NTT

NTT Information Sharing Platform Laboratories
NTT情報流通プラットフォーム研究所

セマンティックWeb技術を用いた社内情報の連携

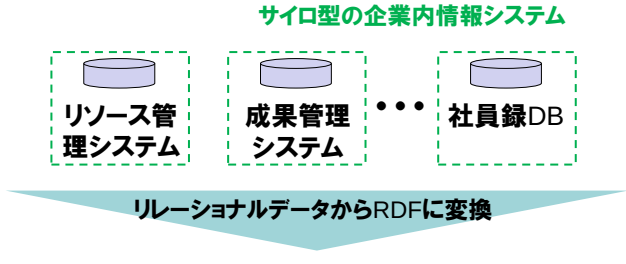
森田 大翼、飯塚 京士
（日本電信電話株式会社 NTT情報流通プラットフォーム研究所）

セマンティックWebコンファレンス2012
2012年3月8日（木）

NTT

RDFを用いた企業内情報活用

- 企業内に蓄積されている電子情報は増加の傾向にある
- 企業内情報を検索・分析し、ナレッジとして再利用するニーズが高まっている
- しかし、多くの情報は企業ナレッジとして有効に活用できていない
  - 企業内システムは各機能に個別最適なサイロ型であることが多い

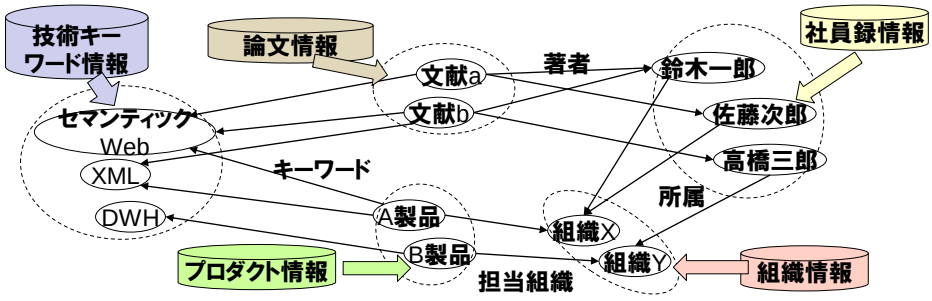


RDFでデータを連結し、企業内情報の様々な関連性を横断的に検索することを可能にする

NTT

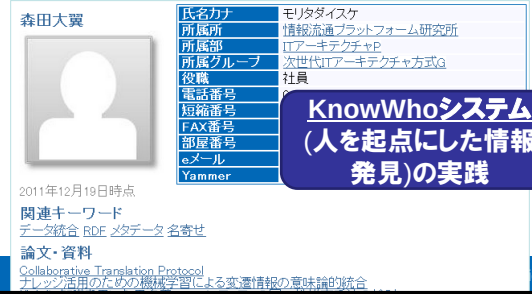
企業内情報活用の実践

・企業内情報を繋げると、有用なナレッジを得ることができる



企業内情報の連携により、
- 効果的な情報の検索
- 意外な関係性の発見

を可能にする

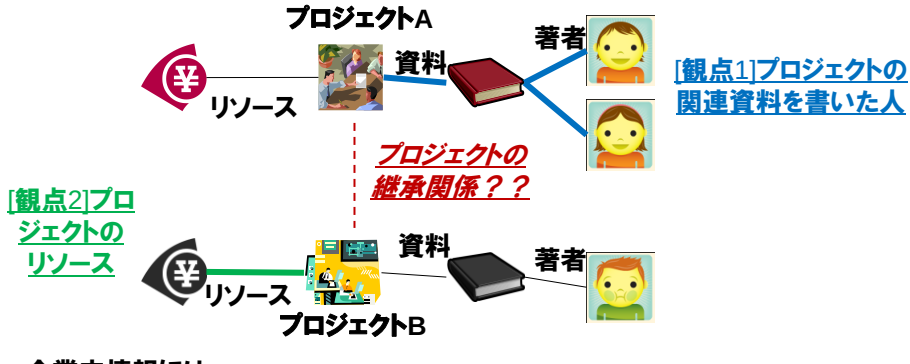


KnowWhoシステム
(人を起点にした情報発見)の実践

NTT

企業内情報活用における問題

