CiNii (Articles) の名寄せ

国立情報学研究所
大向 一輝
Twitter: @i2k

はじめに

NII

- 総合目録構築のラディカルなアプローチ
  - すでにある書誌を利用する
  - 個別の館は独自に（ばらばらに）目録を作る
  - 個別の目録を自動的に統合して総合目録を作る
  - 実現可能性は？
- 課題
  - 同じものを同じものとして認識する（同定）
  - 別のものを同じものとして認識しない（分類）
  - 名寄せ問題

名寄せ問題

NII

- 情報空間内の複数のエンティティを同一のものとみなすこと
  - エンティティ（個物）：人・モノ・概念…
  - 単一のデータベース内での多重登録の解消
  - 異なるデータベースに存在するエンティティを同一視
- アプローチ
  - ID
  - コンテンツ
  - 自己申告

IDによる名寄せ

NII

- エンティティを特定可能な識別子
  - 図書・雑誌：ISBN・ISSN
  - デジタル資源：DOI・URI
  - 商品：JANコード・ASIN
- 信頼できるID管理システムが必須
  - エンティティの出現と同時に（あるいは先行して）IDが発行される
  - IDの再利用や複雑な付番ルールがない
  - 管理組織・管理システム
    - 日本図書コードセンター・DNS…
- IDのない世界、ID管理システムから漏れているものへの対応
- 予稿集・同人誌…

コンテンツによる名寄せ

NII

- 異なるエンティティにおけるコンテンツの一致、あるいは類似性によって第三者が同一性を判定する
  - 内容そのもの
  - メタデータ：タイトル・著者名・発行年月日…
- 名寄せの主体
  - 典拠：組織活動による人手での名寄せ
    - ワークフローへの組み込み
    - 教育・研修制度
- 本質的な困難さ
  - 本当に同一なのか？という疑問に答えられない
  - 主体となる組織・コミュニティへの信頼によって成立

自己申告による名寄せ

- 情報の作成主体による名乗り
- 主体への信頼によって成立
  - なりすまし・ミスの可能性
  - 第三者による認証に基づいて信頼性を担保
    - 公的認証
    - コスト
    - 例：ねんきん特別便

