【VLDB2011勉強会】

Session 2: Entity Matching

担当：白川真澄（大阪大学）

Entity Matching

- **Entity Matching: How Similar Is Similar**
  - *Jiannan Wang (Tsinghua University), Guoliang Li (Tsinghua University), Jeffrey Xu Yu (Chinese University of Hong Kong), Jianhua Feng (Tsinghua University)*
- **Large-Scale Collective Entity Matching**
  - *Vibhor Rastogi (Yahoo! Research), Nilesh Dalvi (Yahoo!), Minos Garofalakis (Technical University of Crete)*
- **Linking Temporal Records**
  - *Pei Li (University of Milan – Bicocca), Xin Dong (AT&T Labs), Andrea Maurino (University of Milan – Bicocca), Divesh Srivastava (AT&T Labs)*

Entity Matching: How Similar Is Similar

Jiannan Wang (Tsinghua University), Guoliang Li (Tsinghua University), Jeffrey Xu Yu (Chinese University of Hong Kong), Jianhua Feng (Tsinghua University)

概要

- DBの二つのエントリが同じエンティティを指すか否か
  ⇒ 二つのエントリの類似度が閾値以上か否か

名前	電話番号	メールアドレス
Jeffrey Yi	852	Jeff
Jeffery Yi	852-111111	Jeffyi@abc.com

← 同じ人？別人？

- **何をもって「似ている」とするか？**
  - 様々な**属性**，**類似度計算手法**，**閾値**

上記の三つの要素を定式化し，
教師データから最適なルールを導出

  - 入力：ルール集合，正解データ集合
  - 出力：最適なルール

入力と出力の例

ルール1: 名前の[類似度]が[0,1] && 電話番号の[類似度]が[0,1]
ルール2: 名前の[類似度]が[0,1] && メールアドレスの[類似度]が[0,1]

名前の編集距離の逆数が0.2以上 && 電話番号のJaccardが0.8以上
名前の編集距離の逆数が0.33以上 && メールアドレスのJaccardが0.7以上

教師データの正解率を最大化する類似度，閾値の導出

どうやって類似度・閾値を最適化？

- ヒューリスティックな三段階の手法でNP完全問題を解決

- SiFi-Greedy（≒貪欲法）
  - まずは貪欲法で個別の属性に対するルールを最適化
  - 冗長な類似度計算手法および閾値を効率よく削減（論文参照）
- SiFi-Gradient（≒最急降下法）
  - 同じルールに属する別の属性の影響を考慮
    閾値の近いルールのみを考慮しつつルールを修正
- SiFi-Hill（≒山登り法）
  - Gradientとほぼ同じだが，他の全てのルールを考慮

Large-Scale Collective Entity Matching

Vibhor Rastogi (Yahoo! Research), Nilesh Dalvi (Yahoo!),
Minos Garofalakis (Technical University of Crete)

概要

- “Collective” – 「集合的な」
  - Naiveとはおよそ対照的な意味，全体との関連を考慮
  - NP完全なので大規模になると近似とか経験則が用いられる

