社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2008－DD－69（17）
2008 / 11 / 28

Webリソースからのオントロジ自動構築法の基礎的検討

村田 智紘† 秋元 良仁†‡ 亀山 渉†*

†早稲田大学大学院国際情報通信研究科 〒367-0035 埼玉県本庄市西富田大久保山1011

E-mail: †tomo-m@moegi.waseda.jp, ‡ryoji@fuji.waseda.jp, *wataru@waseda.jp

あらまし 近年，Web上の多くのリソースがデータベースとして組織化・公開され，様々な用途で活用されつつある．しかしながら，各々のデータベースは異なるメタデータスキーマで構成されるため，横断的かつ複合的なデータの利用は困難な状況にある．異なるスキーマ間での情報交換を実現するためには，スキーマを構成する要素間の意味的な類似性を発見し，それを活用することが必要不可欠となる．本稿では，概念間の関係を記述する一方式であるシソーラスをベースとする英語辞書WordNetとWeb上の大規模百科事典であるWikipediaを相補的に組み合わせ，スキーマ間の意味的類似性をオントロジとして導出する手法について基礎的な検討を行う．

キーワード シソーラス，オントロジ，WordNet，Wikipedia

A Basic Consideration on Automatic Construction Method of Ontology from Web Resource

Tomohiro MURATA† Ryoji AKIMOTO†‡ and Wataru KAMEYAMA†*

†Graduate School of Global Information and Telecommunication Studies, Waseda University

1011 Okuboyama, Nishitomita, Honjo-shi, Saitama, 367-0035 Japan

E-mail: †tomo-m@moegi.waseda.jp, ‡ryoji@fuji.waseda.jp *wataru@waseda.jp

Abstract Recently, web resource has increased. And, it is being organized and used by several methodologies. In such a situation, a technology is necessary to extract useful information according to requirements from various web resources by using metadata. However because various metadata schema has been defined, it is difficult to exchange different metadata. In this paper, we propose the concept of ontology construction methodology applied metadata exchanging that is combined thesaurus named WordNet and online web encyclopedia named Wikipedia.

Keyword Thesaurus, Ontology, WordNet, Wikipedia

1. はじめに

情報技術の急速な進展に伴い，人類によって創出される情報量は爆発的に増加している．例えば，2008年10月現在，レスポンスを返すWebサーバは約1億8200万に上っており[1]，各サーバが複数のWebページを管理生成しているとすれば，世界中で数億から数百億のWebリソースが存在することになる．

また，Web上に存在する様々なリソースを組織化し，公開することで利用に供する試みも盛んに行われている．例えば，世界で発行される科学技術雑誌に限ってみても，約18万の逐次刊行物のうち，約3万4500誌がオンライン，約6500誌がCD-ROMによる発行であり，重複を除いてみても約3万7500誌が電子的に出版されている[2]．

このような状況において，膨大な情報の中からユーザにとって有益な情報を探し出すには，組織化された情報群を横断的かつ複合的に利用する技術が必要となる．組織化された情報群を横断利用するためには，メタデータ[3]を用いた情報の統合整理が必要となる．従来よりメタデータはその重要性が認識されており，メタデータを用いた様々な活用手法が検討されている．しかしながら，各々のメタデータは予め定義されたメタデータスキーマに基づいて構成されるため，横断的かつ複合的なデータの利用は困難な状況にある．異なるスキーマ間での情報交換を実現するためには，スキーマを構成する要素間の意味的な類似性を発見し，それを活用することが必要不可欠となる．

本稿では，要素間の概念関係を記述する一方式であるシソーラスであるWordNet[4]と，Web上のオンライン百科事典であるWikipedia[5]を相補的に組み合わせ，スキーマ間の意味的類似性をオントロジとして導出する手法を検討する．

以下，2章で関連研究について述べる．3章でスキーマ間における概念の類似性発見方式の提案を行い，4章でシステム構成を示す．5章で実験の手法を述べ，6章でまとめと今後の課題について述べる．

