【VLDB2010勉強会】

Session 17: Data Extraction

担当: 加藤 誠(京都大学)

Towards the Web of Concepts: Extracting Concepts from Large Datasets

Aditya Parameswaran (Stanford University, United States of America), Hector Garcia-Molina (Stanford University, United States of America), Anand Rajaraman (Kosmix Corporation, United States of America)

- データベースから「concept」を抽出
- concept: 「人気」があって，「簡潔」なk-gram
  - e.g. **Milky way, Harry potter, Pirates of the Caribbean**
  - **人気**: 数多く言及されている
  - **簡潔**: 同じことを表すのに最小の文字列

- 人気度(k-gram) = (k-gramのデータベース内出現回数)
- 簡潔度(k-gram) = ((k-1)-gramや(k+1)-gramとの人気度比)
  - e.g. 簡潔度("Harry potter")
    = 人気度("Harry potter") / 人気度("The harry potter")
  - k-gramと(k-1)-gramや(k+1)-gramとの比のminやmaxを利用

手法概要

- 簡潔度フィルタでの工夫
  - 「k > 2のとき，conceptと判断されたk-gram $t_1t_2\cdots t_k$の部分文字列である(k−1)-gram $t_1t_2\cdots t_{k-1}$ と$t_2t_3\cdots t_k$ が両方ともconceptであることはない」(**Wikipediaの分析より実証済み**)
    - e.g.「Pirates of the Caribbean」の部分文字列「Pirates of the」と「of the Caribbean」は両方ともconceptでない
  - これを利用してフィルタを強化

実験と感想

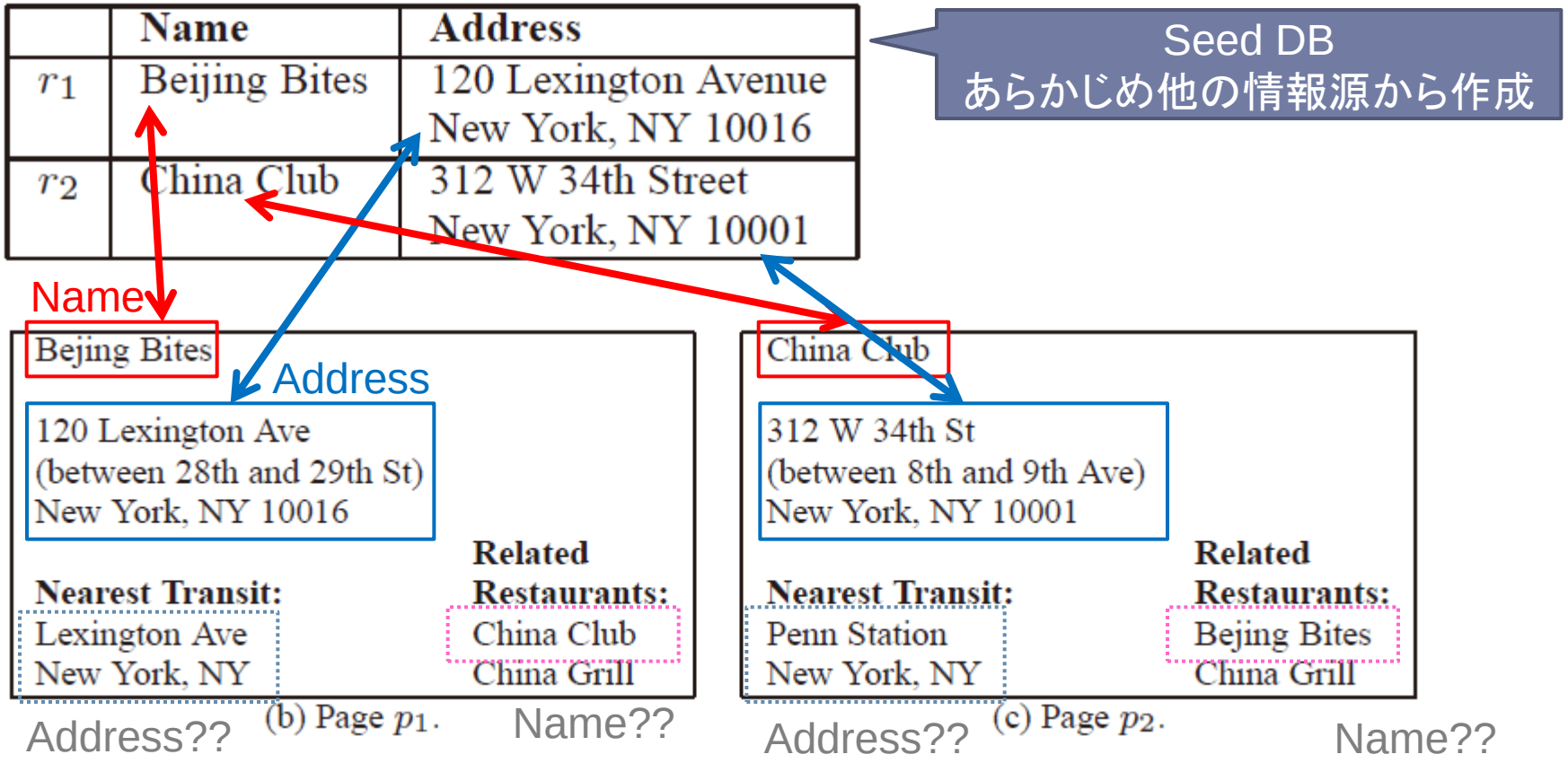
- 実験
  - AOLクエリログ（36M）よりconcept抽出
  - WikipediaのページタイトルとAmazon MTで評価
  - 適合率: 0.95 (> 0.92 by C-value Method)
  - 得られたconcept数: 25,000

