Business School
UNIVERSITY OF COLORADO DENVER

Information Systems Program

Module 4
Data Integration Concepts, Processes, and Techniques

Lesson 6: Quasi Identifiers and Distance Functions for Entity Matching

Lesson Objectives

- Discuss usage of quasi identifiers
- Explain features of distance functions used in entity matching
- Evaluate simple examples of edit distance
- Reflect on relationship between quasi identifiers and distance functions

Quasi Identifiers

- Used in entity matching
- Almost unique in combination
- Examples
  - Name components
  - Location components
  - Profession
  - Birthdate
  - Race

Distance Functions

- Determine amount of space between records or values
  - Determine distance between combination of quasi identifier values
  - Determine distance between two quasi identifier values
- Text distance
  - Important for quasi identifiers containing text
  - Many applications besides entity matching

Edit Distance

- Common distance function for text
- Operations to transform two text values
  - Delete a character
  - Insert a character
  - Substitute one character for another
- Minimal number of operations

Edit Distance Example

Saturday → Sunday

1. Sturday (delete “a”)
2. Surday (delete “t”)
3. Sunday (substitute “n” for “r”)

1. Suturday (substitute “u” for “a”)
2. Sunurday (substitute “n” for “t”)
3. Sunrday (delete “u”)
4. Sunday (delete “r”)

Phonetic Distance Functions

- Many applications in law enforcement
- Codes words into standard consonant sounds
- Widely available in DBMSs and data integration tools
  - Soundex: 6 consonant sounds
  - Metaphone: 16 consonant sounds

Phonetic Matching Examples

- Soundex
  - Soundex(Assistance) = A223
  - Soundex(Assistants) = A223
- Metaphone
  - Metaphone(Assistance) = ASSTNS
  - Metaphone(Assistants) = ASSTNTS

Summary

- Quasi identifiers for entity matching
- Use inexact matching on quasi identifiers with text
- Edit and phonetic distance functions
- Text distance functions provided by tools for data mining and data integration tools as well as DBMSs