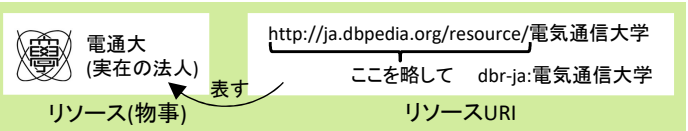
複数の大規模RDFデータを用いたWeb文書の意味的検索

情報・ネットワーク工学専攻 兼岩研究室 山中佑紀

大規模RDFデータと意味的関係性

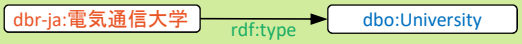
RDFとは | Resource Description Framework



事実(関係性) 電気通信大学は大学の一つである。

RDFトリプル (dbr-ja:電気通信大学, rdf:type, dbo:University)
主語 述語 目的語

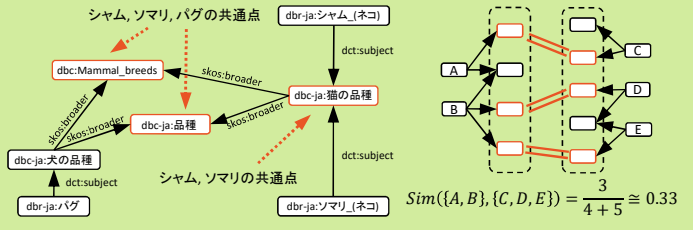
RDFグラフ



提案手法：集結パスと類似度関数

集結パス
- 複数のリソースから、1つの共通のリソースへ至るRDFパスの集合
- 共通のリソース(共通点)を中心にした関係性を示す構造

類似度
- 2つのリソース(またはリソース集合)間にある共通点の数の割合
- リソースの間の関係性の強さを表す数値



リンクトデータと同値類リソース

リンクトデータセット

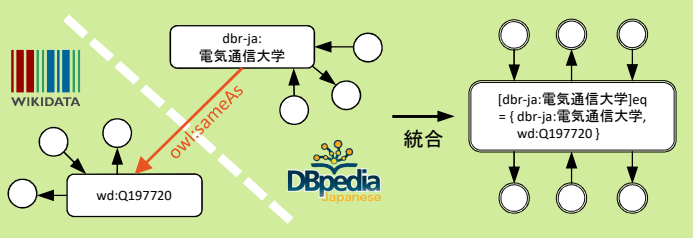
リンクトデータ
- 「グラフ構造」と「共通のURI」というRDFの2つの特徴によって、別々に作成されたデータを組み合わせて使うことができる

DBpedia
- Wikipediaから半自動生成されるリンクトデータセット
- 日本語版：96万項目, 英語版：467万項目 (それぞれ2016年4月版)

Wikidata
- Wikimediaのデータ統合基盤として作られている知識ベース
- 全言語共通 4269万項目(2018年1月現在)

問題提起

- 使用するURIを共通にすることは必須ではないため、同じリソースに対応するURIがリンクトデータセット間で異なることがよくある
- それらを同じものとして扱わないと、パス検索などの結果に複数のリンクトデータセットを組み合わせて使う効果が現れない



提案手法：同値類リソースの統合

同値関係プロパティ
- 2つのURIが同じリソースを意味していることを示すプロパティ(述語)
- 例：owl:sameAs, skos:exactMatch

同値類リソースの統合
- 同値関係プロパティで結ばれた、同じリソースを表すことが示されているURIの集合を一つに統合して扱う
- この集合のことを同値類リソースと呼ぶ

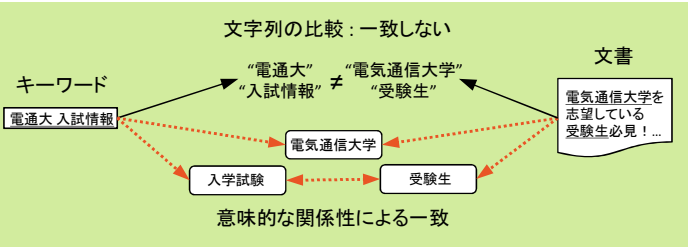
実験と結果
- ランダムに抽出した2つのリソースURI間のRDFパスを、使用するRDFデータの数を変えながら検索する実験を行った
- 統合を行うことで、検索結果の数に複数のリンクトデータセットを組み合わせて使う効果が現れるようになった

文書検索への応用

文書の意味的なキーワード検索

通常の文書のキーワード検索
- キーワードと文章の文字列比較なので、以下のような弱点がある
  - 単純な文字列の比較ではマッチしない文書の検索に弱い
  - 検索に予備知識やテクニックが必要

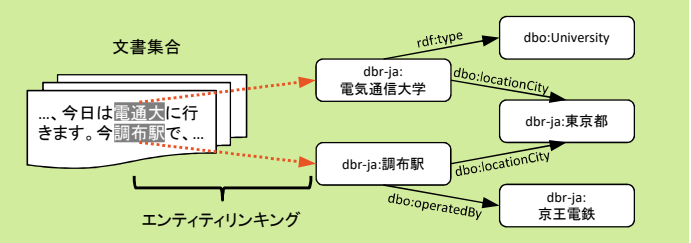
文書の意味的なキーワード検索
- キーワードと文章の間の意味的な関係性を元に文書を検索する
- 通常のキーワード検索の弱点を克服することが目的



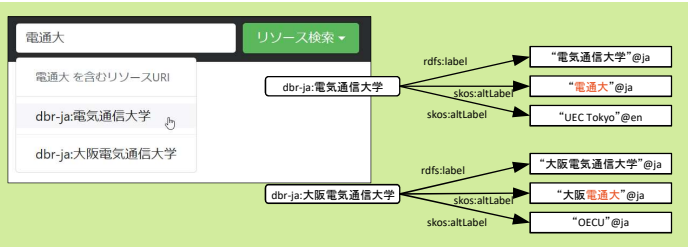
提案する文書検索システムの処理

日英2言語版のDBpediaとWikidataを関係性の情報源として用いて、意味的に文書を検索するシステムを実装した。

1. 検索対象とする文書を、Web上から収集する。収集したそれぞれの文書に対して、文中の単語・固有名詞が表している物事(リソース)と同じリソースを表すような、RDFデータ上のリソースURIを推定する。一般に、この処理はエンティティリンキングと呼ばれる。



2. 利用者が検索キーワードを入力する。システムは入力されたキーワードが表しうるリソースの一覧を利用者に提示し、利用者はその中から意図したものを選ぶ。



3. それぞれの文書に対して、文中に含まれるリソース集合と選択されたリソース集合の類似度を計算し、値が上位の文書を選び出す。
4. 選ばれた文書とキーワード、それぞれに含まれるリソース集合間の集結パスを検索する。
5. 文書と集結パスを検索結果としてグラフ表示する。