ICS 624 Spring 2011

Entity Resolution with Evolving Rules

Preface to Steven Whang's slides

Asst. Prof. Lipyeow Lim

Information & Computer Science Department

University of Hawaii at Manoa

Facebook

Sharon Adler

Born on July 5, 1942

Activities and Interests

Other The Peninsula Bangkok, XML Prague, Barack Obama, The GEICO Gecko, Lori's Jewels, Wisconsin Colleges, Samantha's Skin Spa, I'm not yelling....I'm Jewish....That's how we talk...., Added 4.6 billion USD to the Veterans Administration budget to recruit and retain more mental health professionals

Basic Information

Sex Female

Contact Information

Email scaenator@gmail.com

LinkedIn

Sharon Adler (2nd)

Senior Manager at IBM Research

Providence, Rhode Island Area | Computer Software

Current	• **RSM Emeritus - retired** at **IBM Research**
Past	• Chair - XSLT WG at W3C • Chair - XSLT WG at W3C • Mother at XML & SGML Community see all...
Connections	**126** connections
Public Profile	http://www.linkedin.com/pub/sharon-adler/11/705/b74

Are these two pages referring to the same person ?

Entity Resolution (ER)

- comparison shopping
- mailing lists
- classified ads
- customer files
- counter-terrorism

r1

Name	Address	Credit Card	Phone
Sharon	RI	123122	303-123-9989

r2

Name	Affiliation	Phone
Sharon	IBM	303-123-9989

Are these two records referring to the same entity ?

Entity Resolution Problem Statement

- Given a table of records (about some entities), partition the records according to the "entities" that they refer to.

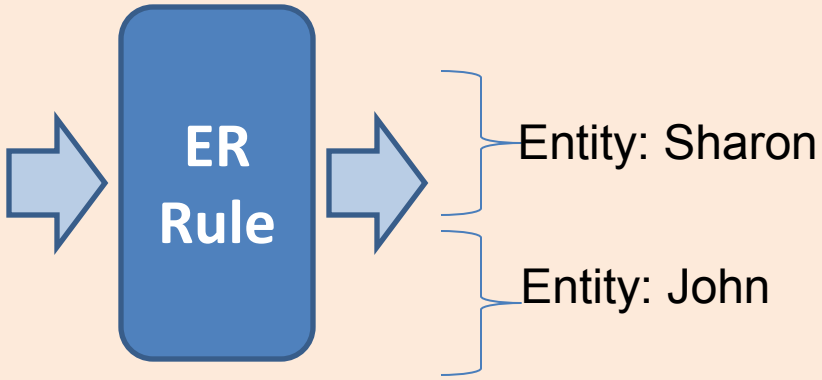
Name	Address	Credit Card	Phone
Sharon	RI	123122	303-123-9989
Sharon	NY	122223	303-123-9989
John	NY	333222	212-222-4433
John	NJ	222333	212-222-4433

Entity: Sharon

Entity: John

ER Rules

Name	Address	Credit Card	Phone
Sharon	RI	123122	303-123-9989
Sharon	NY	122223	303-123-9989
John	NY	333222	212-222-4433
John	NJ	222333	212-222-4433



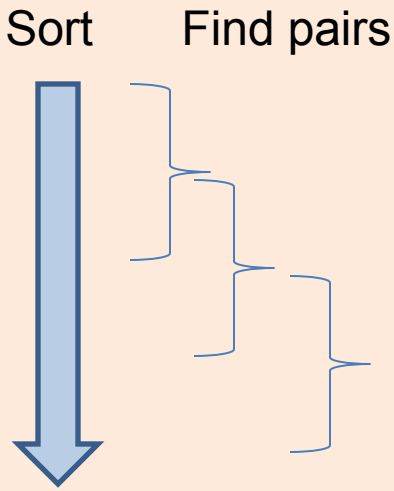
- ER Rules: ER algorithm that computes the mapping of records to entities ("partition")
- Match-based: boolean rules like
  - "if the name in the two records are the same, then they belong to the same partition"
- Distance-based: uses a distance function

