS ALREADY REQUESTED A FREE SAMPLE.

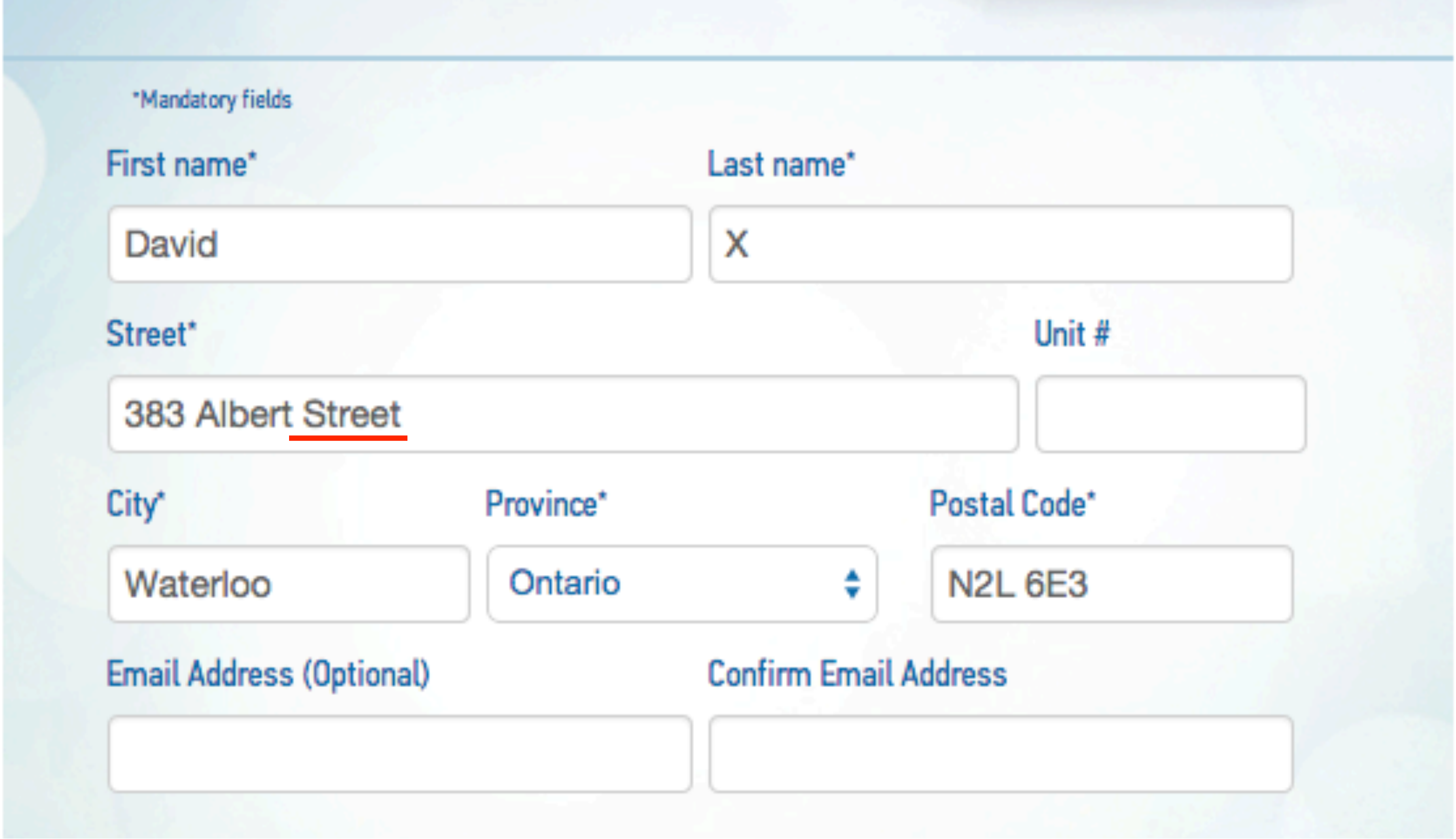
Feel the difference with Colgate Sensitive Pro-Relief*

- Rub the toothpaste directly on the sensitive tooth with your fingertip and gently massage for 1 minute.
- Eat or drink something that triggers your tooth sensitivity, and discover the instant relief and freshness of Colgate Sensitive Pro-Relief* toothpaste.

