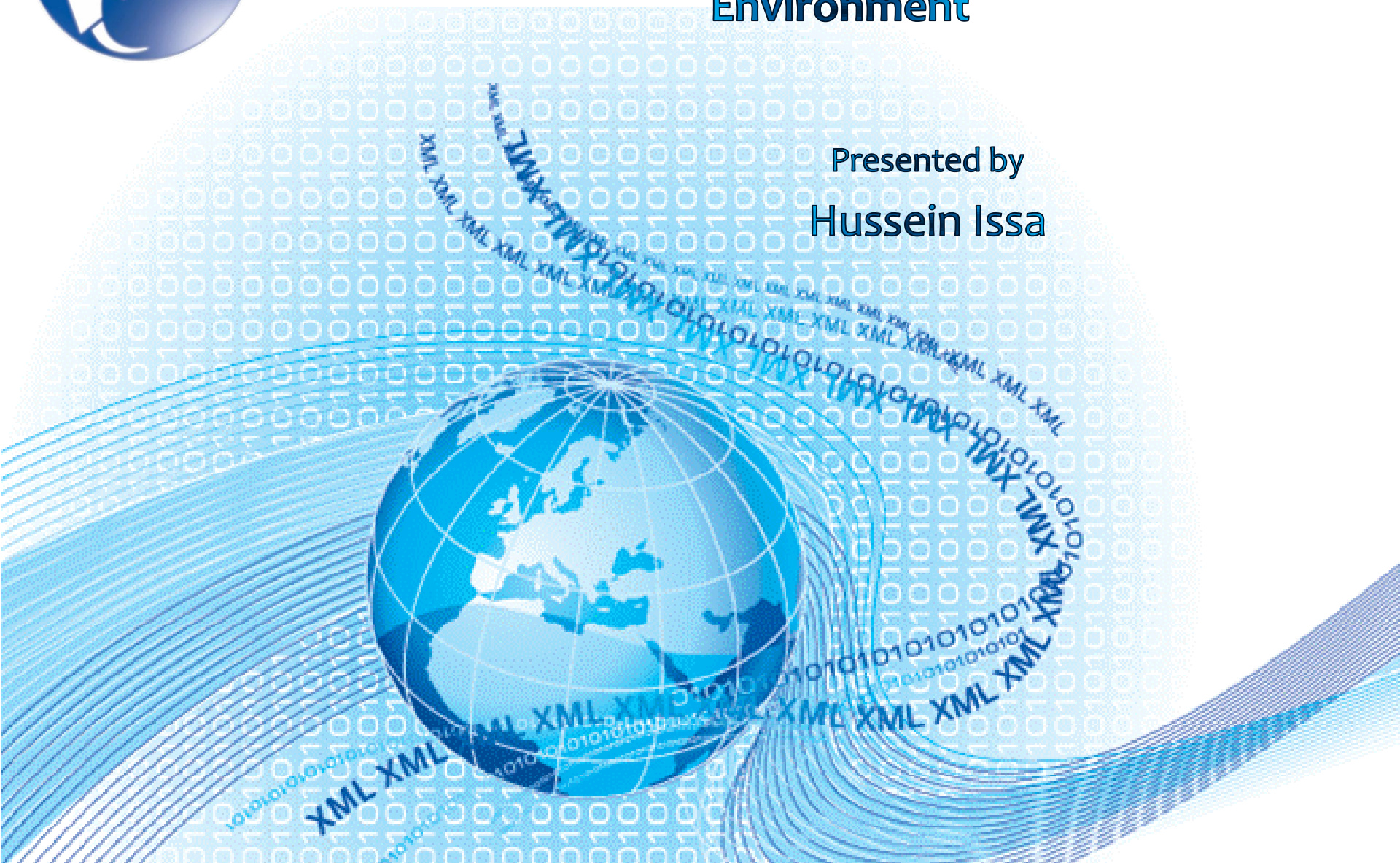
Application of Duplicate Records Detection Techniques to Duplicate Payments in a Real Business Environment

Presented by

Hussein Issa



Outline

- Introduction
- Research Question and Contribution
- Duplicate Records
- Duplicate Payments
- Methodology
- Data Set
- Findings

Introduction

- Data is the cornerstone and base of business operations
- Data size increased mainly due to electronization and Globalization of businesses
- Data is gathered from different sources, which gave rise to new problems that affected the quality and accuracy of this data
- One critical problem is **duplicate records** where a real object has multiple representations in a database
  - Example Duplicate payments

Research Question and Contribution

- **Research Question:**

Compare different algorithms and techniques used in the Computer Science literature, apply them to the problem of duplicate payments, and determine the optimal method.

