Query-time Entity Resolution

Indrajit Bhattacharya
Lise Getoor

Presentation: Sean Goldberg
Overview

- Introduction
- Entity Resolution and Queries: Formulation
- Collective Entity Resolution and Relational Clustering
- Analysis of Collective ER Using Relational Clustering
- Collective Resolution for Queries
- Adaptive Query Expansion
- Experiments
- Conclusions
Introduction

- The Entity Resolution Problem
- ER Aliases
- Why is Entity Resolution Useful?
- Common Ways of Doing ER
  - Limitations
- Query Time Entity Resolution (QTER)
  - Basic Steps
The Entity Resolution Problem

Steven M. Smith

“Steve Smith”

“S. Smith”

Stephen Smith

“Stephen Smith”

Sam Smith

“Sam Smith”

Samuel Smith

“S.M. Smith”

“S. Smith”
ER Aliases

- Deduplication
- Fuzzy Match Problem
- Reference Reconciliation
- Entity Resolution
- Object Consolidation
- Record Linkage
Why is ER Useful?

- Resolution of Duplicates
- Consolidation
- Maximize Information Content
- Disambiguation
- Identification
Why is ER Useful?

- Animation
Common Approaches to ER

- Attribute Similarity (traditional)
  - Exact Match
  - String Edit Distance
  - Cosine Similarity
  - TF-IDF
  - Clustering or Feature Extraction
  - Secondary Source Information

- Relational Similarity
  - Affiliation or Co-Authorship
  - Collective Resolution
Limitations to Current Approaches

- Traditional approaches use only string matching on attributes
  - Fields can be written in many different ways
    - abbrev., spelling errors, nicknames
  - Some schemas may be completely different
  - Matching threshold difficult to determine
- Requires resolution of entire database
  - Computationally hard
  - Difficult to adjust to persistent data
Query Time Entity Resolution

- Allow user to query unresolved or partially resolved database
- Resolve only *relevant* entities pertaining to specific queries on-the-fly
Basic Steps

- Extract relevant records by a recursive expansion technique
- Select 'most informative' records using an adaptive algorithm
- Resolve selected records collectively using *relational clustering*
Overview

- Introduction
- **Entity Resolution and Queries: Formulation**
- Collective Entity Resolution and Relational Clustering
- Analysis of Collective ER Using Relational Clustering
- Collective Resolution for Queries
- Adaptive Query Expansion
- Experiments
- Conclusions
Entity Resolution and Queries: Formulation

- Definitions
- Running Example: Citation Matching
- ER Queries
Running Citation Example

A Mouse Immunity Model
- W Wang
- C Chen
- A Ansari
- r1
- r2
- r3

Measuring Protein-bound Fluoxetine
- L Li
- C Chen
- W Wang
- r6
- r7
- r8

A Better Mouse Immunity Model
- W Wang
- A Ansari
- r4
- r5

Autoimmunity in Biliary Cirrhosis
- W W Wang
- A Ansari
- r9
- r10
- r7
ER Queries

• Consider \( R = \{ r \} \) and \( A = \{ a \} \)
• Return entities based on attribute
  - \( Q(R.A=a) \)
  - Should only return unique entities
• Return records based on entity
  - \( Q(R.A=r1.A) \) such that \( E(R)=E(r1) \)
Overview

- Introduction
- Entity Resolution and Queries: Formulation
- Collective Entity Resolution and Relational Clustering
- Analysis of Collective ER Using Relational Clustering
- Collective Resolution for Queries
- Adaptive Query Expansion
- Experiments
- Conclusions
Collective Entity Resolution and Relational Clustering

- **ER Approaches**
  - Attribute-based Entity Resolution (A)
  - Naive Relational Entity Resolution (NR)
  - Collective Entity Resolution

- **Relational Clustering**
  - Attribute-based Similarity
  - Relational Similarity
  - Clustering Algorithm

- Issues with Collective Resolution for Queries
Attribute-based ER

- Attributes of records used for comparison
- Considered co-referent if similarity above a certain threshold $T$
- Similarity comparison methods are the same as those discussed earlier
Attribute-based ER

A Mouse Immunity Model
- W Wang
- C Chen
- A Ansari

A Better Mouse Immunity Model
- W Wang
- A Ansari

Measuring Protein-bound Fluoxetine
- L Li
- C Chen
- W Wang

Autoimmunity in Biliary Cirrhosis
- W Wang
- A Ansari
Naive Relational ER

- Considers attributes of related references
  - Related references are those connected by hyper-edges
- Generally leads to a strong increase in performance when combined with attribute-based ER
- Can still lead to resolution errors
Relational ER