Let your friends and followers know about this great offer!

To close this box, click X in the upper right hand corner

Example – Real World Example



*Mandatory fields

First name*: David

Last name*: X

Street*: 383 Albert Street

Unit #:

City*: Waterloo

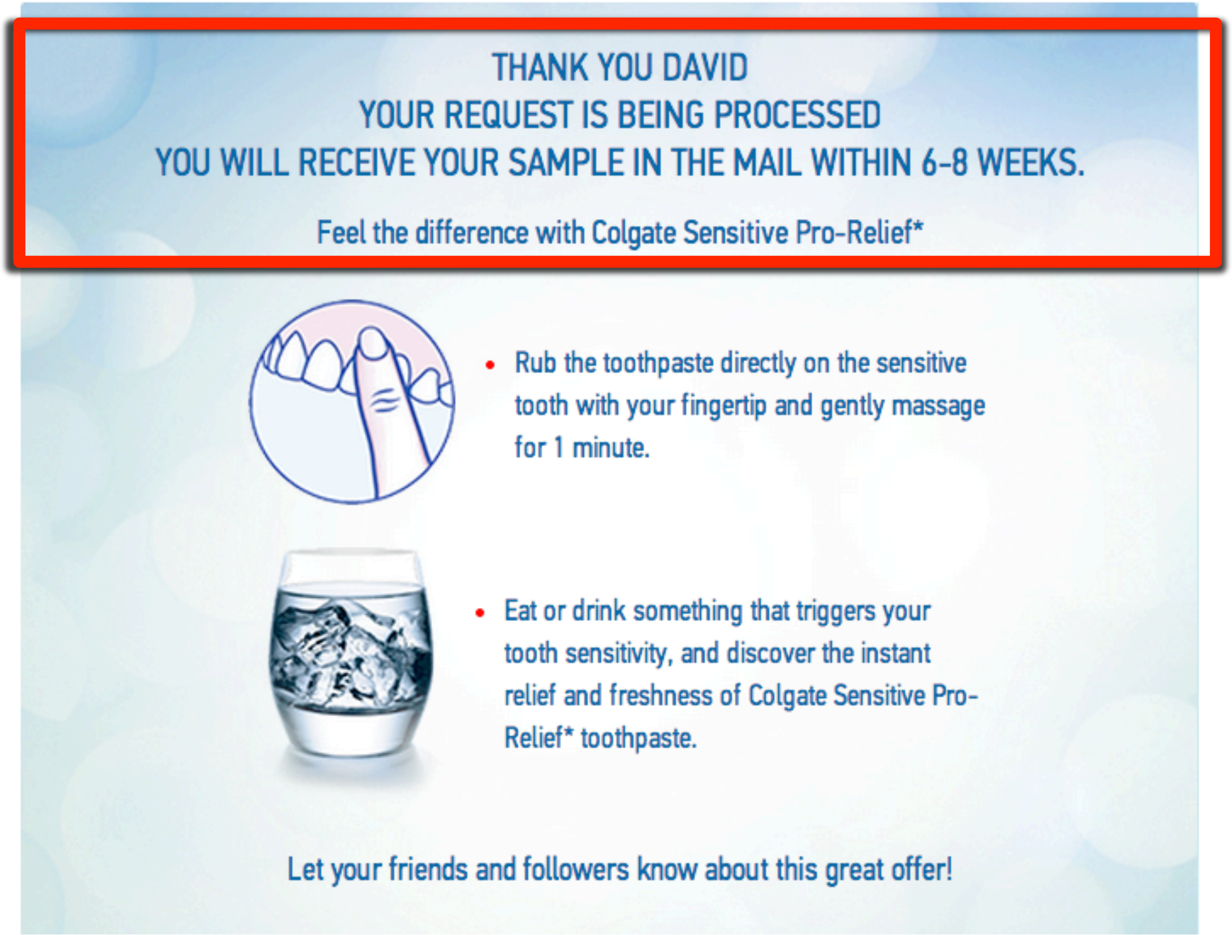
Province*: Ontario

Postal Code*: N2L 6E3

Email Address (Optional):

Confirm Email Address:

Example – Real World Example



THANK YOU DAVID
YOUR REQUEST IS BEING PROCESSED
YOU WILL RECEIVE YOUR SAMPLE IN THE MAIL WITHIN 6-8 WEEKS.