- **Paper's Contribution**

The literature in Computer Science field is abundant with studies that address duplicate records detection. However, in the accounting literature, only few academic papers targeting this issue exist. This paper contributes by adapting a technique from the Computer Science literature and applying it to an accounting problem.

Duplicate Records

Problem History

- Duplicate records detection was known before as record matching or record linkage (Tepping 1968,Newcombe 1988)
- Used as a fraud detection tool (Cottini et.al, 1995)
- Most proposed algorithms in the literature are domain-specific

Duplicate Records

Duplicate Detection Methods

Generalized framework described by Weis and Neumann (2005):

- Phase 1: Candidate definition
  - Decide which objects to be compared
  - E.g. No need to compare name and address
- Phase 2: Duplicate definition
  - Criteria for 2 candidate duplicates to be considered actual duplicates
  - E.g. With duplicate payments, the variable AMOUNT is important, size of client company is not
- Phase 3: Actual duplicate detection
  - Specifying how to detect duplicates candidates and find which ones are true duplicates.

Duplicate Records

Duplicate Detection Methods – Continued

Two general methods are used:

1. Exact matching:
   - 2 records in a dataset are *identical* in all fields

2. Fuzzy (near-identical) matching (Weis et.al., 2008):
   - 2 records have *similar* values for certain relevant fields, and they are considered duplicates based on similarity criteria and threshold
   - May happen due to data entry mistakes, or different value formats. E.g. Address format, date format

Duplicate Payments

Problem Description

- Electronization ➔ large datasets ➔ development of systems ➔ ERPs, DSSs, KESs, automated audit tools (Chou et.al., 2007)

- Output quality of these systems depends on the quality of input data

- Duplicate payments can result from human errors, object presentation (Checks paid to ***Rutgers*** vs. ***Rutgers University***), systematic errors (*date format*, *name format*), or **fraud**

Duplicate Payments

Problem Implications

- Frequent problem
  - Medicaid: $9.7 million detected, real amount estimated to be around $31.1 million (Novello report, 2004)
  - Veterans Affairs: 2 duplicate payments = $1.08 million (General Inspector report, 1997)

- Some commercial agencies specialize in detecting duplicates and charge 15%-50% of amounts detected

Methodology

Methodology

Most studies follow the generalized framework described earlier, and the majority use a 3-way (*AMOUNT, DATE, VENDOR*) or a 4-way match (*AMOUNT, DATE, VENDOR, INZVOICE #*).

Data Description

- 2 files: (July 2008 – June 2010)
- **Dataset 1:** includes information on payments to telecom carriers and has 21606 records and 8 variables
- **Dataset 2:** includes information on checks payments and has 47683 records and 51 variables
- Software used: ACL

Data Set

Data Validation

- **Method followed:** described by Kimball and Caserta (2009):
  - ***Parsing, Data transformation, and Data standardization***

- **Dataset 1:**
  - No transformation was needed
  - No missing values

- **Dataset 2:**
  - Many irrelevant attributes and filters were applied based on company feedback to fix this problem
  - Many missing values (fixed after applying filters)
  - Some attributes needed standardization (amounts format)

Findings

Carriers Dataset

- *(Carrier ID) + Effective Date + Amount gave* 82 candidate duplicates (at least one is confirmed to be duplicate)
- *(Carrier ID) + Entered Date + Amount gave* 168 candidate duplicates

Oracle Fin Dataset

- 4-way match (DATE, AMOUNT, VENDOR, INVOICE ID) gave 33 records
- Interesting finding: Refunds (>12000 records)

Thank You!