- 感想
  - 問題も手法もあまり新しいようには見えない
  - ただし，手法の妥当性や複雑性などの分析が素晴らしい

Exploiting Content Redundancy for Web Information Extraction

Pankaj Gulhane (Yahoo! Research Labs, Bangalore, India), Rajeev Rastogi (Yahoo! Research Labs, India), Srinivasan Sengamedu (Yahoo! Research Labs, Bangalore, India), Ashwin Tengli (Microsoft India Development Center, Bangalore, India)

- Webからのスキーマ抽出・属性抽出

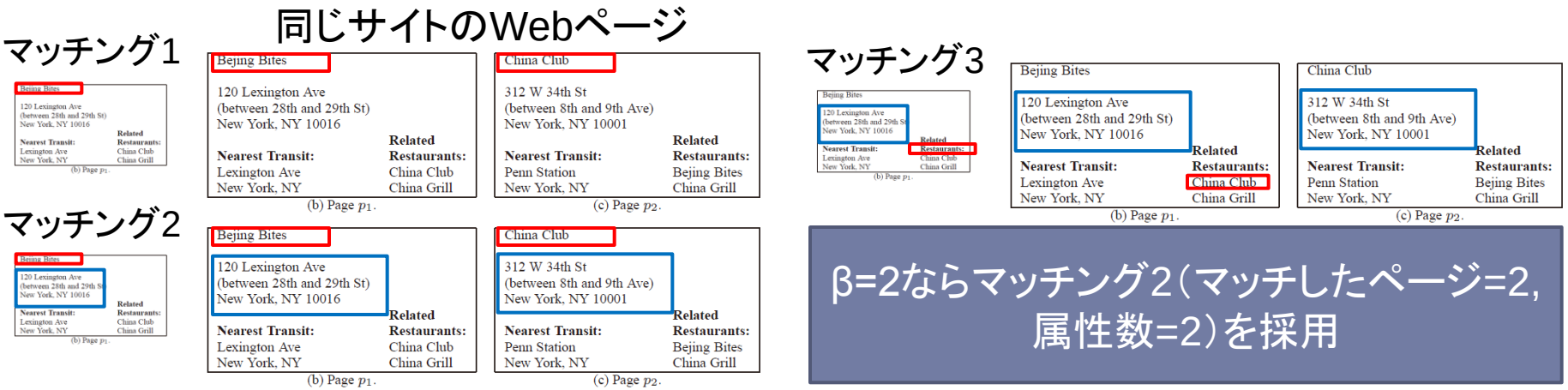


	Name	Address
r_1	Beijing Bites	120 Lexington Avenue New York, NY 10016
r_2	China Club	312 W 34th Street New York, NY 10001

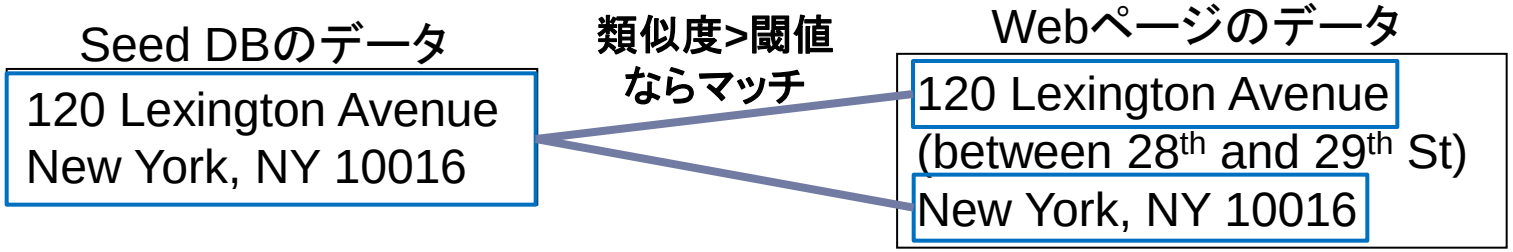
(b) Page p_1. (c) Page p_2.