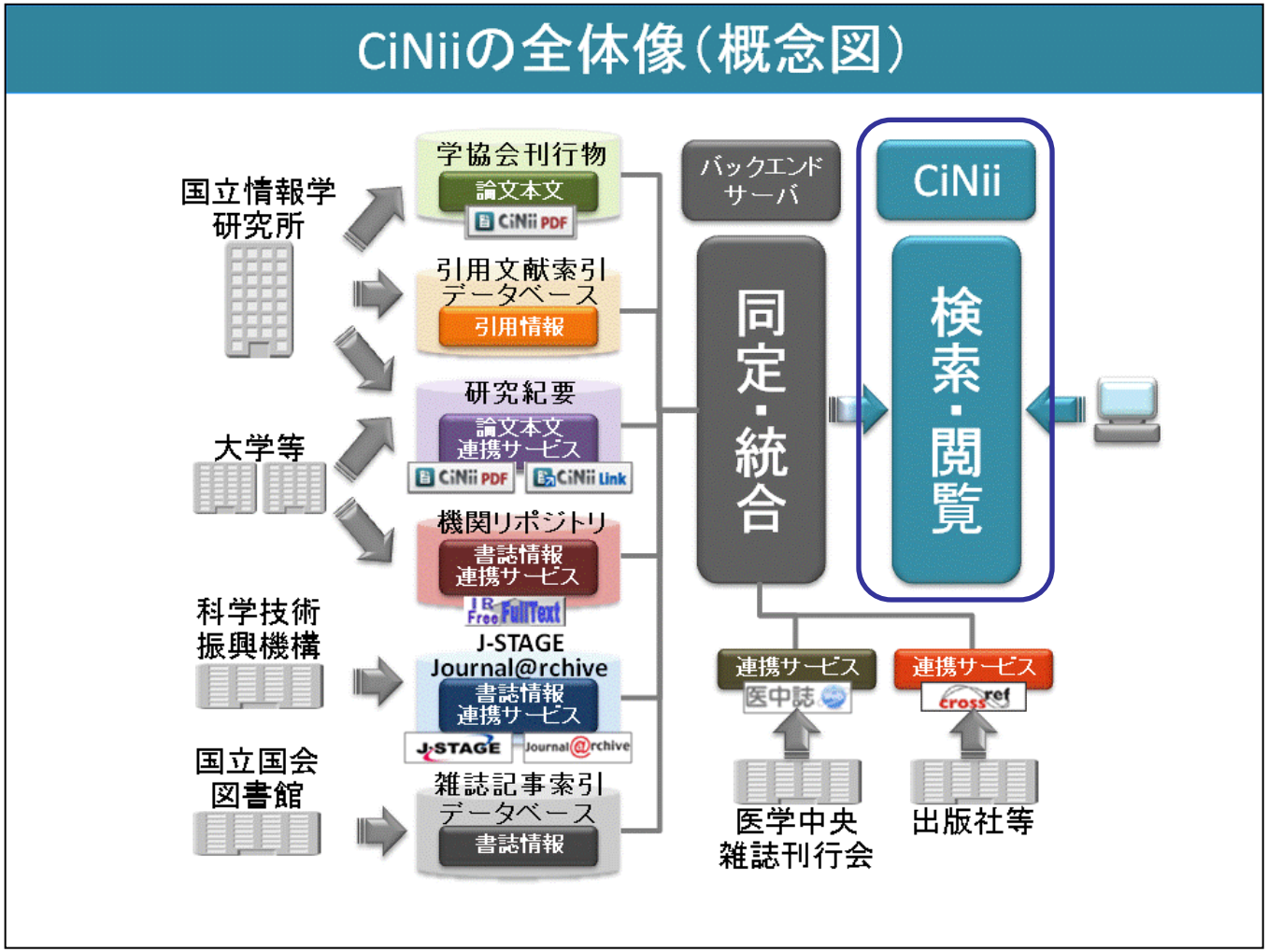
CiNiiの概要

NII

- 書誌検索＋本文提供サービス
  - 複数の中小～大規模データベースの書誌情報を統合して提供
    - ELS（学協会・紀要）400万
    - CJP引用 2400万
    - 雑索 800万
    - J-Stage 200万
    - 機関リポジトリ 50万
    - 学会・出版社…
  - 週あたり10万件ペースで増加

CiNiiとは

- 自動・手動処理を組み合わせて同定・統合



CiNiiとは

NII

- データ数・増加数・更新頻度（2009.8）

データベース名	データ数	年間増加数	更新頻度	本文	料金
NII-ELS学協会刊行物	約303万件	約18万件	週次	○	一部有料
NII-ELS研究紀要	約87万件	約3.5万件	週次	△	無料
引用文献索引データベース	書誌：約154万件 引用：約1661万件	書誌：約14万件 引用：約161万件	10回/年	×	無料*1
雑誌記事索引データベース	約827万件	約40万件	週次	×	無料
機関リポジトリ	約30万件	不定	週次	○	無料
J-STAGE/Journal@rchive	約8万件	不定	数回/年	○	無料
CiNii合計*2	約1239万件	約70万件	週次		

＊1　参考文献/被引用文献の閲覧は制限あり。
＊2　重複データが統合されるため、単純合計とは一致しない。

CiNiiの概要

NII

- データの由来上、複数の情報源に同じエンティティが存在する可能性
  - 例：ELSと雑索はほぼ包含関係
- 各データベースに共通のIDはない
  - 将来的にはジャパンリンクセンター発行のDOI？
- 異なるメタデータ記述ルール
  - 完全一致しない
  - 著者が複数名である時の記述ルールなど
- 分量的に人手での名寄せは現実的でない

機械による名寄せ

NII

- 名寄せの自動化に向けて
  - 情報学研究の成果を活用
  - 人間の知識をコンピュータに転移する
  - ワークフローへの組み込み・実運用が課題
- 知識ベース
  - if – then ルールの集合体
    - 正しい知識が記述できれば確実に動作する
  - ルールベースで完全に記述できるか？
    - 経験・勘・常識が多く含まれる場合には難しい
    - 適用範囲が限定的

機械による名寄せ

NII

- 機械学習
  - 既知の事例からモデル（ルール）を導く
  - 新しい事例をモデルに適用する
  - 適用範囲の広さ
  - モデルの質が悪ければ悪い結果が出る
- 二値分類問題への変換
  - 任意のペアが同一か否かを判定する