A Mouse Immunity Model
- W Wang
- C Chen
- A Ansari

A Better Mouse Immunity Model
- W Wang
- A Ansari

Measuring Protein-bound Fluxetine
- L Li
- C Chen
- W Wang

Autoimmunity in Biliary Cirrhosis
- W W Wang
- A Ansari
Collective ER

- Takes into account *resolution decisions* for related references
- Leads to recursive procedure
- Computationally more challenging
Collective ER

A Mouse Immunity Model
W Wang
C Chen
r1
A Ansari
r2
r3

A Better Mouse Immunity Model
W Wang
A Ansari
r4
r5

Measuring Protein-bound Fluoxetine
L Li
C Chen
W Wang
r6
r7
r8

Autoimmunity in Biliary Cirrhosis
W W Wang
A Ansari
r9
r10
h1
h2
h3
h4
Relational Clustering

- Cluster references according to their entities
  - Assign cluster label r.C to record
  - Take relationships into account
- Initialize all references to their own cluster
- Iteratively merge similar clusters of references
Similarity Measure

- \( \text{sim}(c_i, c_j) = (1-a) \times \text{sim}_A(c_i, c_j) + a \times \text{sim}_R(c_i, c_j) \)
  - \( c_i \) - \( i \text{th} \) cluster
  - \( a \) - combination weight
  - \( \text{sim}_A \) - attribute similarity between clusters
  - \( \text{sim}_B \) - relational similarity between clusters
Attribute-based Similarity

- Assigns pairs of attributes a value between 0 and 1
- Total similarity between clusters is a weighted linear combination of similarities over different attributes
Relational Similarity

- Similarity of *cluster neighborhoods*
  - Neighborhoods defined by hyper-edges $h$
  - Hyper-edges connect clusters (entities) to other clusters
  - $c.H$ denotes hyper-edge set for each cluster
- Neighborhood of cluster $c$
  - $Nbr(c)$ defined by all records within all clusters connected by a hyper-edge
- Relational Similarity Measure
  - $\text{simR}(c_i, c_j) = \text{Jaccard}(Nbr(c_i), Nbr(c_j))$
Relational Similarity

- Example of clusters connected by hyper-edge
Clustering Algorithm

1. Find similar references in R using blocking
2. Initialize clusters using bootstrapping
3. For clusters \( ci, cj \) such that \( \text{similar}(ci, cj) \)
4. Insert \( \langle \text{sim}(ci, cj), ci, cj \rangle \) into priority queue
5. While priority queue not empty
6. Extract \( \langle \text{sim}(ci, cj), ci, cj \rangle \) from queue
7. If \( \text{sim}(ci, cj) \) less than threshold, then stop
8. Merge \( ci \) and \( cj \) to new cluster \( c_{ij} \)
9. Remove entries for \( ci \) and \( cj \) from queue
10. For each cluster \( ck \) such that \( \text{similar}(c_{ij}, ck) \)
11. Insert \( \langle \text{sim}(c_{ij}, ck), c_{ij}, ck \rangle \) into queue
12. For each cluster \( cn \) neighbor of \( c_{ij} \)
13. For \( ck \) such that \( \text{similar}(ck, cn) \)
14. Update \( \text{sim}(ck, cn) \) in queue
Collective ER for Queries

- Collective ER improves accuracy significantly for offline cleaning
- Attribute and Naive Relational can be applied at query-time easily
- Collective ER much more difficult
  - Need to identify all references relevant for answering a query
  - Need to perform resolution on these references
  - Unconstrained form leads to a deep recursion
Overview

- Introduction
- Entity Resolution and Queries: Formulation
- Collective Entity Resolution and Relational Clustering
- Analysis of Collective ER Using Relational Clustering
- Collective Resolution for Queries
- Adaptive Query Expansion
- Experiments
- Conclusions
Analysis of Collective ER Using Relational Clustering

- Recall and Precision Definitions
- Influence of Attributes
- Influence of Relationships
- Recall and Precision Equations
Definitions

- **Recall**
  - For entity $e_i$, how many pairs of references are correctly assigned to the same computed cluster?

