2N-05

知識ベースと外部情報源の統合利用環境

大森雄基* 北川博之† 天笠俊之‡
* 筑波大学大学院 システム情報工学研究群
† 筑波大学 国際統合睡眠医科学研究機構 ‡ 筑波大学 計算科学研究センター

1 背景と問題

知識を構造化して保存する RDF 形式の知識ベースは，様々な知識処理における重要な情報源として活用が進められつつある．RDF 形式の知識ベースは，現実の物事を IRI として表現した**エンティティ**，IRI や値の関係を主語 s・述語 p・目的語 o のトリプル $s \xrightarrow{p} o$ の形で集積して保存しており，全体としてはグラフ構造を持つ．これに対するクエリ言語として，SPARQL がある．一方，気象情報や運行情報といった即時性が重要な情報や専門的な知識など知識ベースではない情報源も多い．このため，知識ベースと外部情報源の連携利用が重要となっている．しかし，外部情報源は問合せと返り値に独自の形式を持つため，知識ベースと外部情報源の統合利用は一般に，極めて煩雑となる．一方，クエリ言語 SPARQL で様々な応用に対応した知識利用のために独自のユーザ定義述語を設定可能 [2] であるが，この機能は外部情報へのアクセスにも転用可能である．

本研究では，**ユーザ定義述語を利用して，外部情報源をあたかも知識ベースと一体であるかのように扱える統合利用環境のアーキテクチャ**を示す．また，外部情報源からの返り値に対して，知識ベース中の適切なエンティティを選択する手法について述べる．

2 関連研究

2.1 Federation

SPARQL を用いて複数の知識ベースを統合的に検索する技術は Federation と呼ばれ，CostFed[3] 等がある．SPARQL クエリの再構成と知識ベースの選択を行うことが基本技術である．しかし本研究では，知識ベースだけではなく非知識ベースの外部情報源を連携させるという問題を扱うため，これに合わせたアーキテクチャを設計する．

An Integrated Environment for Knowledge Bases and External Information Sources.
Yuuki OHMORI*, Hiroyuki KITAGAWA†and Toshiyuki AMAGASA‡
* †‡University of Tsukuba
1-1-1 Tennodai, Tsukuba 305-8573, Japan
* omori.yuki.sm@alumni.tsukuba.ac.jp {†kitagawa, ‡amagasa}@cs.tsukuba.ac.jp

2.2 エンティティリンキング

エンティティリンキングとは，自然言語文中の現実の物事に対応する語句などに対して，知識ベース中の同一の事物を指すエンティティを割り当てる行為である．TAGME[4], DoSeR[5] といった手法がある．

本研究では自然言語ではなく，外部情報源の返した値を統合するという問題を扱い，このためのエンティティリンキング法を設計する．

3 提案手法

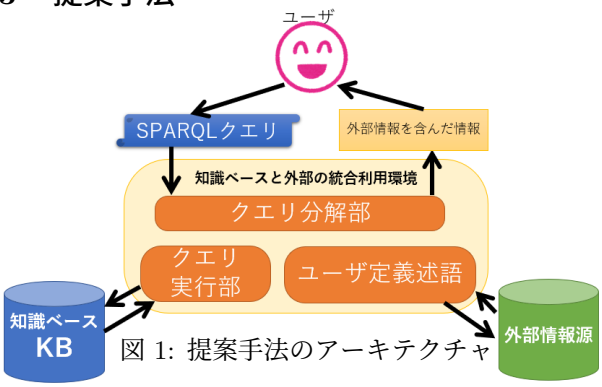

図 1: 提案手法のアーキテクチャ

SPARQL 問合せ例

```
SELECT DISTINCT * WHERE {
?x rdfs:label "Dan Hirschberg"@en .
?x rdf:type dbo:Scientist . ?x dbo:knownFor ?k .
?x dbp:workplaces ?w . ?w dbo:city ?c .
?x <m:co-author> ?y. }
```

提案手法のアーキテクチャを図 1 に示す．また，問合せ例を上に示す．ユーザ定義述語（以下 UDP と記す．例では<m:co-author>）を含むクエリを検索すると，UDP が外部情報源から情報を取得し，外部情報を含んだ検索結果を返す．これは以下のステップで行う．

1. UDP 以外の部分について、SPARQL クエリを実行
2. UDP に関する外部情報源から情報を取得
3. エンティティの割り当て
4. 最終的な問合せ結果の導出

Step 1：UDP 以外のクエリを実行 ユーザが送ったクエリから UDP（p_u とする）を含む部分を削除したうえで，通常の SPARQL クエリとして評価する．この結果を R_Q とする．また、R_Q 中に出現した p_u の主語を x_i とする．また、x_i と同じタプルに出現するエンティティの集合を Q_i とする．