Sorted Neighborhood (SN)

Avoids comparing all $O(n^2)$ pairs of records by:

- Sorting records based on some column(s)
- Comparing all pairs of records in a sliding window
- Merging connected components into entities

Name	Address	Credit Card	Phone
Sharon	RI	123122	303-123-9989
Sharon	NY	122223	303-123-9989
John	NY	333222	212-222-4433
John	NJ	222333	212-222-4433



HC_B: Hierarchical Clustering Boolean

- Similar to bottom-up hierarchical agglomerative clustering
- Merge two clusters if a boolean comparison rule **B** returns true.
- Apply rule **B** on *one chosen* tuple in each of the two clusters

Name	Address	Credit Card	Phone
Sharon	RI	123122	303-123-9989
Sharon	NY	122223	303-123-9989
John	NY	333222	212-222-4433
John	NJ	222333	212-222-4433

B(r1,r2) = true if r1.name=r2.name

HC_{BR}: Hierarchical Clustering Boolean

- Same as HC_B except in how comparison is evaluated.
- Apply rule **B** on *all pairs of* tuples in each of the two clusters
- Merge clusters if **B** is true on at least one pair

Name	Address	Credit Card	Phone
Sharon	RI	123122	303-123-9989
Sharon	NY	122223	303-123-9989
John	NY	333222	212-222-4433
John	NJ	222333	212-222-4433

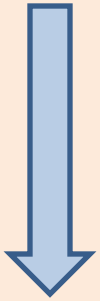
B(r1,r2) = true if r1.name=r2.name

ME: Monge-Elkan Clustering

- Sort records according to some column(s)
- Initialize an empty fixed length queue of clusters
- Scan through sorted records and match each record to clusters in queue
- If record matches existing cluster, move cluster to front
- Else make record into a new cluster at front of queue
- If queue is full, last cluster is dropped

Name	Address	Credit Card	Phone
Sharon	RI	123122	303-123-9989
Sharon	NY	122223	303-123-9989
John	NY	333222	212-222-4433
John	NJ	222333	212-222-4433

Sort

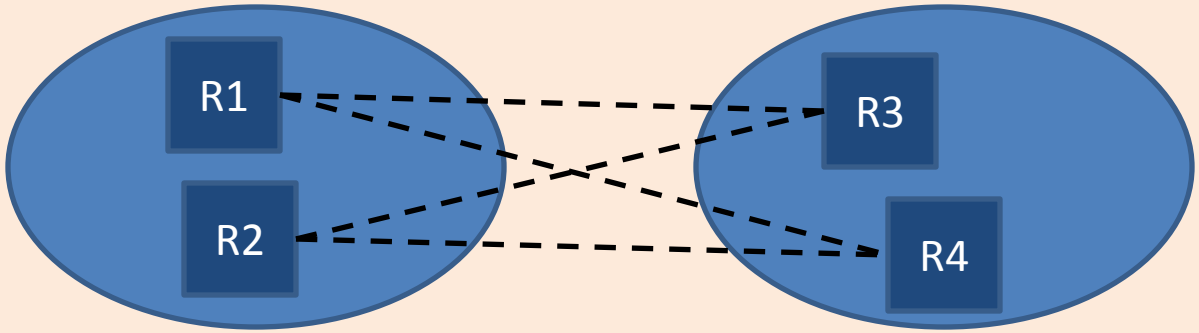


Queue

Sharon

Distance-based ER Algorithms

- Similar to bottom-up hierarchical agglomerative clustering with different variations on how distance is computed from two clusters
- **HC_{DS} Single-link** : smallest possible distance between two records from the two clusters
- **HC_{DC} Complete-link** : largest possible distance between two records from the two clusters



Evolving Rules