

Improving Grouped-Entity Resolution using Quasi-Cliques

PENNSTATE



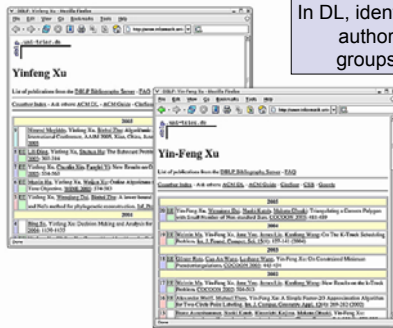
Byung-Won On, Ergin Elmacioglu, Dongwon Lee, Jaewoo Kang*, Jian Pei+

Penn State University, USA
 *Korea University, Korea
 +Simon Fraser University, Canada

The ER & GER Problems

- Entity Resolution Problem
 - Identifying matching entities that refer to the same real-world object
 - Main building block in many data applications
- Grouped-Entity Resolution
 - Entities have a group of repetitive elements
 - One can exploit the repetitive elements for improving ER accuracy
 - We propose to use Quasi-Clique

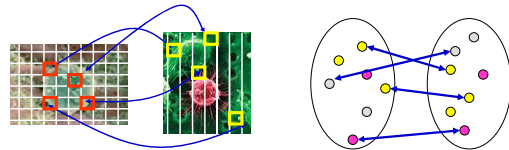
GER Example



In DL, identifying matching authors with their groups of citations

GER Examples

- In IRS, Identifying matching tax payers using their family information
- In Multimedia, retrieving matching images with $m \times n$ grids



Landscape

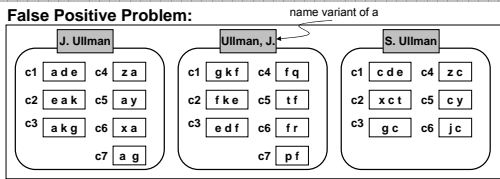
- Abundant research on related problems
- Also known as:
 - DB: approximate join, merge/purge, record linkage
 - DL: citation matching, author name disambiguation
 - AI: identity uncertainty
 - LIS: name authority control

Landscape

- In a nutshell, existing approaches often do:
 - For two entities, $e1$ and $e2$, capture their information in data structures, $D(e1)$ and $D(e2)$
 - Measure the distance or similarity between data structures: $dist(D(e1), D(e2)) = d$
 - Determine for matching:
 - If $d < threshold$, then $e1$ and $e2$ are matching entities
- Work well for common applications
- Ours do ER better when
 - Entities have structures (ie, repetitive groups) that we can exploit using graphs

Using Graphs

False Positive Problem:



- Our graph-based approach:
- Overcome the limitation of existing distance metrics
- Unearth the hidden relationships in contents
- Use Quasi-Clique to measure the strong relations

IEEE ICDM 2006

7

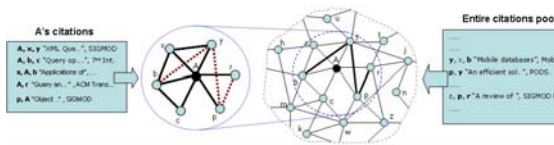
Using Graphs

- Represent entity e_1 as graph g_1 using common tokens
 - Author: co-author
 - Venue: common venues
 - Title: common keywords
- Superimpose the graph g_1 onto base graph B_1 to get a final graph representation G_1
 - Author: entire collaboration graph as B_1
 - Venue: entire venue similarity graph as B_1
 - Title: entire token co-occurrence graph B_1
- Measure the similarity of two entities e_1 and e_2 w.r.t. G_1 and G_2

IEEE ICDM 2006

8

Superimposition

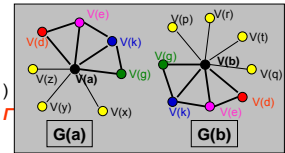


IEEE ICDM 2006

9

Quasi-Clique

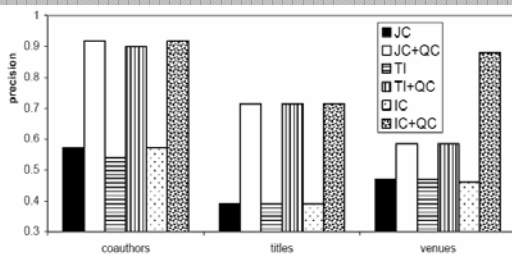
- Graph G
 - $V(G)$: set of vertices
 - $E(G)$: set of edges
- Γ -quasi-complete-graph ($0 < \Gamma \leq 1$)
 - Every vertex in G has **at least** Γ
- $V(S) (\subseteq V(G))$
 - $G(S)$: Γ -Quasi-Clique
 - => If $V(S)$ forms the graph satisfying Γ -quasi-complete-graph
 - $G(S)$: Clique
 - => If $\Gamma=1$
- Use **Quasi-Clique (QC)** to measure contextual distances
 - E.g., Function $QC(G(a), G(b), \Gamma=0.3, S=3)$



IEEE ICDM 2006

10

ACM Dataset



Precision:
 • k results are returned
 • r of k are name variants
 • $\text{precision} = r / k$

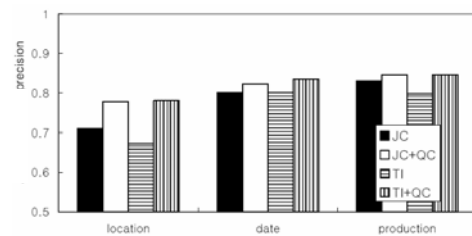
Attributes

- JC : Jaccard similarity
- JC+QC : JC + Quasi-Clique
- TI : TF/IDF Cosine similarity
- TI+QC : TI + Quasi-Clique
- IC : IntelliClean (venue hierarchy)
- IC+QC : IC + Quasi-Clique

IEEE ICDM 2006

11

IMDB Synthetic Dataset



IEEE ICDM 2006

12

Conclusion

- Many ER problems have entities with a group of repetitive elements
- Our Quasi-Clique based method exploits them to achieve improved accuracy
- Further improvement using **Groups** instead of Graphs
 - Group Linkage
 - Will appear in ICDE 2007