日本語Wikipediaオントロジーの構築および検索システムの実装

Building up a Japanese Wikipedia Ontology and a Search System

森田武史† 桜井慎也† 玉川奨† 和泉憲明‡ 山口高平†
Takeshi Morita† Shinya Sakurai† Susumu Tamagawa† Noriaki Izumi‡ Takahira Yamaguchi†
†慶應義塾大学
‡独立行政法人 産業技術総合研究所
†Keio University
‡National Institute of Advanced Industrial Science and Technology

要旨

近年，即時更新性・語彙網羅性に優れたオンライン百科事典 Wikipedia がオントロジー構築のためのリソースとして注目を集めている. しかしながら, Wikipedia はユーザ参加型という性質上, 厳密な体系化が行われていないため, オントロジーへ直接結び付けることは難しい．そこで我々はこれまで，Wikipedia から大規模なオントロジー（Wikipedia オントロジー）を構築する手法を提案してきた．本論文では，これまで構築してきた Wikipedia オントロジーの全体像および問題点を紹介するとともに, Wikipedia オントロジーを活用した検索システムの実装について述べる.

1. はじめに

オントロジーの中でも幅広い分野の一般的知識を記述した汎用（言語）オントロジーは，現在では英語版としては WordNet，日本語版としては EDR 電子化辞書がよく知られており，自然言語処理やセマンティック Web 分野における貢献度は高い．しかし，これらのオントロジーは膨大な時間とコストをかけて人手で構築されているため, 固有名詞も含め, 新しい語彙定義への即時対応が難しいのが現状である.

一方，近年，ハイパーリンクやフィードを活用した半構造化情報資源が広がりをみせている．中でも即時更新性・語彙網羅性に優れたオンライン百科事典 Wikipedia がその代表例であり，オントロジー構築のためのリソースとして注目を集めている [1-2]．しかしながら，Wikipedia はユーザ参加型という性質上，厳密な体系化が行われていないため，オントロジーへ直接結び付けることは難しい．そこで我々はこれまで，Wikipedia から大規模なオントロジー（Wikipedia オントロジー）を構築する手法を提案してきた [3-5]．本論文では，これまで構築してきた Wikipedia オントロジーの全体像および問題点を紹介するとともに，Wikipedia オントロジーを活用した検索システムの実装について述べる.

2. 日本語Wikipediaオントロジー

日本語 Wikipedia オントロジーを構築するために，我々はいくつかの手法を日本語 Wikipedia（2008 年 5 月時点）に適用し，関連関係，シノニム，クラス-インスタンス関係，Is-a 関係，Infobox トリプル，プロパティ定義域の 6 種類の関係を抽出した.

[3]では，Wikipedia マイニングの手法を用いて関連関係を抽出した．また，292,036 のリダイレクトリンクを利用してシノニムの定義を行った. [4]では, 約 8,300 の一覧記事に対してスクレイピングを行い, クラス-インスタンス関係を抽出した．また，カテゴリーツリーに対して 2 種類の文字列照合（後方文字列照合および前方文字列照合部除去）を適用することで， Is-a 関係を抽出した．[5]では，Infobox のテンプレート名とカテゴリ名の照合を行うことにより，[4]で提案したカテゴリーツリーに対する文字列照合では抽出できなかった Is-a 関係を抽出した．また，Infobox を持つ記事名，Infobox の項目，Infobox の項目の値の関係を Infobox トリプルとして抽出した. 最後に, Infobox のテンプレート名をクラス, Infobox のテンプレートの中で定義されている項目（属性）名をプロパティとみなして，プロパティ定義域の抽出を行った．表 1 にこれまで構築してきた日本語 Wikipedia オントロジーの各関係数を示す.