・繋がっていないデータからはナレッジを抽出できない！
  – 下図の場合、プロジェクトの継承関係は繋がっていない

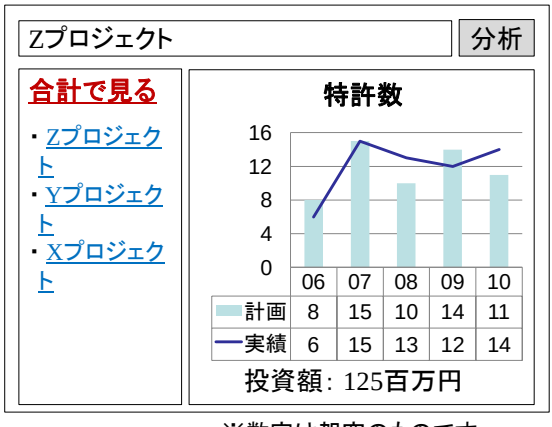


企業内情報には、
- 四半期・年度などのある一定期間毎に発行されるデータが多い
- その期間毎のデータ間のつながりが管理されてない場合が多い

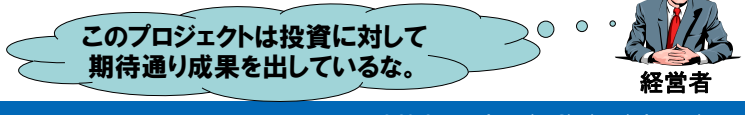
という、活用の障壁となる問題がある

継承関係を繋げることによるメリット例

・関連する情報を一元的に集計、分析することを可能にする



※数字は架空のものです



本研究の課題

・期間の移り変わりにより名称やIDなどの属性が変化するデータが，**実態として同一である関係**に基づくデータ統合に着目する

Zプロジェクト

Yプロジェクト
（2008～2009年）

Xプロジェクト
（2006～2007年）

名称は異なるが、実態として同一業務を行なっているプロジェクトのデータの統合

定義

「変遷を伴う同義性」:
異なる時間軸上で，実体として同一である関係

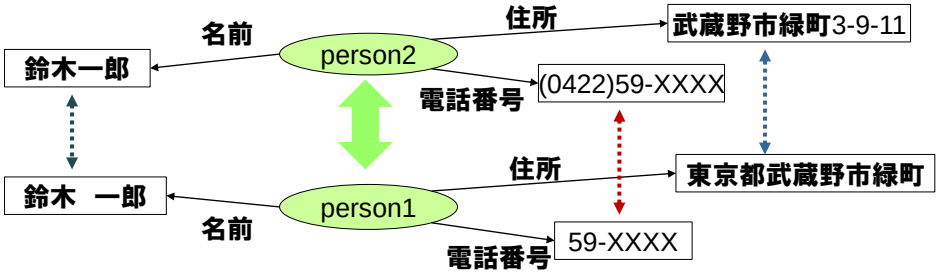
変遷を伴う同義性に基づく情報統合を本研究の課題とする

従来のデータ統合技術

目的 現実世界で一つの存在であるものを、データ上でも一つで表現したい

前提 データ間の同じ属性値の文字列は類似している

アプローチ 属性値間の文字列類似度を元に、データの統合を行う