- **Precision**
  - For cluster $c_i$, how many pairs of references assigned to $c_i$ correspond to the same underlying entity?
Influence of Attributes

- T-similar
  - Two references are T-similar if their similarity is above some threshold $T$
- Attribute-based approach assigns references to same cluster iff they are T-similar
Influence of Attributes

- Running Example - precision and recall
Influence of Attributes

- **Attribute Identification Probability**
  - $a_{I}(e,T)$ - probability that a pair of references chosen randomly from those corresponding to entity $e$ are T-similar to each other
  - Same as Recall $R(e,T)$

- **Attribute Ambiguity Probability**
  - $a_{A}(e1,e2,T)$ - probability that a pair of references chosen randomly such that one corresponds to entity $e1$ and another to $e2$ are T-similar to each other
  - Same as Imprecision $I(e1,e2,T)$
Influence of Attributes

- Show probabilities on running example
Influence of Relations

- Improves results that are not T-similar
- D-similar references for some $D < T$ considered as candidates for relational clustering
- Similarity increased if multiple hyper-edges connect the same clusters

- Insert figure 3 from paper
Influence of Relations

- Let \((r_i, r_j)\) co-occur through hyper-edge \(h\)
- Let \((r'_i, r'_j)\) co-occur through hyper-edge \(h'\)
- \((r_i, r'_i)\) are more similar if \((r_j, r_j')\) are clustered together
  - \(h\) and \(h'\) are an *identifying relationship* for entity \(e_i\)
- The clustering could still be incorrect
  - \(h\) and \(h'\) are an ambiguous relationship
Influence of Relations

- Identifying Relation graphic
- Ambiguous Relation graphic
Influence of Relations

- Identifying Relationship Probability
  - \( r_I(e, D) \) - probability that a randomly chosen pair of D-similar references corresponding to entity \( e \) has identifying relationships \( h \) and \( h' \) with some other entity

- Ambiguous Relationship Probability
  - \( r_A(e_1, e_2, D) \) - probability that a randomly chosen pair of D-similar references corresponding to entities \( e_1 \) and \( e_2 \) has ambiguous relationships \( h \) and \( h' \) with some other pair of entities
Recall

- Recursive dependence on neighbor entities
- Sum of attribute and relational recall

\[ R(e,T,D) = a_i(e,T) + [1-a_i(e,T)] \times r_i(e,D) \times R[N(e),T,D] \]
Imprecision

- Recursive dependence on neighbor entities
- Sum of attribute and relational imprecision

\[
l(e_1,e_2,T,D) = a_A(e_1,e_2,T) + [1-a_A(e_1,e_2,T)] \times r_A(e_1,e_2,D) \times R[N(e_1),N(e_2),T,D]\]
Recall vs. Precision

- Tradeoff
  - Recall increases when using Relational Clustering
  - Imprecision also increases
- Balance of $r_A$ vs. $r_I$ determines which one increases faster
- Relational Clustering has a negative effect when there are more ambiguous relationships than identifying relationships
Overview

- Introduction
- Entity Resolution and Queries: Formulation
- Collective Entity Resolution and Relational Clustering
- Analysis of Collective ER Using Relational Clustering
- **Collective Resolution for Queries**
- Adaptive Query Expansion
- Experiments
- Conclusions
Collective Resolution for Queries

- Two phases: Extraction and Resolution
- Expansion Operators
  - Attribute
  - Hyper-edge
- Level 1 Expansion
- Naive Analysis
- Level 2 Expansion
- Further Expansion
Resolution Process

- Only want to resolve references pertaining to a query and any references it's dependent on
- Two Phases
  - Extraction Phase - identifies all relevant references that need to be resolved for answering query
  - Resolution Phase - extracted entities collectively resolved with relational clustering
Attribute Expansion Operator $X_A$

- Given attribute $A$ and value $a$
  - $X_A$ returns all references whose $r.A$ attributes exactly match or are D-similar

- Level-0 References
  - All references D-similar to a query attribute
  - $\text{Rel}_0(Q) = X_A(a,D)$
Hyper-edge Expansion Operator $X_H$

- Given a set of references $R$
  - $X_H$ returns all references that share hyper-edges with the set $R$
- Level-1 References
  - All references that co-occur with level-0 references using collective resolution
  - $\text{Rel}_1(Q) = X_H(\text{Rel}_0(Q))$
Further Expansion

- Alternate $X_A$ and $X_H$ operators
- Level-2 References
  - Compare attributes for Level-1 references
  - $\text{Rel}_2(Q) = X_A(\text{Rel}_1(Q))$
- Level-3 References
  - Second-order neighbors that co-occur with Level-2 references
  - $\text{Rel}_3(Q) = X_H(\text{Rel}_2(Q))$
- And so forth...
Example
Overview

