Module L12
Data Integration Part 1

Presented by: John R. Talburt
Acxiom Chair of Information Quality, UALR
Chief Science Officer, Black Oak Analytics, Inc.
September 28, 2016
Overview

• Module L12 Objective
  • Provide a broad overview of the role and challenges around data integration with special attention to entity-based data integration (EBDI)

• Topics
  • What is data integration
  • What are data integration “keys” – source key, hash key, maintained keys
  • What is entity-based data integration (EBDI)
  • What is entity resolution (ER)
  • How is entity ER evaluated
  • Probabilistic matching versus deterministic matching
Why This Is Important

Relationship to Information Quality and Data Governance
Multiple Sources of Information

• “multiple sources of the same information” is #1 in list of top 10 root causes of DQ problems [Journey to Data Quality, Lee, et al]
Data Integration

• “Using data from more than one source in a single application”
• In almost all cases, data from different records are combined by agreement on a common value – the data integration “key”
• Types of data integration keys
  • Intrinsic key – a common value found in each source, e.g. “Twitter handle”
  • Hash key (composite key, intrinsic key) – a value made by combining several values in each source, e.g., “John”, “Talburt” → “JTALBU”
  • Maintained (managed) key – an external value appended to records by a separate system that determines which records should be joined, e.g. “patient master index”
Maintained Keys for Critical Information

• Intrinsic keys are not always present
• Hash keys may not produce the required accuracy of integration
• Systems that automate the management of entity identity information are often the best solution
• These are systems called
  Entity Identity Information Management (EIIM) Systems
Which belong together?

J. Talburt
UALR T00020975

J Talburt
White Dr, Batesville, AR

John Talburt
700 Pierce, LR, AR

John Talburt
Hazelton Rd, Pea Ridge, AR

J R Talburt
Old Country Ln, Cabot, AR

John Talburt
18 Chaparral Ln, LR, AR
xxx-xx-1519

Jerry Talburt

DOB 3/1961

(c) 2016 iCDO@UALR
Entity References

• An entity reference is a collection of entity attribute-value pairs for a specific entity
  • (name) John Doe, (dept) Trust Dept, (org) ABC Bank
  • (name) Marriott Hotel, (city) Cambridge, (state) MA

• References to the same entity are said to be equivalent references
  • John Doe, 123 Main St, Anyville, AR
  • J Doe, 9/12/1965, ABC Co., Anyville, AR
Entity-Based Data Integration (EBDI)

• The creation of almost every information product (IP) includes a step for entity-based data integration (EBDI)
  • Entities are real-world objects (customers, products, places)
  • First step is to decide which records are for the same entity
• Only after that can you decide how to integrate and reconcile the remaining information
Three Related Concepts

- Entity resolution (ER)
- Entity identity information management (EIIM)
- Master data management (MDM)
Entity Resolution (ER)
Entity Resolution (ER) is the process of determining whether two references in an information system are referring to the same real-world object or to different objects (Talburt, 2011).

If they refer to same real-world object, they are said to be “Equivalent”.

- Record-linking
- Record-deduplication
- Data matching
- Co-reference problem
- Semantic resolution
ER is widely used in many areas

- Homeland Security
- Banking and Finance
- Healthcare
- Customer Data Management
Equivalence versus Matching

• Not all equivalent references match
  • “Mary Doe, 123 Oak” and “Mary Smith, 345 Elm” could be the same customer, married and moved

• Not all matching references are equivalent
  • “Jim Jones, 67 Pine” and “James Jones, 67 Pine” could be different customers, father and son

• Entity resolution is about equivalence, not matching
Linking References

• An ER system shows its decision that two references are equivalent by giving them both the same identifier value (link identifier)

• This is called linking

• ER errors
  • False Positive: Linking two references that are not equivalent
  • False Negative: Not linking two references that are equivalent
Measures of Linking Outcomes

- False Negative Rate = FN/(TP+FN)
- False Positive Rate = FP/(TN+FP)
- Precision (P) = TP/(TP+FP)
- Recall (R) = TP/(TP+FN)
- Accuracy = (TP+TN)/(TP+FP+TN+FN)

D = All pairs
L = Linked pairs
E = Equivalent pairs
Truth Set Evaluation

• Build truth set (golden records, certified matches)

• Same techniques but one-time effort
  • Randomly select seed records from set
  • Try to find all true positive matches
    • Wildcard matches, random selections
  • Give cluster a separate set of “true” links
  • Apply ER process, then compare differences between ER links and true links
Sampling ER Results

• False positives can be found by inspecting clusters of linked records
  • Large Clusters
  • High-Entropy Clusters

• False negatives are harder, use combination of
  • Wild card search
  • Random selection
  • Compound key splits of clusters
Benchmarking

• Comparing current outcomes to results believed to represent accurate results using the same data input

• Useful for
  • Comparing rule changes for the same system during rule development
  • Comparing results from two different systems
Talburt-Wang Index

\[ TWi = \frac{\sqrt{|A| \cdot |B|}}{|V|} \]

Where

• \(|A|\) is the number of cluster created by ER Process 1
• \(|B|\) is the number of cluster created by ER Process 2
• \(|V|\) is the number of overlaps between Process Clusters

• Will always give a value between 0 and 1
• Only has value of 1 when clusters are identical
For One-Time Integration, ER is Enough