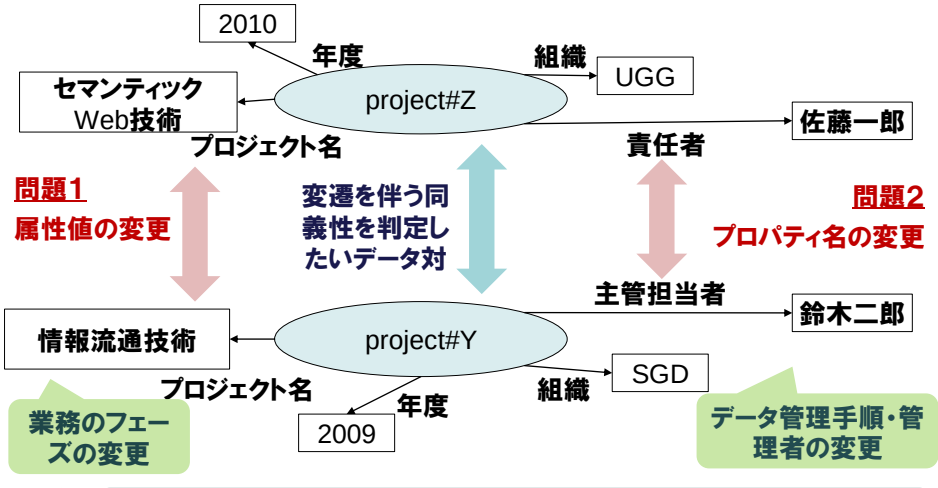


判定ルールを人手で作成する手法
[Shen 2005], [Whang 2009] など

判定ルールを機械学習する手法
[Bilenko 2003], [Chaudhuri 2007] など

変遷を伴う同義性によるデータ統合

前提 属性値が時間の経過に伴って、変更が生じる場合がある



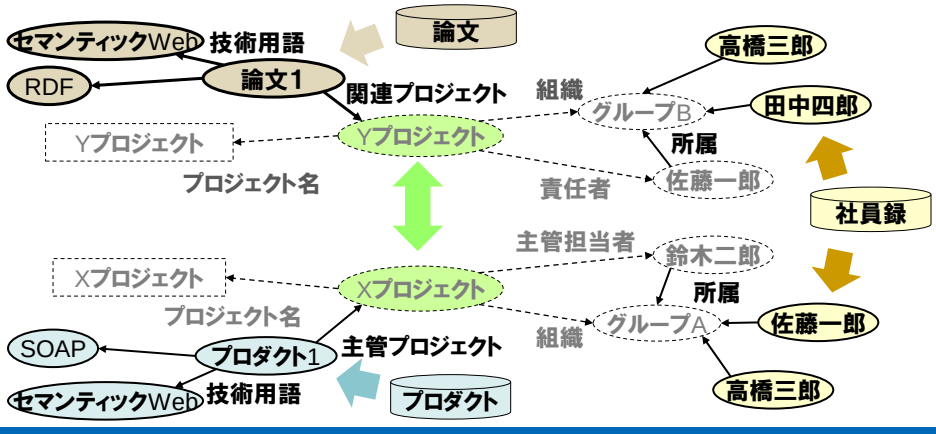
従来手法とは異なるアプローチが必要である

本研究のアプローチ (1/2)

アプローチ1：周辺情報を、変遷を伴う同義性を判定する情報として利用

周辺情報：組織のリレーショナルデータの繋がりを利用して得られる関連情報

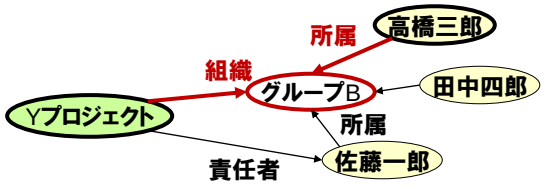
⇒ 直接の属性値以外の情報を利用する



本研究のアプローチ (2/2)

アプローチ2：繋がりの意味論を考慮したリソース対の特徴抽出

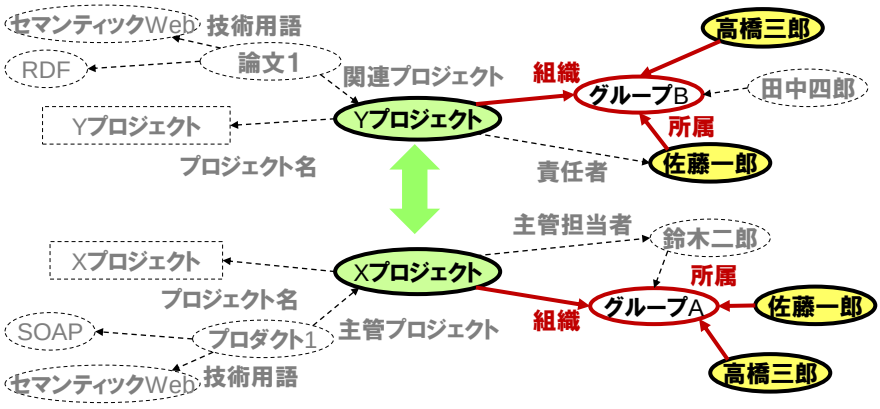
- **（本研究における）意味論**
  - 2つのデータ間を繋ぐアークのラベル（プロパティ）の繋がり
    - 例）「Yプロジェクト」と「高橋三郎」は、「組織：所属」というプロパティで繋がっている
      - 繋がりは矢印の方向の順向き・逆向きの両方とも辿ることができる
    - 機械的に解釈できる
  - 人間が解釈すると上記例は「高橋三郎は、Yプロジェクトの担当組織の所属人物である」となる。