- どうやって**大規模**かつ**集合的**なEntity Matchingを**現実的な計算時間**で行う？

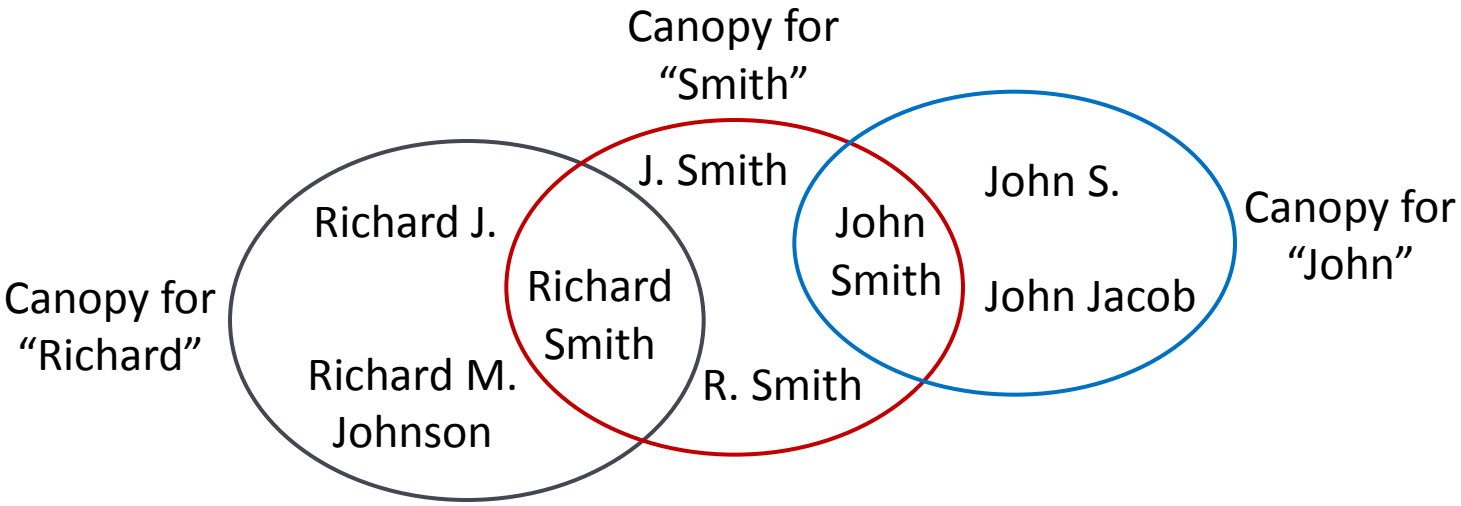
  類似している名前同士でクラスタ群を作り，
  クラスタ間で情報をやりとりすることで実現

  - 既存手法では数千程度のエンティティにしか適用不可
    ⇨ 数百万のエンティティに対しても適用可能に

Canopyを用いたEntity Matching（既存手法）

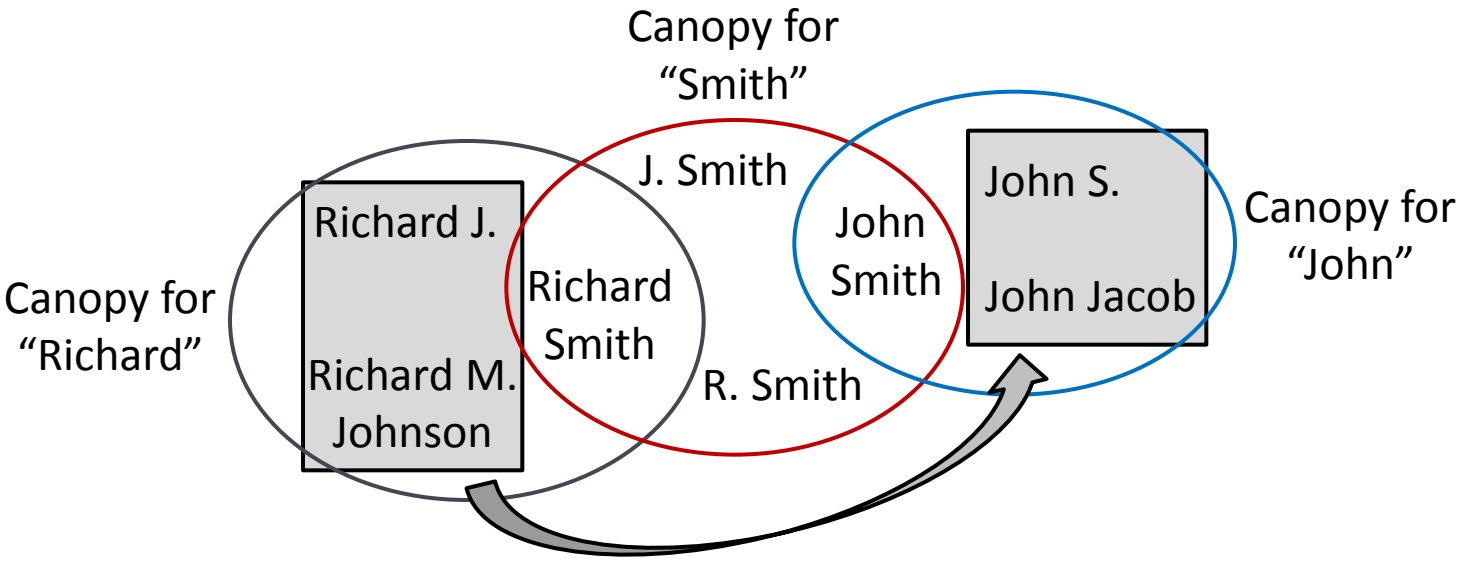
- String MatchingによりエントリをCanopy（クラスタ）に分割
  - 全ペア計算→Canopy内のみのペアの計算（Pair-wiseな手法の場合）
  - Canopy内で既存のCollectiveな手法を適用
  - でもCanopy内だけなんて，そんなの全然Collectiveじゃない！