Feel the difference with Colgate Sensitive Pro-Relief*

- Rub the toothpaste directly on the sensitive tooth with your fingertip and gently massage for 1 minute.

- Eat or drink something that triggers your tooth sensitivity, and discover the instant relief and freshness of Colgate Sensitive Pro-Relief* toothpaste.

Let your friends and followers know about this great offer!

Example – Media Dataset

ID	ArtistName	TrackName
1	The Doors	LA Woman
2	Doors	LA Woman
3	The Beatles	A Little Help from My Friends
4	Beatles, The	With a Little Help From My Friend
…	…	…
7	4th Elemynt	Ears/Eyes
8	4th Elemynt	Ears/Eyes – Part II
9	4th Elemynt	Ears/Eyes – Part III
10	4th Elemynt	Ears/Eyes – Part IV
11	Aaliyah	Are You Ready
12	AC DC	Are You Ready

M. Bilenko. RIDDLE: Repository of information on duplicate detection, record linkage, and identity uncertainty. http://www.cs.utexas.edu/users/ml/riddle/index.html

Example – Media Dataset

ID	ArtistName	TrackName	
1	The Doors	LA Woman	Duplicates
2	Doors	LA Woman	
3	The Beatles	A Little Help from My Friends	Duplicates
4	Beatles, The	With a Little Help From My Friend	
…	…	…	
7	4th Elemynt	Ears/Eyes	Not Duplicates
8	4th Elemynt	Ears/Eyes – Part II	
9	4th Elemynt	Ears/Eyes – Part III	
10	4th Elemynt	Ears/Eyes – Part IV	
11	Aaliyah	Are You Ready	Not Duplicates
12	AC DC	Are You Ready	

M. Bilenko. RIDDLE: Repository of information on duplicate detection, record linkage, and identity uncertainty. http://www.cs.utexas.edu/users/ml/riddle/index.html

Introduction - Motives

- Customer Data
  - Prevent unnecessary costs in promotional material
- Company Data
  - Incorrect data analysis, such as counts on product

Machine Learning - Overview

- Leverage a branch of AI, called Machine Learning, to eliminate duplicates

- Use data to train algorithms into performing a task

- Run the algorithms on databases to clean the data

Machine Learning - Overview

1) Supervised Learning

2) Unsupervised Learning

Machine Learning - Supervised

1) Supervised Learning
   - Uses well defined training data to teach algorithm
   - May be difficult to obtain training data
   - Needs “domain knowledge”

Machine Learning - Unsupervised

2) Unsupervised Learning

- Relies on distance function detect duplicates
- Involves clustering of data

Duplicate Elimination Strategy

- Use edit distance to detect fuzzy duplicates

- **Edit distance:** Quantify similarity between strings, based on:
  - Insertion
  - Deletion
  - Substitution

  - E.g. Yellow -> Jello is 1 substitution and 1 deletion

- Can assign a distance metric between tuples

Edit Distance: https://web.stanford.edu/class/cs124/lec/med.pdf

Duplicate Elimination Strategy

- Baseline: “Global Threshold” to eliminate duplicates
- E.G. tuples are duplicates if: # of changes < X

Example – Media Dataset

ID	ArtistName	TrackName	
1	The Doors	LA Woman	Duplicates
2	Doors	LA Woman	
3	The Beatles	A Little Help from My Friends	Duplicates
4	Beatles, The	With a Little Help From My Friend	
…	…	…	
7	4th Elemynt	Ears/Eyes	Not Duplicates
8	4th Elemynt	Ears/Eyes – Part II	
9	4th Elemynt	Ears/Eyes – Part III	
10	4th Elemynt	Ears/Eyes – Part IV	
11	Aaliyah	Are You Ready	Not Duplicates
12	AC DC	Are You Ready	

M. Bilenko. RIDDLE: Repository of information on duplicate detection, record linkage, and identity uncertainty. http://www.cs.utexas.edu/users/ml/riddle/index.html