NTT 意味論に基づくリソース対の特徴抽出(1/3)

- 共通の周辺情報に対して、共通の意味論を利用する(方式 1)
  - 共通の周辺情報:「佐藤一郎」、「高橋三郎」
  - 意味論:
    - 研究テーマY:「組織:所属」
    - 研究テーマX:「組織:所属」

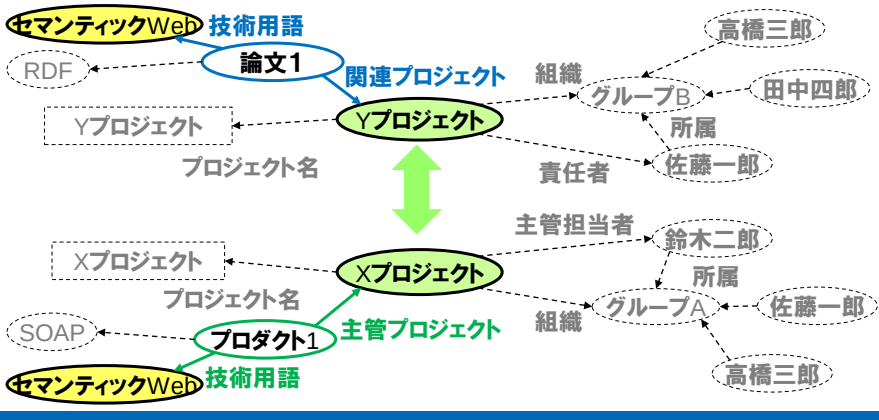
両研究テーマに対して、担当組織に属している人が共通に2人いる



NTT 意味論に基づくリソース対の特徴抽出(2/3)

- 共通の周辺情報に対して、異なる意味論を利用する(方式 2)
  - 共通の周辺情報:「セマンティックWeb」
  - 意味論:
    - 研究テーマY:「(論文の)関連研究テーマ:技術用語」
    - 研究テーマX:「(プロダクトの)主管研究テーマ:技術用語」

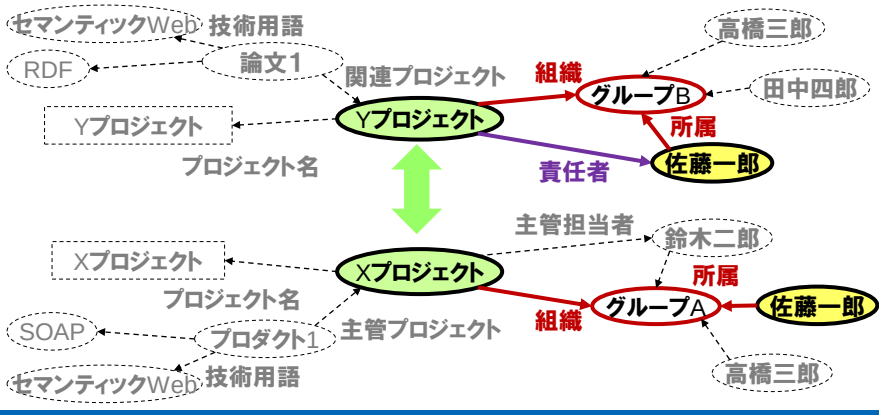
意味論は異なるが、共に関連する技術である



意味論に基づくリソース対の特徴抽出(3/3)