Canopy間で情報交換してCollectiveに



Canopy間のメッセージ交換

- **新たな発見があったら他のCanopyに知らせる**
- 例えば共著者の関係から以下のルールがあるとする
  - Match(Richard J., Richard M. Johnson) → Match(John S., John Jacob)
  - Richard J.とRichard M. Johnsonが同じ人物であると判明した場合，他のCanopyに情報を送信 → John S.とJohn Jacobは同一人物



Linking Temporal Records

Pei Li (University of Milan - Bicocca), Xin Dong (AT&T Labs),
Andrea Maurino (University of Milan - Bicocca),
Divesh Srivastava (AT&T Labs)

概要

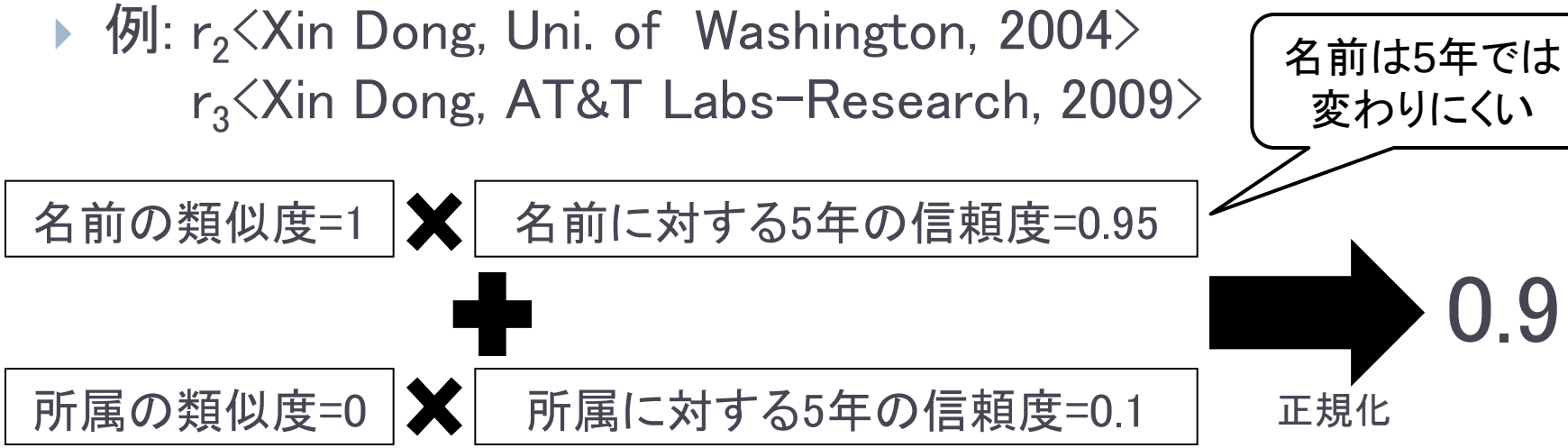
- Temporal＝時間的な（≠一時的な）
- 長期的な時間の制約をEntity Matchingに利用
  - 例：文献の発表年月，生没年など