図 1 に日本語 Wikipedia オントロジーの全体像と問題点を示す．現状の日本語 Wikipedia オントロジーには主に五つの問題点がある．一つ目は，クラス階層の上位層において，Wikipedia のカテゴリーツリー

表 1. 日本語 Wikipedia オントロジーの各関係数

関連関係数	1,030,444	**クラス-インスタンス関係数**	331,535
シノニム数	292,036	**Infobox トリプルの数**	511,146
Is-a 関係数	9,970	**プロパティ定義域の数**	9,644

からは上位概念を抽出できていないという問題である．これについては，日本語Wikipediaオントロジーと既存の上位オントロジーや日本語WordNet[1]を統合することにより解決することを考えている．二つ目は，クラス階層における中～低位層において，著名なインスタンスを持つクラスしか抽出できていないという問題である．これについては，目次見出し等に着目した新たなIs-a関係構築手法を検討している．三つ目は，クラス階層がハイブランチ構造になっているという問題である．Wikipediaのカテゴリには「日本の～」や「～出身の人物」など，過剰なカテゴリ分類が行われているため，カテゴリツリーからクラス階層を構築した場合に，ハイブランチ構造になってしまう．これについては，「日本の」や「～出身」など，クラス名の一部分を属性として定義することによって，ハイブランチ構造を解消することを検討している．四つ目は，インスタンスが定義されていないクラスがあるという問題である．現在，一覧記事名から獲得したクラスについては，インスタンスが定義されているが，それ以外のカテゴリツリーから獲得したクラスにはインスタンスが定義されていないものがある．これについては，インスタンスが定義されていないクラスについて，クラスの元になったカテゴリに属する記事の中からクラス-インスタンス関係となる記事を見つける方法を検討している．五つ目は，一覧記事名から獲得したクラスがカテゴリーツリーから構築したクラス階層と融合されていないという問題である．これについては，一覧記事名から獲得したクラス名とカテゴリーツリーから獲得したクラス名を文字列照合により結合する方法を検討している．

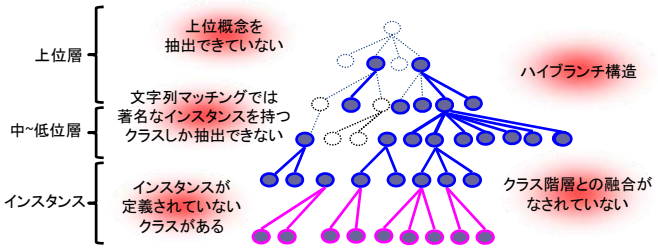

図 1. 日本語 Wikipedia オントロジーの全体像と問題点

3. 日本語Wikipediaオントロジー検索システムの実装

本研究では，日本語 Wikipedia オントロジーを検索するためのシステムを実装した．本章では，システム構成，サーバーサイドプログラム，検索インターフェースについて述べる．

3.1. システム構成

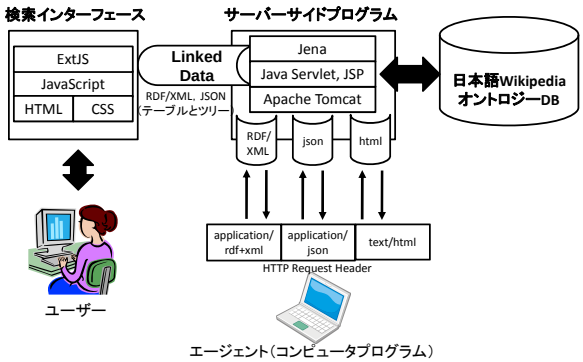

図 2. 日本語 Wikipedia オントロジー検索システムの構成

図 2 に日本語 Wikipedia オントロジー検索システムの構成を示す．日本語 Wikipedia オントロジー検

[1] http://nlpwww.nict.go.jp/wn-ja/

索システムは，日本語 Wikipedia オントロジーデータベース (DB)，サーバーサイドプログラム，検索インターフェースの三つから構成される．

日本語Wikipediaオントロジー DBは，Jena Semantic Web Framework for Java（以下，Jena）[2]のAPIを用いてあらかじめOWL (Web Ontology Language)形式の日本語Wikipediaオントロジーおよびインスタンスを MySQLに格納している．（現在，Is-a関係およびクラス-インスタンス関係のみを格納している）