Duplicate Elimination Strategy

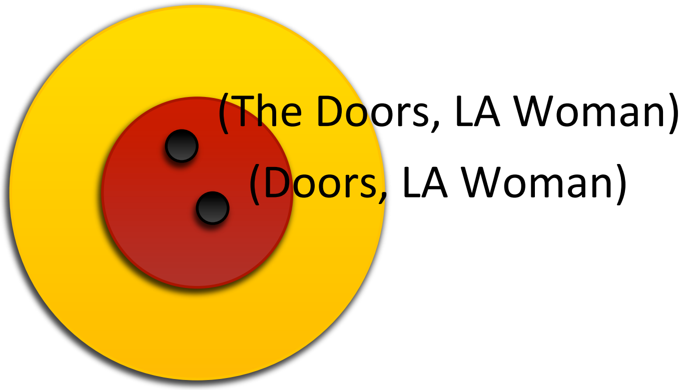
Fuzzy Duplicates are:

1) Duplicate tuples are ‘closer’ to each other than to others
    - A “compact set” (CS criteria)

2) The local neighborhood of duplicate tuples is sparse
    - A “sparse neighborhood” (SN criteria)

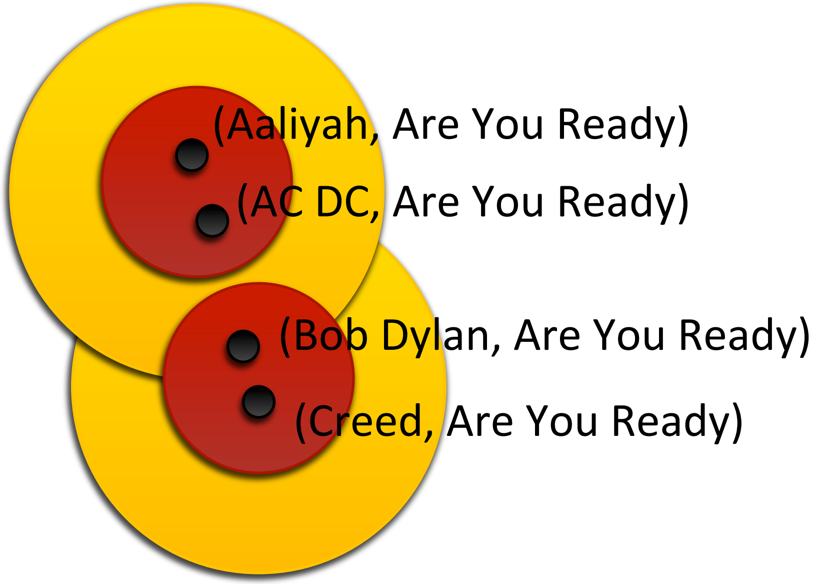
Duplicate Elimination Strategy

Red = Compact Set Criteria
Yellow = Sparse Neighborhood Criteria



(The Doors, LA Woman)

(Doors, LA Woman)



(Aaliyah, Are You Ready)

(AC DC, Are You Ready)

(Bob Dylan, Are You Ready)

(Creed, Are You Ready)

DE Problem

Formal Definitions

CS Criteria:

- Given a set S of tuples from relation R
- Each tuple in S, called v, is closer to tuples v', in S, than any other tuples v'' in R-S

SN Criteria:

- Neighborhood:
    - sphere of radius 2nn(v), (2x distance of closest neighbor)
- Sparse Neighborhood:
    - if # of tuples in Neighborhood < c

DE Problem

Partition R into a minimum number of groups $\{G_1,..,G_m\}$ for all $\mathbf{G_i}$ so:

1) $\mathbf{G_i}$ is a compact set
2) $\mathbf{G_i}$ is a sparse neighborhood
3) The size of $\mathbf{G_i} \leq$ K

OR

The diameter of $\mathbf{G_i} \leq$ Theta

c: positive threshold value
K: positive integer
Theta: positive real number

DE Algorithm

Sample implementation:

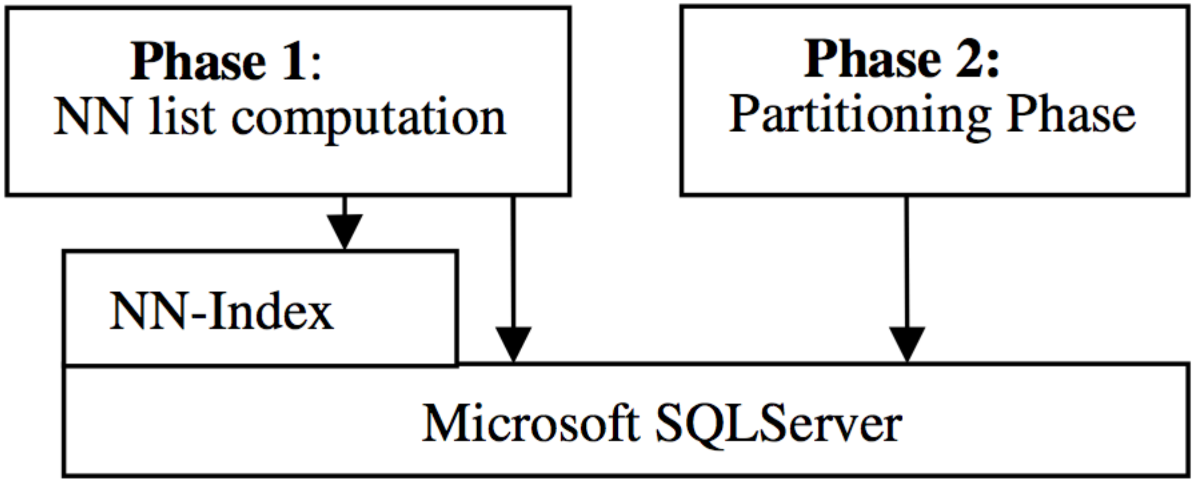


Figure 3: Architecture

Figure from: S. Chaudhuri, V. Ganti, and R. Motwani. Robust Identification of Fuzzy Duplicates . In Proc. Int'l Conf. on Data Engineering (ICDE'05), 2005, pp. 865-876.

DE Algorithm

Phase 1:

- Find the nearest neighbors for each tuple
  - the K nearest

    OR
  - within certain radius, Theta

- **Paper assumes a database indexed for distance** between neighbors
  - Index based on Exact Distance is very difficult
  - Index using an approximate / probabilistic method

DE Algorithm

Phase 2:

- Partition input relation into minimum number of compact SN sets

- The resulting partitions are the fuzzy duplicates

- Solution is unique based on parameters:
  - c threshold
  - K value or Theta distance

DE Algorithm - Impact on Database

Phase 1 – NN List Computation:

- Database needs to be indexed in a certain way

Phase 2 - Partitioning Phase:

- Most processing is done using SQL queries
- Avoids moving large amounts of data between client & server

Evaluation

RIDDLE Repository:

Internal Datasets:

- Media[artistName, trackName]
- Org[name, address, city, state, zipcode]

Public Datasets:

- Restaurants[Name]
- BirdScott[Name]
- Census[LastName, First name, Middle initial, Number, Street]

M. Bilenko. RIDDLE: Repository of information on duplicate detection, record linkage, and identity uncertainty. http://www.cs.utexas.edu/users/ml/riddle/index.html

Evaluation

1) Recall

- **“Fraction of true pairs of duplicates identified by an algorithm”**
- How many fuzzy duplicates can be identified?
- Higher the better

2) Precision

- **“Fraction of tuple pairs an algorithm returns which are truly duplicates”**
- How many of the duplicates tagged, are fuzzy duplicates?
- Higher the better