機械学習

NII

- ペアA
  - 大向一輝, 武田英明: 学術情報サービスの現状と展望（「特集：メタデータを斬る」）, 情報処理, Vol.50, No.1, 2012.
  - 大向一輝, 武田英明: 学術情報サービスの現状と展望, 情報処理, 50(1).
- ペアB
  - 大向一輝ほか: 学術情報サービスの現状と展望（「特集：メタデータを斬る」）, 情報処理, Vol.50, No.1, 2012.
  - 大向, 竹田: 学術情報サービスの昨日と明日, 情報処理, 40(1).
- 二値分類問題への変換
  - 「条件1、条件2、…、条件n → 同一 or 同一でない」の形に
  - 条件（素性）
  - 第一著者が一致・共著者が一致・タイトルが包含関係…

機械学習

NII

- ペアA：1, 0, 1, 1, … , 1 → 1
- ペアB：0, 0, 1, 0, … , 0 → 0
- ペアC：…… → 0
- …
- ペアn：…… → 1

- 確率モデルへの落とし込み
  - SVM（サポートベクタマシン）
  - ニューラルネット
  - 遺伝的プログラミング
  - …

機械学習

NII

- 結果の評価
  - 精度（Precision）・再現率（Recall）・F値
    - 精度と再現率はトレードオフ
  - その他
    - 偽陽性・偽陰性・確信度…

CiNiiの書誌名寄せ

NII

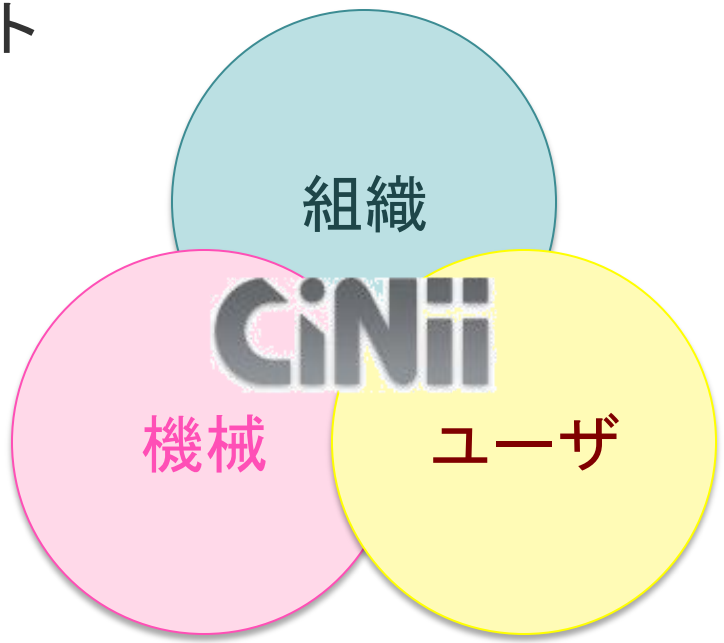
- 判別すべき候補（週次）
  - 10万（新規）対4000万（既存）のペア
    - すべてのペアを対象にすると計算時間が足りない
    - 検索エンジンでペア候補を絞る
- 確信度によって自動分類できるものとできないものを分別
  - できないものは手動判別へ
- NII書誌IDを付番
  - 後日名寄せされたものはリダイレクト情報で対応