手法概要

- **Seed DBとマッチしたページ数 がβ 以上で最大属性数のマッチングを採用**



- **属性値のマッチング方法（Strong Similarity）**
  - 文字列マッチする部分だけで類似度計算



実験と感想

- Restaurant(17sites)・Bib.(7sites)のデータを抽出

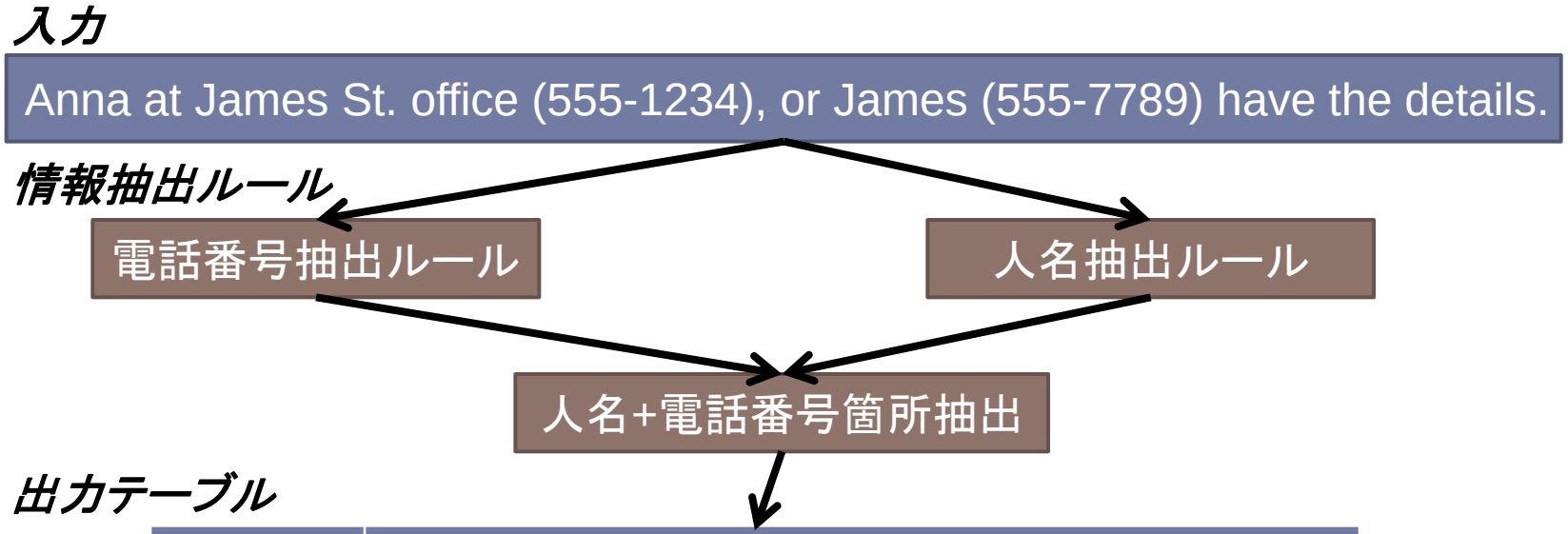
Restaurant		Bibliography	
Attribute	Precision	Attribute	Prec
Name	78.26	Title	96.14
Address	99.74	Author	98.12
Phone	100.00	Source	100.00
Payment	100.00		
Cuisine	100.00		

- 感想
  - マッチング箇所の発見が非常にテクニカル
    - 一方で難解すぎてこれ以上どうやって精度を上げるかわからない
  - スキーマ発見・データ抽出の研究はよく見るが,
    サイト数が多くない(7, 17sites)なら手動やった方が良さそう

Automatic Rule Refinement for Information Extraction

Bin Liu (University of Michigan), Laura Chiticariu Vivian Chu (IBM Almaden Research Center), H. Jagadish (University of Michigan), Frederick Reiss (IBM Almaden Research Center)

