1N-6

Wikipediaからセマンティック Webオントロジーの半自動構築 ～シラバス推論・検索への応用と評価～

立川 正和[*1]　松原 直樹[*2]　若木 利子[*2]

芝浦工業大学大学院 工学研究科[*1]　芝浦工業大学 システム工学部[*2]

1 はじめに

近年，Web ポータルサイトや e-Learning 等の分野において、オントロジーを利用した情報検索や自然言語処理のニーズが高まり，有効性が示されている．しかしオントロジー構築のための用語分類や精査はその分野の人手による作業のため，膨大な時間とコストを要する．ところでセマンティック Web 技術の進展より，2004 年に Web 上にオントロジーを表現するための OWL (Web Ontology Language) が W3C により勧告された．他方，今日，Web 上で Wikipedia[8] や WordNet[9] などの様々な辞書や事典が公開されている．特に Wikipedia は，Wiki という Web ベースの CMS (Contents Management System) を用いて構築されるフリーの百科事典であるため，各記事は Wiki で決められたテンプレートに基づいて登録，かつ記事 (概念) どうしがハイパーリンクで参照可能で，コンテンツが Wiki の CMS で管理されている．本研究では，このような Wiki の CMS に着目し，Wikipedia のコンテンツから，OWL 言語 [1] で表現されたオントロジーの半自動構築を試みた．さらに, 本学学生向けの履修支援を目的としたシラバス推論・検索システム [3] を開発，シラバスには記述されていない学術分野のオントロジーを Wikipedia のコンテンツより半自動生成し，それとシラバスから構築したシラバス・オントロジーと統合して，統合オントロジーのシラバス推論・検索上における有用性を確認した．

2 WikipediaとOWLオントロジ

Wikipedia のコンテンツ編集用テンプレートとして, 記事の内容 (項目名, 本文, 関連項目へのリンク及び所属カテゴリ) が書かれる項目のページと, その記事の親カテゴリを登録するカテゴリページがある．(図 1)

オントロジーはライトウェイト (Light-weight) オントロジーとヘビィウェイト (Heavy-weight) オントロジーの 2 種類に分類される [4]．前者はウェブドキュメントのオントロジー，つまり検索エンジンが持っているトピックや語彙の階層やメタデータ関連の概念階層が代表例であり，後者は概念の厳密な意味的構造を規定するものである．本研究では前者については，Wiki の CMS で構築されたカテゴリ階層や記事 (項目) 間の関連性を対象とし，後者はシソーラスと想定する．シソーラスは同義語・多義語, 上位語/下位語の情報を持つ．Wikipedia コンテンツから自動構築するオントロジーのイメージ図を図 2 に示す．なおシソーラスの構築は，現在，手動入力である．本研究で開発したオントロジー構築ツール (図 3) は，ダウンロードした Wikipedia の HTML 文書の (1)～(4)，(a)～(c) の各内容 (図 1 参照) について，表 1 に示す変換規則に従い，OWL DL で表現されたオントロジーを半自動生成する．Wikipedia の各カテゴリは OWL クラスとして表し，カテゴリの階層を subClassOf を用いてクラス階層として構築する (表 1 の (c) 参照)．Wikipedia の各項目は個体として表し, 項目の所属カテゴリとの関係を個体とクラスとする．プロパティには mainName, assoLink, synonym がある．mainName は項目と項目名を関連付けるプロパティである．assoLink は項目と別の項目を関連付けるプロパティでその 2 つが関連項目であることを表す．Wikipedia で片方向へのリンクしかなかった場合でも項目同士に関連があると考え対称性が指定される．synonym は項目名と手動で入力した別名を関連付けるプロパティでその 2 つが同義語であることを表す．項目名に対して複数の別名がある場合, 全ての名前同士が互いに同義語であるため synonym を対称性及び推移性プロパティとする．