Results

Performs somewhat better than baseline

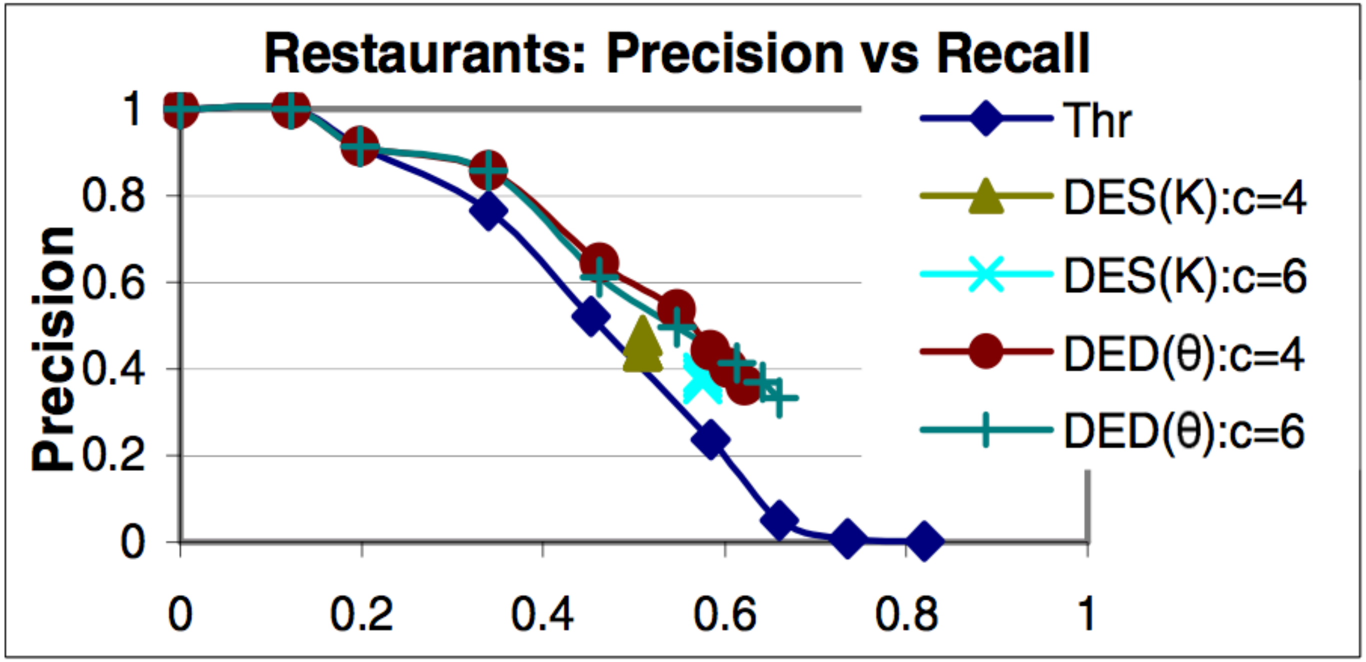


Figure from: S. Chaudhuri, V. Ganti, and R. Motwani. Robust Identification of Fuzzy Duplicates . In Proc. Int'l Conf. on Data Engineering (ICDE'05), 2005, pp. 865-876.