- 「情報抽出ルール」を改善するために，どこを直したらよいかユーザに提示（Data provenanceと深い関係）

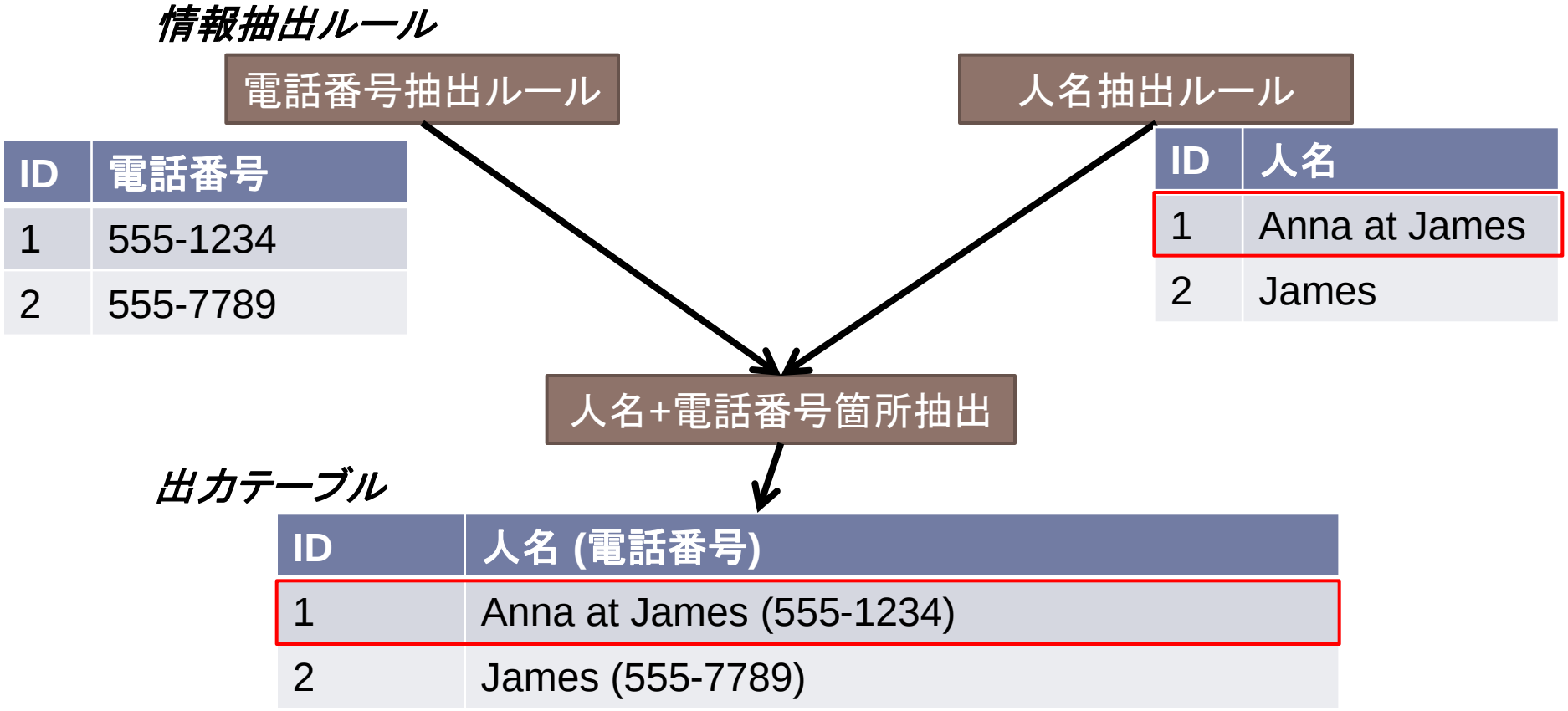


ID	人名 (電話番号)
1	Anna at James (555-1234)
2	James (555-7789)

Miss!!

High-level Changeの発見

- High-level Change: どこのルールのどの出力を削除したら結果が良くなるか？



Low-level Changeの発見

- Low-level Change: どこのルールをどう直したら結果が良くなるか？（精度が良くなる順に提示）

人名抽出ルール

ID	人名
1	Anna at James
2	James

修正案

人名フォーマット「First (Middle) Last」を追加

人名の長さを制限

- **問題**
  - 修正案の組み合わせ数が無限
    （修正による影響箇所を考慮した精度の計算が必要）
- **解決方法**
  - High-level Change発見時にどのデータの変化が結果にどのように影響を与えるかを保持し利用