CiNiiの著者名寄せ

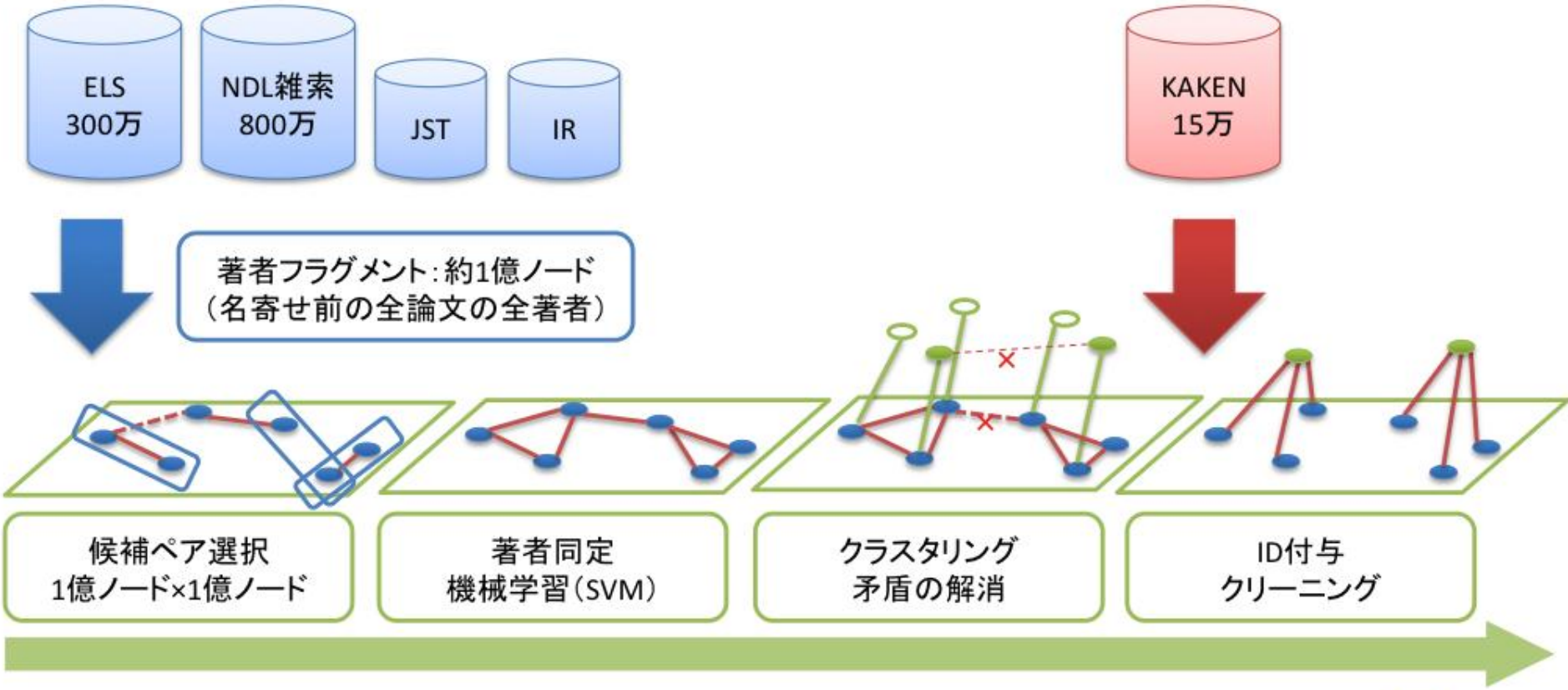
- 構造化されてこなかった情報への対応
  - 既存データに手をつけるコスト
    - 分量・ワークフロー
  - 完全性の保証
- 代表的な例：著者ID
  - 高まる重要性
    - 個人の業績管理
    - 国際競争（ResearcherID・ORCID）
  - 著者名典拠がない
    - 論文の著者名は膨大かつロングテール
    - 同姓同名・旧姓・タイプミス…

CiNii著者検索

- NII著者ID (NRID) の導入
  - 科研費番号＋機械処理による著者へのID付与
  - 著者ごとのページを生成
- NRIDベースの論文検索機能
  - 著者名→IDリスト→論文リスト
  - APIの提供
- 新たなデータ生成・管理モデル
  - 研究成果の活用
  - ユーザーフィードバック



CiNii著者検索の概要



- 著者ごとの論文リスト
- 同姓同名問題の解決
- フィードバック
- 業務フローへの組み込み
- API提供

NII著者ID（NRID）と
論文ID（NAID）の対応表

NRID
800万

CiNii著者検索

- ALS（Author Linking System）
  - i-Linkage（NII相澤教授）の大規模・実運用システム
  - CPU32コア・メモリ320GB・計算時間5日（全件処理）
- フィードバック（同一人物の報告）
  - 機械処理だけで100%の精度を得ることは不可能
  - あらかじめフィードバックを織り込んだシステム・アルゴリズム設計
    - 例：過統合より未統合を指摘する方が簡単
  - 実績：6217件（4月1日～7月15日）
    - Researchmap経由で研究者本人からのフィードバックも可能に

信頼とは何か

- 名寄せの主体によって信頼の意味が異なる
  - 組織：精度は高いが網羅性は不明
  - 機械：網羅性は高いが精度は不明
  - 自己申告：どちらにもない情報に基づくが信じてよいのか？
- 必要十分な精度・網羅性を最小のコストで実現する
  - 異なるアプローチのベストミックス