- Introduction
- Entity Resolution and Queries: Formulation
- Collective Entity Resolution and Relational Clustering
- Analysis of Collective ER Using Relational Clustering
- Collective Resolution for Queries
- **Adaptive Query Expansion**
- Experiments
- Conclusions
Adaptive Query Expansion

- Necessity
- Adaptive Expansion Estimation
  - Adaptive Hyper-edge
  - Adaptive Attribute Expansion
- Ambiguity Estimation
Unconstrained Expansion

- Expansion can continue for many levels
- Relevant record size becomes very large
  - Resolution can not be done in query time
- Problem: All co-occurrences treated equally
- Solution: Only select *most important* co-occurrences
Adaptive Expansion

A Mouse Immunity Model
- W Wang
- C Chen
- A Ansari

A Better Mouse Immunity Model
- W Wang
- A Ansari

Measuring Protein-bound Fluoxetine
- L Li
- C Chen
- W Wang

Autoimmunity in Biliary Cirrhosis
- W W Wang
- A Ansari
Adaptive Hyper-Edge Expansion

- Set an upper-bound $h_{\text{max}}$ on number of new references $h$-expansion can generate
  - $|X_H(\text{Rel}_i(Q))| \leq h_{\text{max}}|\text{Rel}_i(Q)|$
- Given $h_{\text{max}}$, choose the least ambiguous references from $X_H$
  - Provides most informative evidence
  - Sort $h$-expanded references
  - Choose top $k = h_{\text{max}}|\text{Rel}_i(Q)|$
Adaptive Attribute Expansion

- Set an upper bound $a_{\text{max}}$ on number of new references $a$-expansion can generate
  - $|X_A(\text{Rel}_i(Q))| \leq a_{\text{max}}|\text{Rel}_i(Q)|$
- Given $a_{\text{max}}$, choose the **most** ambiguous references from $X_A$ before expansion
  - Pick ones which need more evidence
  - Expand after to compare attributes on most ambiguous
Ambiguity Estimation

- Naive Estimate
  \[ \text{Amb}(r.A1|r.A2) = \frac{|d(b_{R.A2}(p_{R.A1=r.A1}(R)))|}{|R|} \]
  - \(|d(b_{R.A2}(p_{R.A1=r.A1}(R)))|\) - number of distinct values of A2 among references with value r.A1 for A1
  - \(|R|\) - total number of references extracted

- If A2 is not correlated with A1, it is a good assumption that distinct values of A2 correspond to distinct entities
Ambiguity Estimation

• Better Estimate
  - \( \text{Amb}(r.A1) = \frac{|p_{R.A1=r.A1}(R)|}{|R|} \)
  - \( |p_{R.A1=r.A1}(R)| \) - number of references with value \( r.A1 \) for \( A1 \)
  - \( |R| \) - total number of references extracted

• Naive estimate is not a good indicator since number of references with an attribute does not always correlate to the number of entity labels for an attribute
1. Initialize RSet to {}
2. Initialize depth to 0
3. Initialize FrontierRefs to {}
4. While depth < d*
   5. If depth is even or 0
      6. \( R = X_A(FrontierRefs) \)
      7. else
      8. \( R = X_H(FrontierRefs) \)
   9. FrontierRefs = R
10. Add FrontierRefs to RSet
11. Increment depth
12. Return RSet
Overview

- Introduction
- Entity Resolution and Queries: Formulation
- Collective Entity Resolution and Relational Clustering
- Analysis of Collective ER Using Relational Clustering
- Collective Resolution for Queries
- Adaptive Query Expansion
- **Experiments**
- Conclusions
Experiments

- Data Sets
- Real Data
  - Relevant Size vs. Resolution Time
  - Entity Resolution Accuracy for Queries
  - Reducing Time w/ Adaptive Expansion
- Synthetic Data
- Limitations
Relevant Size vs. Resolution Time
Relevant Size vs. Resolution Time
Entity Resolution Accuracy for Queries
Entity Resolution Accuracy for Queries
Reducing Time w/ Adaptive Expansion
Reducing Time w/ Adaptive Expansion
Synthetic Data
Synthetic Data
Limitations
Overview

- Introduction
- Entity Resolution and Queries: Formulation
- Collective Entity Resolution and Relational Clustering
- Analysis of Collective ER Using Relational Clustering
- Collective Resolution for Queries
- Adaptive Query Expansion
- Experiments
- Conclusions
Conclusions

- Relational clustering with collective resolution improves upon traditional attribute-based approaches.
- Adaptive expansion process converges quickly with much fewer references needing to be resolved.
- Collective entity resolution can be done at query-time without sacrificing accuracy.
  - Proven on two real bibliographic data sets.