Step 2：外部情報源からの情報取得　UDP 内の関数で、外部情報源のアクセスを行う．Step 1 で得た x_i を外部情報源インタフェイスに適した形式へ変換し $o(x_i)$ とする．これを用いて，外部情報源を検索したとき，それの返り値の集合を M_i とする．

Step 3：エンティティの割り当て　Step 2 で得た外部情報源の結果について，知識ベース中のどのエンティティに当たるか，もしくは無いかを判定する．各サブステップについては次に述べる．

Step 3-1：文脈エンティティの更なる抽出　x_i から n 回の述語を辿って到着可能なエンティティの集合を E_n^i とする．ただし，重複がある場合は最小の n の E_n^i のみに含ませる．これを N 番目まで取得する．

Step 3-2：候補エンティティの選出　Step 2 で得た外部情報源の結果 $m_{ij} \in M_i$ に対して，候補エンティティの集合 C_{ij} を，語句と候補エンティティとの対応辞書を用いて選び出す．辞書は，英語版 Wikipedia のアンカーリンク情報から作成する．

Step 3-3：最適なエンティティの決定　x_i 自身，Q_i，$E_1^i, E_2^i, \ldots, E_N^i$ をあわせて文脈エンティティ $ctx(x_i, N)$ と定義する．個々のエンティティ間の関連度は，対応する Wikipedia ページのアンカーリンク先の重複数とする．$ctx(x_i, N)$ の各エンティティに対して C_{ij} の各要素との関連度を合計し，関連度スコアを算出する．最適エンティティ y_{ij} は，関連度スコアが最大のものとする．ただし，閾値以下の場合は無しと判定する．

以上から，p_u の目的語 y_{ij} が取得できたので，これをタプルとしたものの集合 R_{UDP} を得る．

3.1　Step 4：最終的な問合せ結果の導出

Step1 の結果 R_Q と、Step3 の最適エンティティ R_{UDP} を結合して、最終的な問合せ結果をユーザに返す．

4　実験

4.1　実験環境

実験データには，知識ベースには英語版 DBpedia を用いた．また，外部情報源は DBLP を用いた．実験に用いた UDP は，主語 $?x$ について，その名前を示す述語を用いて名前 $o(?x)$ を取得し，論文データから共著者名を取得するものである．以下 4 つのクエリを作成・実行し，リンキング結果を検証した．正解判定は目視で行った．(1) 知識ベース中のある研究者を指定し，その共著者を取得するクエリ．(2) 特定分野の研究者を知識ベース中から探し，共著者を取得するクエリ．(3) 非情報系の指定した本の著者を知識ベース中から探し，共著者を取得するクエリ．(4) 茨城県の大学に所属する研究者を知識ベース中から探し，共著者を取得するクエリ．

4.2　実験結果

実験結果を表 1 に示す．なお評価は，UDP の主語とリンキング結果である目的語の組の集合について，リンキング結果と真の出力を比較したものを集計した．

表 1: リンキング結果：$N = 3$，閾値 4

	データ数	Precision	Recall	Accuracy
Query1	28	0.75	0.33	0.68
Query2	35	0.70	0.64	0.75
Query3	12	0.00	0.00	0.56
Query4	61	0.50	0.43	0.86

4.3　考察

エンティティ間のグラフ関係が豊富な場合は正しく判定でき，知識ベース中に存在しないエンティティの却下の割合も高い．一方で，エンティティ間のグラフ関係が疎な場合に，誤却下や誤ったエンティティに誘導される事例があった．

より高度なリンキングのためには，エンティティ間のグラフ関係だけでなく，エンティティに関連付けられた文字列を利用した判定も必要になると考えられる．

5　まとめ

本研究では、知識ベースと非知識ベース外部情報源の統合利用アーキテクチャを提案した．また、この問題に合わせたエンティティリンキングを設計した．今後の課題は，より多様な UDP を用いた検証，エンティティリンキングの高度化，UDP が複数ある場合のクエリ処理等である．

6　謝辞

本研究の一部は，JSPS 科研費 JP19H04114，AMED ムーンショット型研究開発事業による．

参考文献

[1] RDF 1.1 Concepts and Abstract Syntax, W3C Recommendation 25 February 2014 , https://www.w3.org/TR/rdf11-concepts/

[2] 3.3 Magic Properties, SPIN - Modeling Vocabulary, W3C Member Submission 22 February 2011 https://www.w3.org/Submission/spin-modeling/#spin-magic

[3] Muhammad Saleem, et al., "CostFed: Cost-Based Query Optimization for SPARQL Endpoint Federation", Procedia Computer Science, Volume 137, 2018.

[4] Paolo Ferragina and Ugo Scaiella, "TAGME: on-the-fly annotation of short text fragments (by wikipedia entities)", CIKM, 2010.

[5] Stefan Zwicklbauer, et al., "DoSeR - A Knowledge-Base-Agnostic Framework for Entity Disambiguation Using Semantic Embeddings." ESWC, 2016.