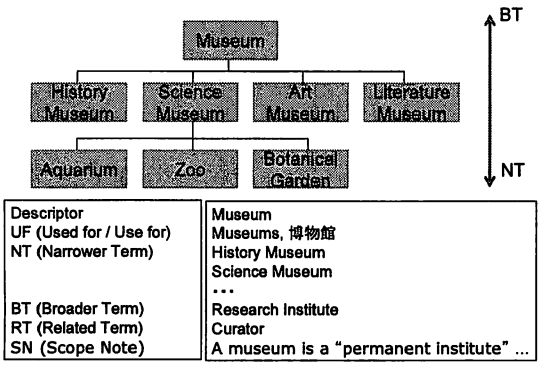


図 1　シソーラスの例

```
<?xml version="1.0" encoding="UTF-8"?>
<noun xml:id="5455460">
 <gloss>
 the person or thing chosen or selected; "he was my pick for mayor"
 </gloss>
 <word-form>choice</word-form>
 <word-form>pick</word-form>
 <word-form>selection</word-form>
 <hypernym part-of-speech="noun" target="5453619"/>
 <frames part-of-speech="verb" target="653781"/>
 <hyponym part-of-speech="noun" target="5455670"/>
 <hyponym part-of-speech="noun" target="5455968"/>
 <hyponym part-of-speech="noun" target="5456920"/>
</noun>
```

図 2　WordNet を用いた XML 記述例

2. 関連研究

2.1. シソーラスによる概念関係の記述

概念間の関係構造を記述する方式の一つにシソーラスがある．シソーラスは，1986 年に ISO2788 として標準化された語彙体系辞書のことであり，語彙を等価関係，階層関係，関連等に分類し，体系的にデータベース化しておくことで，全文検索における表記揺れの吸収や類語表現等に利用される．

シソーラスの例を図 1 に示す．図中の Descriptor とは概念を象徴する語を指す．Descriptor に対し，同義語として UF（Used for / Use for），広義語として BT（Broader Term），狭義語として NT（Narrower Term），関連語として RT（Related Term），補完的な記述として SN（Scope Note）等が定義される．各語を定義に沿って配置することで，語彙の意味的な体系化がなされる．

WordNet[4]は，英語の大規模概念辞書であり，シソーラスに見られる関係を用いて構築される．一般的に語彙は品詞別に文法的な取扱いが異なることから，品詞別に語彙の関連を定義する仕様となっている．

WordNet では，まず語彙を synset と呼ばれる同義語のカテゴリに分類する．synset は約 15 万語あり，品詞別に上位語（hypernym），下位語（hyponym），同族語（包含関係により holonym と meronym に区別される）等の関係が定義される．

```
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:skos="http://www.w3.org/2004/02/skos/core#">
 <skos:Concept
  rdf:about="http://www.example.com/concepts#mammals">
  <skos:prefLabel>mammals</skos:prefLabel>
  <skos:broader
   rdf:resource="http://www.example.com/concepts#animals"/>
 </skos:Concept>
 <skos:Concept
  rdf:about="http://www.example.com/concepts#animals">
  <skos:prefLabel>animals</skos:prefLabel>
  <skos:narrower
   rdf:resource="http://www.example.com/concepts#mammals"/>
 </skos:Concept>
</rdf:RDF>
```

図 3　SKOS の RDF 記述例

WordNet は文書の自動解析に応用されることを想定して作られた概念辞書であり，近年では，Web サーバ・フレームワークに組み込んで検索語彙の揺れを吸収するアプリケーションや，WordNet から XML 形式で概念関係を抽出し，それを利用することで検索結果として意味的に近い文書を返すセマンティック XML アプリケーション等に応用する試みも行われている*[6]*．図 2 に WordNet を用いた XML 記述例を示す．

WordNet は英語の概念辞書であるため，日本語処理を目的としたアプリケーションへの応用を考える際には，語の翻訳作業が必要となる．これまでにも，和英辞書 EDICT を用いて機械翻訳する試み[7]等がなされている．

SKOS(Simple Knowledge Organization System)[8]は，シソーラス表現に RDF を用いて概念スキームを表現するためのモデルである．シソーラスを RDF ボキャブラリで表現できるよう工夫がなされており，要件として図書館や文書館等で使用されている既存のシソーラス体系を簡易に Web 上で利用できるように，RDF に変換することが想定されている．

2008 年 10 月現在，SKOS の仕様ステータスは W3C の Working Draft の Last Call 中であり，現在も議論が行われている．

図 3 に SKOS の RDF 記述例を示す．SKOS ではシソーラスの Descriptor に skos:Concept, BT に skos:broader, NT に skos:narrower 等，シソーラスとほぼ 1 対 1 の関係でボキャブラリが定義されている．

2.2. Wikipedia を用いた概念形成

Wikipedia[5]は共同文書編集ソフトウェアである Wiki を用いて様々な人が共同で構築することができる大規模百科事典である．2008 年 10 月現在，英語 260 万項目，日本語 50 万項目を超える様々な分野の記事を

カバーしている.市販の百科事典は記事数が約 10 万項目であることから比較しても膨大な記事数をカバーしていることが分かる. Wikipedia は記事（概念）同士がハイパーリンクで互いに参照されており，また，1 つの記事には 1 つの URL が割り当てられているため，インターネット上において記事を一意に参照できるという特徴を持つ.