Results

Performs the same as baseline

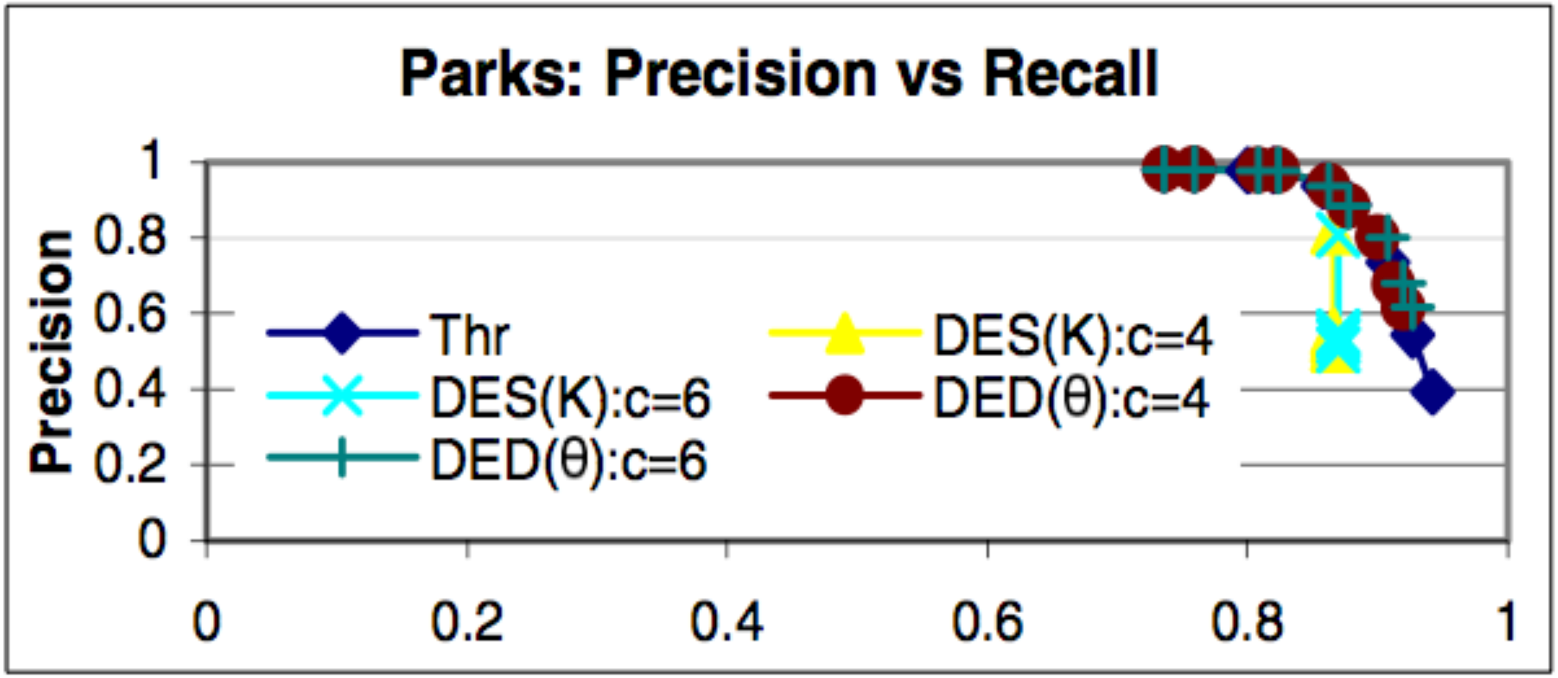


Figure from: S. Chaudhuri, V. Ganti, and R. Motwani. Robust Identification of Fuzzy Duplicates . In Proc. Int'l Conf. on Data Engineering (ICDE'05), 2005, pp. 865-876.