時間の経過とともに変わりやすい情報を学習

時間の制約を用いてクラスタリング

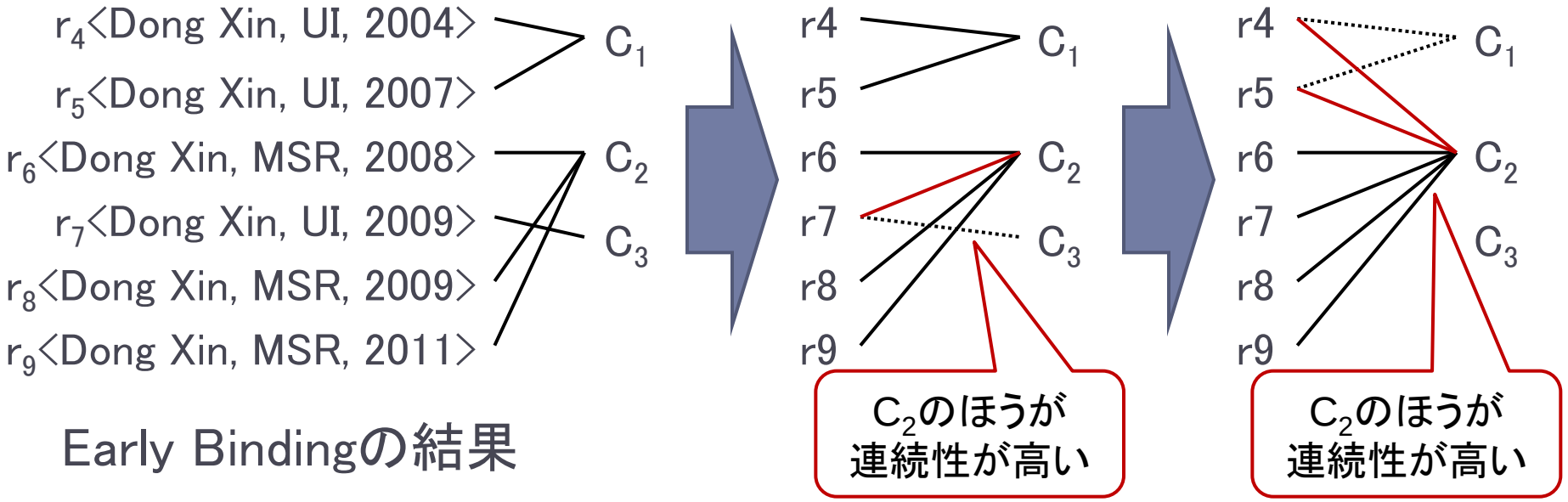
Decayという考え方

- 長い期間経ってると，文字列が一致してるとか一致してないとかいう情報があまり信頼できなくなる
- Disagreement Decay
  - University of Washington (01-07)
  - AT&T Labs-Research (07-)
- Agreement Decay
  - Adam Smith: (1723-1790)
  - Adam Smith: (1965-)
- Decayを教師データから学習し類似度の重みとして利用
  - 例: r_2<Xin Dong, Uni. of Washington, 2004>
    r_3<Xin Dong, AT&T Labs-Research, 2009>



時間的クラスタリング

- Early Binding: 時系列に処理，既にあるクラスタに併合
- Late Binding: 確率が最大となるようクラスタを生成
- Adjust Binding: EMアルゴリズム的（一番精度が良い）
  - 値の一貫性（類似度）および時間的連続性に基づく



Early Bindingの結果