サーバーサイドプログラムは，Java Servlet，JSP，Jenaを用いて実装し， Apache Tomcat上で動作している．サーバーサイドプログラムは，クラスおよびインスタンスを識別するURIにアクセスがあった場合に，エージェント向けのLinked Data[3]を生成し，返す機能を実装している．

検索インターフェースは，ExtJSライブラリ[4]を用いて実装されている． 検索インターフェースでは，はじめにユーザが検索キーワードを入力すると，キーワードに関連する日本語Wikipediaオントロジー内のクラスまたはインスタンスを識別するURIに変換する．サーバーサイドプログラムでは，検索インターフェースから問い合わせのあったクラスまたはインスタンスに関連するステートメントを日本語 Wikipediaオントロジー DBに問い合わせて，RDF/XMLデータおよびJSONデータを返す．検索インターフェースは，サーバーサイドプログラムから返されたJSONデータを元に，テーブルやツリーの構造で Wikipediaオントロジーのクラスおよびインスタンスにおけるステートメントや階層関係を表示する．また，ソースコードとしてRDF/XMLデータを表示することもできる．

以下では，サーバーサイドプログラムと検索インターフェースの詳細について説明する．

3.2. サーバーサイドプログラム

サーバーサイドプログラムは以下の仕様に従って, 日本語 Wikipedia オントロジー内のクラスまたはインスタンスに関する RDF/XML データ，JSON データ （テーブル用とツリー用），HTML ページを返すように実装した．クラスを表す URI にアクセスすると，クラスのラベル，Wikipedia カテゴリページへのリンク （foaf:page プロパティの値），対象クラスからルートクラスまでのパス，対象クラスに属するインスタンスを含むデータが取得できる．インスタンスを表す URI にアクセスすると，インスタンスのラベル，Wikipedia 記事へのリンク（foaf:page プロパティの値），対象インスタンスが属するクラス，対象インスタンスが属するクラスからルートクラスまでのパスを含むデータが取得できる．クラスまたはインスタンスを表す URI にアクセスすると，HTTP Request Header の Accept の値によって，ブラウザからのアクセスかエージェント（コンピュータプログラム）からのアクセスかを識別し，人間向けの HTML ページまたはエージェント向けの RDF/XML および JSON データを返すようにしている．

- クラスまたはインスタンスを表す URI:
  http://(ホスト名)/wikipedia_ontology/(class|instance)/(クラス名|インスタンス名)
- RDF/XML データを返す URL
  http://(ホスト名)/wikipedia_ontology/(class|instance)/data/(クラス名|インスタンス名).rdf
- JSON(テーブル用またはツリー用)データを返す URL
  http://(ホスト名)/wikipedia_ontology/(class|instance)/json_(table|tree)/(クラス名|インスタンス名).json
- HTML ページを返す URL
  http://(ホスト名)/wikipedia_ontology/(class|instance)/page/(クラス名|インスタンス名).html

本システムでは，上記 URI の(class|instance)の後に rdfs_inference を追加することで，RDFS 推論ルールにより導き出されたステートメントの集合に対して，クラスおよびインスタンスを取得することができるようになっている．例えば，現状の日本語 Wikipedia オントロジーでは「アナウンサー」クラスには直接インスタンスが定義されていない．しかし,「アナウンサー」クラスの下位クラスにはインスタンスが定義された「日本のアナウンサー」クラスなどが定義されているため，「http://(ホスト

[2] http://jena.sourceforge.net/
[3] http://linkeddata.org/
[4] http://extjs.co.jp/

名)/wikipedia_ontology/class/rdfs_inference/data/日本のアナウンサー.rdf」にアクセスすることで，「アナウンサー」クラスのインスタンスを取得することが可能となる．