Results

Performs much better than baseline

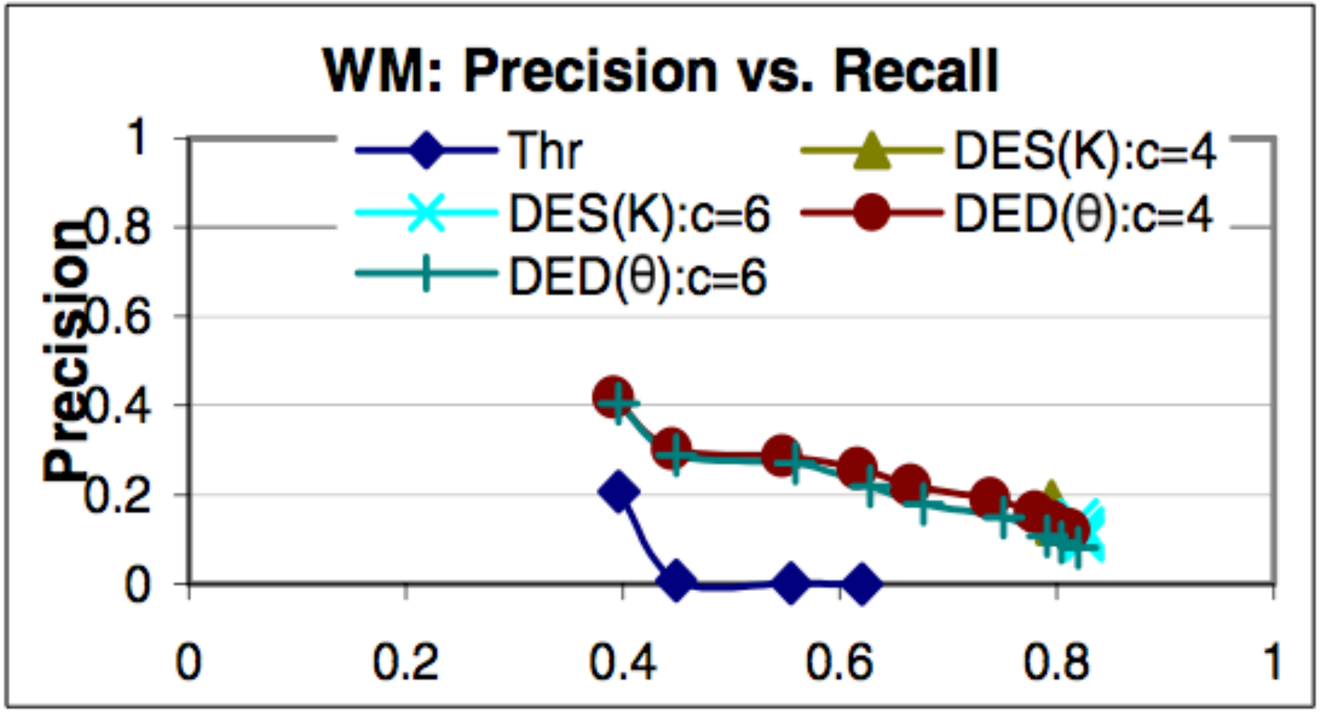


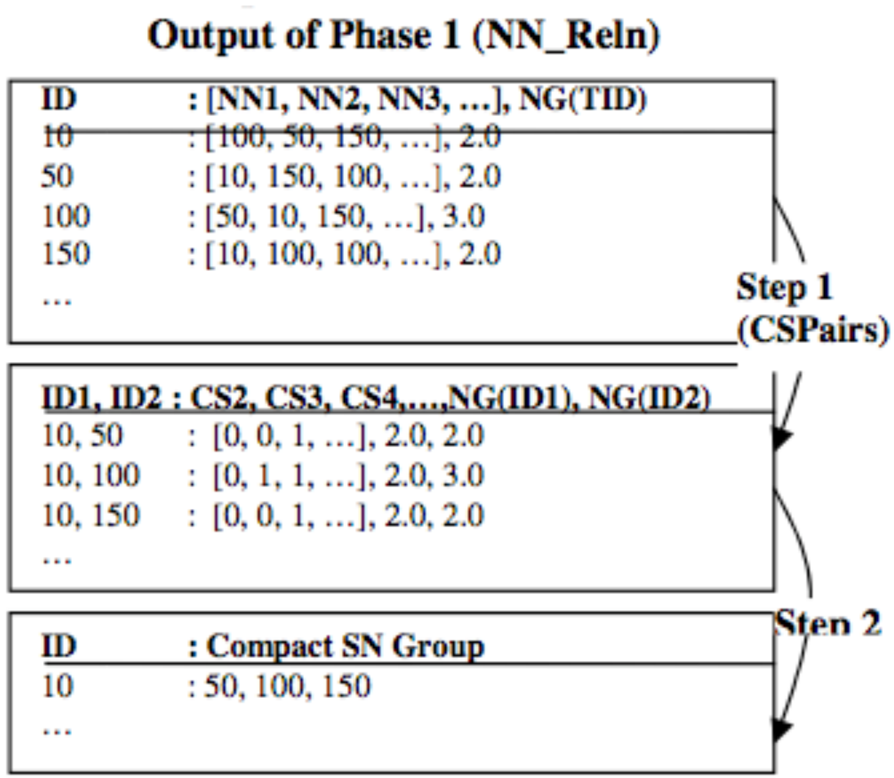
Figure from: S. Chaudhuri, V. Ganti, and R. Motwani. Robust Identification of Fuzzy Duplicates . In Proc. Int'l Conf. on Data Engineering (ICDE'05), 2005, pp. 865-876.

Thanks

Thanks for Listening!

Appendix

Set = {10 50 100 150}



Output of Phase 1 (NN_Reln)

ID	: [NN1, NN2, NN3, …], NG(TID)
10	: [100, 50, 150, …], 2.0
50	: [10, 150, 100, …], 2.0
100	: [50, 10, 150, …], 3.0
150	: [10, 100, 100, …], 2.0
…	

Step 1 (CSPairs)

ID1, ID2	: CS2, CS3, CS4,…,NG(ID1), NG(ID2)
10, 50	: [0, 0, 1, …], 2.0, 2.0
10, 100	: [0, 1, 1, …], 2.0, 3.0
10, 150	: [0, 0, 1, …], 2.0, 2.0
…	

Step 2

ID	: Compact SN Group
10	: 50, 100, 150
…	

Figure 6: Example illustrating the partitioning phase