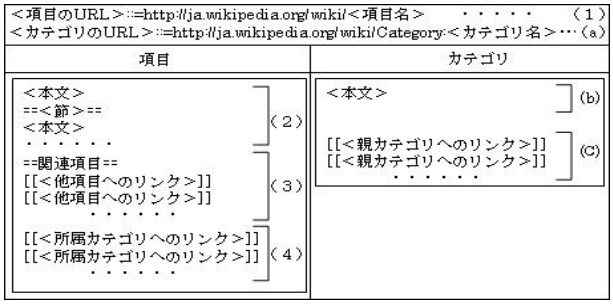

図 1: Wikipedia 文書 (入力) の仕様

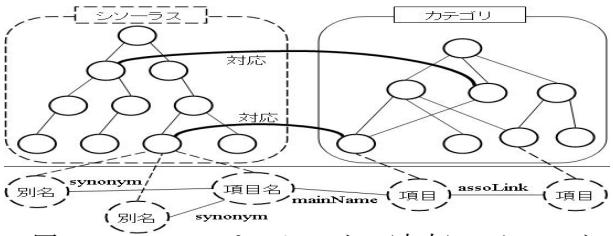

図 2: Wikipedia オントロジー (出力) のイメージ

Semi-automatic generation of Semantic Web Ontologies from Wikipedia -Application to reasoning and searching systems based on Syllabus Ontology-

[*1]M. Tachikawa,[*2]N. Matsubara,[*2]T. Wakaki

[*1]Graguate School of Engineering, Shibaura Institute of Technology

[*2]College of Systems Engineering, Shibaura Institute of Technology

表 1: Wikipedia と OWL の変換対応表

	Wikipedia編集用文書	OWL
(1)	＜項目のURL＞::= http://ja.wikipedia.org/wiki/＜項目名＞	＜項目のURL＞ mainName ＜項目名＞
(3)	==関連項目== [[＜他項目へのリンク＞]] [[＜他項目へのリンク＞]] ・・・・・・	＜項目のURL＞ assoLink ＜他項目のURL＞ assoLink ＜他項目のURL＞
(4)	[[＜所属カテゴリへのリンク＞]] [[＜所属カテゴリへのリンク＞]] ・・・・・・	＜項目のURL＞ type ＜所属カテゴリのURL＞ type ＜所属カテゴリのURL＞
	同義語の情報は 現在は手動入力	＜項目名＞ synonym ＜別名＞ synonym ＜別名＞
(a)	＜カテゴリのURL＞::= http://ja.wikipedia.org/wiki/ Category:＜カテゴリ名＞	＜カテゴリのURL＞
(d)	[[＜親カテゴリへのリンク＞]] [[＜親カテゴリへのリンク＞]] ・・・・・・	＜カテゴリのURL＞ subClassOf ＜親カテゴリのURL＞ subClassOf ＜親カテゴリのURL＞
	クラス名 ・・・ クラス	個体名 ・・・ 個体

3 シラバス推論・検索への応用

3.1 履修支援システムの概要

近年，メディア教育開発センター (NIME) では教育コンテンツに IEEE の LOM(学習オブジェクトメタデータ) を付け LOM に基づいて大学シラバス等の検索を行う NIME-glad[5] を Web 上に公開・運用している。しかし意味を加味した検索が実現されているとは言い難い。そこで我々はシラバス情報をセマンティック Web 技術を用いてオントロジーとして表現，それを用いたシラバス検索システムを開発した [3]。この研究では，NIME の LOM 付きシラバスと本学学修の手引を入力とし，シラバスの各教科の LOM を RDF に自動変換 [2] するとともに，カリキュラムや教科に関する内容オントロジーと構造オントロジーをオントロジー構築ツールの Protege[6] を用いて RDF/RDFS, OWL で表現・構築している．しかし意味を加味した推論検索に必要な学術的カテゴリ分類や同義語情報等が情報的に不足しているので，これらに関するオントロジーを本研究の方法で，Wikipedia から半自動構築を行った．