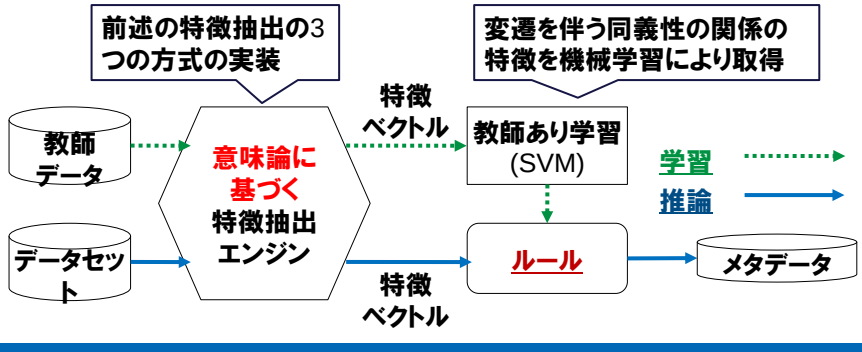
- 共通の周辺情報に対して、複数の意味論を利用する(方式 3)
  - 共通の周辺情報:「佐藤一郎」
  - 意味論:
    - 研究テーマY:「組織:所属」且つ「責任者」
    - 研究テーマX:「組織:所属」

複数の意味論は、関連するデータとして強い繋がりを持つ



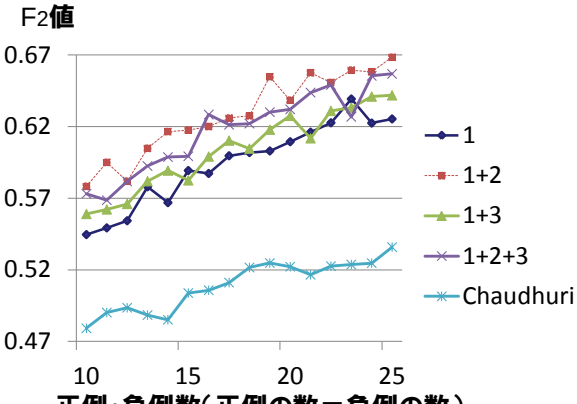
機械学習フレームワーク

- 2つのデータ間に共通する周辺情報の繋がり方(意味論)の特徴を抽出した
- その特徴を用いて、データ連携のためのルールを作成
  - 先行研究:人手で作成 [Shen 2005], 教師あり学習で作成 [Chaudhuri 2007]
  - 大規模・多様な企業内データに対して、高い精度を実現するルールを人手で作ることは現実的に困難
  - 本研究では教師あり学習を検討する



変遷を伴う同義性の判定手法の評価実験

研究所データの変遷を伴う同義性判定実験により、アプローチの有効性を確認した



正例・負例数（正例の数＝負例の数）

方式	1	1+2	1+3	1+2+3
特徴ベクトルの次元	29	121	49	549

- **実験データ**
  - NTT研究所の一部組織の2003～2009年のプロジェクト
  - 581対（変遷を伴うデータは33対）
- **有効性の確認**
  - 方式 1+2の場合、F2値で、0.1～0.13ポイントの精度改善
  - 先行研究は正例・負例数が18辺りで精度改善が停止
- **発見した課題**
  - 方式 1+2+3の次元数が大きすぎることによる精度低下
  - 今後、ヒューリスティクス等を用いた次元削減を検討する

まとめ

- まとめ
  - 企業内システムは各機能に個別最適なサイロ型であることが多く、蓄えられている情報を有効に活用できていない
  - RDF化して情報を連携した社内の実践により、様々な観点で情報を検索・分析できることの可能性を確認すると同時に、期間またがりの情報については繋がりが定義されてない場合が多く、情報活用の障壁となっていることを発見した
  - 本研究では、「変遷を伴う同義性」に基づく情報の統合を課題とした
  - 様々なデータを繋げることによって得られる周辺情報と、その意味的な繋がりの特徴を用いるアプローチを採用した
  - 所内データを用いた実験により、変遷を伴う同義性判定という問題に対して周辺情報と意味論を用いたことが有効なアプローチであったことを明らかにした
- 今後の予定
  - 所内の大規模データ・多様なデータへの適用の有効性確認
  - 本技術の応用を調査し、有効な適用分野の検討