• Quick and dirty analytics
• Bringing records together for a one-time analysis
• No need to “remember” a particular entity for later processing,
  • Anonymous Entities
  • De-Identifier Results
Understand the Data

• Data assessment should be the first step in designing any information process

• Data profiling and “data gazing”*
  • Systematic
  • Random

*Data Quality Assessment, Maydanchik
Selecting Identity Attributes

• Primary identity attributes
  • Ability to discriminate (characteristic, persistent)
  • Higher levels of completeness

• Supporting identity attributes
  • Less unique, less complete, changes
  • Can confirm good match, prevent bad match

• Must work with what you have!
  Especially with multiple sources
Boolean or Scoring Rule

• Two most common types of reference-level matching rules
  • Boolean rules
    • Produce a True (Match) or False (No Match) decision
    • Often called “Deterministic Matching”
  • Scoring rules (Probabilistic)
    • Give a value that represents the relative distance (similarity) between two references
    • Often called “Probabilistic Matching”
Boolean Rule Structure (Deterministic)

To have matching records, all parts of at least one Rule must be satisfied, but satisfying one Rule is enough.
Boolean Rule Example

AND logic

<IdentityRules>
  <Rule Ident="1">
    <Term Item="FirstName" Similarity="SOUNDEX"/>
    <Term Item="LastName" Similarity ="Scan(LR,LETTER,0,ToUpper,SameOrder)"/>
    <Term Item="DOB" Similarity ="Scan(LR,DIGIT,0,ToUpper,SameOrder)"/>
  </Rule>
  <Rule Ident="2">
    <Term Item="FirstName" Similarity ="NICKNAME"/>
    <Term Item="LastName" Similarity ="Scan(LR,LETTER,0,ToUpper,SameOrder)"/>
    <Term Item="DOB" Similarity ="Scan(LR,DIGIT,0,ToUpper,SameOrder)"/>
  </Rule>
  <Rule Ident="3">
    <Term Item="FirstName" Similarity ="Scan(LR,LETTER,0,ToUpper,SameOrder)"/>
    <Term Item="LastName" Similarity ="Scan(LR,LETTER,0,ToUpper,SameOrder)"/>
    <Term Item="Phone" Similarity ="Scan(LR,DIGIT,0,ToUpper,SameOrder)"/>
  </Rule>
</IdentityRules>
Advantages of Boolean Rules (Deterministic Matching)

• Easily understood and formulated
• Changes in one rule do not directly impact other rules
• Can easily “cross-compare” attributes to detect “misfielding” (first-last name switch)
• Easier to align with the match-key index
Scoring Rule Structure (Probabilistic Matching)

<table>
<thead>
<tr>
<th>James</th>
<th>Doe</th>
<th>123</th>
<th>Roosevelt</th>
<th>St</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alias Match</td>
<td>Exact Match</td>
<td>Exact Match</td>
<td>Fuzzy Match</td>
<td>Exact Match</td>
</tr>
<tr>
<td>Jim</td>
<td>Dough</td>
<td>123</td>
<td>Roosevelt</td>
<td></td>
</tr>
</tbody>
</table>

Agree Disagree Agree Agree Missing

Agree Weight: 80 70 25 35 0
Disagree Weight: -10 -5 0 -3 0

Match Threshold: 150

Match: NO MATCH!

Sept 26-29, 2016
(c) 2016 iCDO@UALR
Scoring Rule Example

```xml
<ScoringRule Ident="Example1" MatchScore="800" ReviewScore="600">
  <Term Item="StudentFirst" Similarity="LED(0.80)"
    DataPrep="Scan(LR, Letter, 0, ToUpper, SameOrder)"
    AgreeWgt="300" WgtTable="Ex1SFirst" DisagreeWgt="-20" />
  <Term Item="StudentLast" Similarity="Exact"
    DataPrep="Soundex"
    AgreeWgt="400" WgtTable="Ex1SFirst" DisagreeWgt="-30" />
  <Term Item="StudentDOB" Similarity="Exact"
    DataPrep="Scan(LR, Number, 0, ToUpper, SameOrder)"
    AgreeWgt="350" DisagreeWgt="0" />
  <Term Item="Gender" Similarity="Exact"
    DataPrep="Scan(LR, Alpha, 0, ToUpper, SameOrder)"
    AgreeWgt="20" DisagreeWgt="-50" />
  ....
</ScoringRule>
```
Advantages of Scoring Rule (Probabilistic Matching)

• Provides a finer granularity of matching
• Weights can be assigned at the value level (frequency-based weights)
• Generally out-performs Boolean rules, but
• Difficult to generate and refine the weights
• 100% alignment with blocking may not be possible
Rule Precision and Recall

All record pairs

Pairs linked by the Rule

False positive errors of Rule 1

True Positives

FP

FN

Precision = \frac{TP}{TP + FP}

Recall = \frac{TP}{TP + FN}

False negative errors of Rule 1

All equivalent pairs

True Negatives

Sept 26-29, 2016

(c) 2016 iCDO@UALR
Most ER Systems are Designed for One-time Resolution

Same Clusters, Different Cluster Identifiers
However, Sometimes Identity is Important

• Entities are real-world objects that have distinct identities
• Entities have attributes – their values convey information about the entity
  • Person name, age, ...
  • Product model number, color, ...
• Identity attributes help distinguish one entity from another
Thank You!

• John R. Talburt
• jrtalburt@ualr.edu