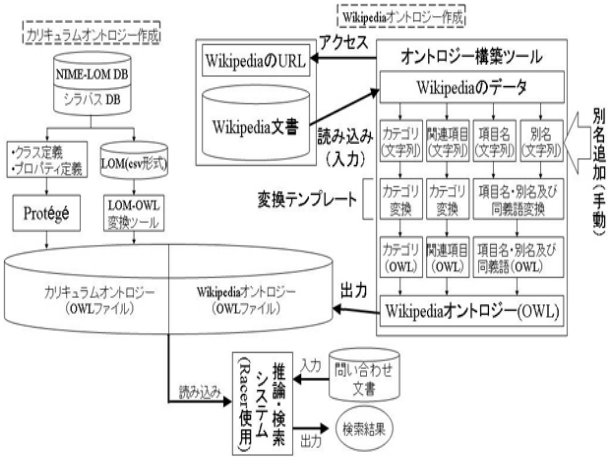

図 3: システム構成図

更に，これら 3 種類のオントロジーを統合し，それに基づくシラバス推論検索システム (図 3 参照) を実現して Wikipedia からの OWL オントロジー構築の有用性評価を試みた．

3.2 推論・検索と評価

統合オントロジーに対して, 推論･検索ツール Racer[7] を用いて本学電子情報システム学科のシラバスに関する質問評価を行った．図 4, 5 でカテゴリと同義語に関する Wikipedia オントロジーを用いた例を示す．各例の上が問い合わせで下が得られた結果である．

```
> (retrieve (?Z ?Y) (and (?X cate:プログラミング)
(?X ?Y prop:mainName)
(?Z ?Y nime-structure:contentKeyword)))
((?Z | データ構造とアルゴリズム |)
        (?Y | データ構造 |))
((?Z | オブジェクト指向言語 |)
        (?Y | オブジェクト指向 |))
((?Z | オートマトンと言語理論 |)
        (?Y |word:コンパイラ |))
 ……
```

図 4: カテゴリオントロジーを用いた件

```
> (retrieve (?Y ?X) (and (word:計算機 ?X
prop:synonym) (?Y ?X nime-structure
:contentKeyword)))
(((?Y 情報処理 I) (?X コンピュータ))
((?Y 計算機アーキテクチュア) (?X コンピュータ))
((?Y 電子情報実験 II) (?X コンピュータ)))
```

図 5: 同義語オントロジーを用いた件

図 4 の例ではカテゴリ階層を用いて関連するキーワードを持つ科目が検索結果として得られ, 図 5 の例では同義語を用いて検索キーワードそのものがシラバスに無くても推論により同義語をキーワードに持つ科目が検索結果として得られている．このように, NIME-glad 等の既存の検索システムでは得ることができない科目もオントロジーを用いた推論により検索が可能になった．

4 おわりに

Wikipedia から構築したオントロジーを用いることで, カテゴリの上位や同義語を利用した推論・検索結果を得ることができた．しかし, 現在はシソーラスや, 同義語の自動構築は行っていない．Wikipedia からのシソーラス群の自動構築が今後の研究課題である．

参考文献

[1] G.Antoniou, F.v.Harmelen: A SemanticWeb Primer, The MIT Press, 2004.
[2] S. Sawai, T. Wakaki, M. Tachikawa, and S. Kunifuji: Multimedia Portals based on Semantic Web Technology for GENES Communities - The Implementation of LOM-RDF Conversion Tools - Proceeding of SWAMM, WWW2006 Workshop, 2006.
[3] 松原 直樹, 立川 正和, 若木 利子: シラバス・オントロジー構築に基づく履修支援システム, 2007 年 3 月 電子情報通信学会東京支部学生会研究発表会発表予定.
[4] 溝口 理一郎: オントロジー工学, オーム社, 2005.
[5] NIME: http://nime-glad.nime.ac.jp/
[6] Protege: http://protege.stanford.edu/
[7] RacerPro: http://www.racer-systems.com/
[8] Wikipedia 日本語版: http://ja.wikipedia.org/
[9] WordNet: http://wordnet.princeton.edu/