3.3. 検索インターフェース

図 3 に検索インターフェースのスクリーンショットを示す. 図 3 上部の検索窓にキーワードを入力し，検索オプションとしてクラスまたはインスタンスを選択し，検索ボタンを押すことによって，検索結果が画面上に表示される．「推論モデルの利用」にチェックを入れることによって，RDFS 推論ルールにより導き出されたステートメントの集合に対して検索を行うことができる．図 3 左側にはクラス階層およびインスタンスが表示される．図 3 中央には，クラスおよびインスタンスに関連するステートメントのリストが表示される．また，複数のクラス名をスペース区切りで検索窓に入力することにより，指定した複数のタイプを持つインスタンスを検索することも可能となっている．

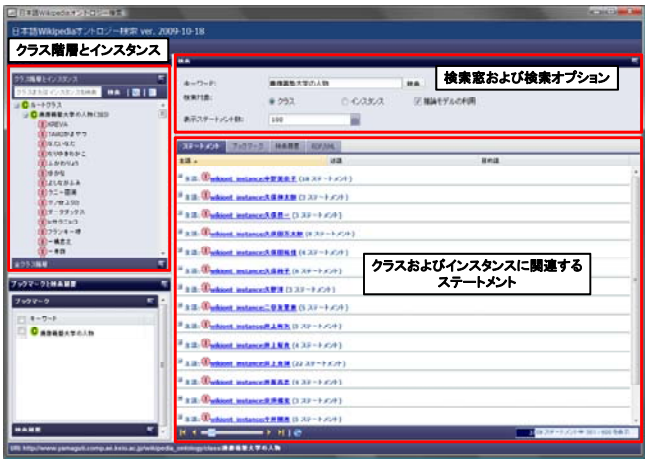

図 3. 検索インターフェースのスクリーンショット

4. まとめと今後の課題

本論文では, 日本語Wikipediaオントロジーの全体像および問題点を紹介するとともに, 日本語Wikipediaオントロジーを活用した検索システムの実装について述べた．今後の課題として，2 章で述べた日本語Wikipediaオントロジーの問題点を解決していく予定である．また，RDFクエリー言語SPARQL[5]などを用いてInfoboxトリプルの検索なども可能にすることにより，Wikipediaコンテンツの分析ができるように検索システムを改良していきたいと考えている．なお，日本語Wikipediaオントロジーおよび検索システムは，SourceForge.jp[6]にて公開する予定である．

参考文献

[1] Sören Auer,Christian Bizer,Georgi Kobilarov,Jens Lehmann1, Richard Cyganiak,Zachary Ives: DBpedia: A Nucleus for a Web of Open Data，idelberg , ISWC/ASWC 2007，LNCS 4825, 2007, pp.722-735

[2] F. M. Suchanek, G. Kasneci, and G. Weikum.: Yago: A Core of Semantic Knowledge. In: Proc. of the 16th Int. Conference on WWW, ACM , 2007，697-706

[3] 手島 拓也, 森田 武史, 和泉 憲明, 山口 高平, "日本語 Wikipedia マイニングと Folksonomy タグに基づく領域オントロジー構築支援", 第 21 回人工知能学会全国大会，2007，1D2-5

[4] 桜井 慎弥, 手島 拓也, 石川 雅之, 森田 武史, 和泉 憲明, 山口 高平, "汎用オントロジー構築における日本語 Wikipedia の適用可能性", 人工知能学会 第 18 回セマンティック Web とオントロジー研究会, 2008，SIG-SWO-A801-06

[5] 桜井 慎弥，手島拓也，森田 武史，和泉 憲明，山口 高平，"Wikipedia オントロジーに基づくドメインオントロジー構築支援環境の実現と評価", 第 23 回人工知能学会全国大会，2009 ，2G1-NFC5-1

[5] http://www.w3.org/TR/rdf-sparql-query/
[6] http://wikipedia-ont.sourceforge.jp/