Wikipedia は大量の記事とハイパーリンクによる語彙間構造を持つことから，Wikipedia を用いて概念の形成や多言語翻訳に応用する研究がある. 伊藤らはリンクに見られる共起性を解析し，シソーラスにおける関連語を自動抽出し，シソーラス辞書を構築する手法を提案している[9]. また，Erdmanm らは言語間リンクのみならず，Wikipedia が持つリダイレクトページやアンカーテキストを利用した翻訳辞書システムを提案している[10].

2.3. オントロジによる概念関係の記述

オントロジとは，概念と概念間の関係性を定義する方式の一つである. 近年では次世代 Web 技術として注目されるセマンティック Web の核技術の一つとして，その構築手法[11]，異なるオントロジの統合手法[12]，辞書や概念形成を目的とした利用手法等[7][13]の研究が行われている.

Web Ontology Language（OWL）[14]は W3C が提唱する Web オントロジ記述言語であり，RDF の語彙拡張として，RDF のトリプル集合として表現される. Web 上のリソースを OWL で記述することで機械的に Web リソース交換を行うことが想定されている.

3. スキーマ間の概念類似性発見方式の提案

本研究では，スキーマの項目間に見られる類似性を発見するため，シソーラスの持つ概念関係記述方式と Wikipedia の持つ広範な記事量を利用することを考える. Wikipedia は階層構造を持ち得ず，WordNet は日本語のデータを持ちえないといった弱点がそれぞれ存在する.

そこで，両者の弱点を相互補完的に行いシソーラスの構築に役立てる方式を提案する.

WordNet には，概念に対して一意に ID が付与されている. そこで，(1)WordNet から概念 ID 及び関連する概念関係性を抽出し，(2)概念 ID は Wikipedia の 1 記事(日本語ページ URL)と対応付ける. また，(3)関連する概念関係性は SKOS に変換する. 更に，(4)Wikipedia の記事と対応付けた概念 ID と変換した SKOS のマージを行う.

4. システム構成

提案したシステムのプロトタイプの実装を行った.

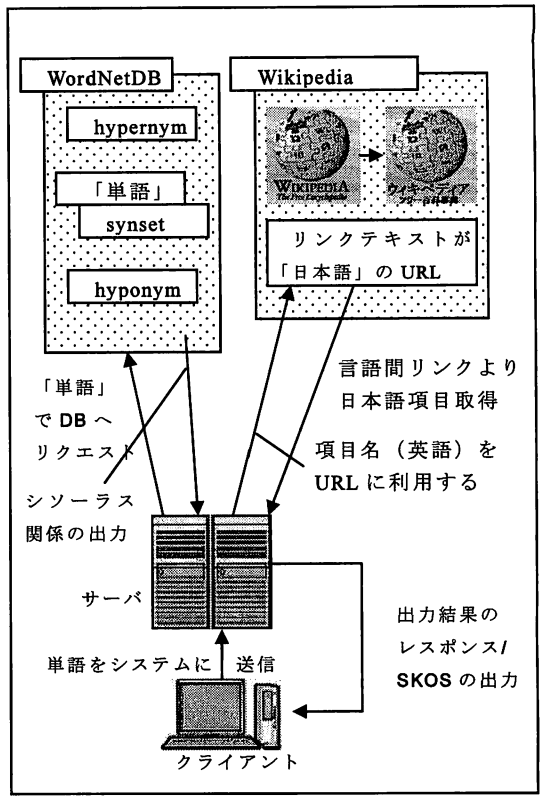


図 4　システム構成図

それぞれのモジュールは Perl を用いて構築した. システム構成図を図 4 に示す. まず，クライアントからシステムに単語を入力し，WordNet のデータベースファイルより入力した単語，その単語の hypernym，hyponym をそれぞれ抜き出しシソーラス関係を出力する. その後得られた項目名を英語版 Wikipedia のベースとなる以下の URL

・http://en.wikipedia.org/wiki/

の後部に付け加え，英語版 Wikipedia の記事の URL とし，WWW::Mechanize モジュールを用い記事データを取得する. 取得したデータにタグ解析をかけ，リンクテキストに「日本語」を含むリンク先の URL を取得し，日本語版 Wikipedia の記事の URL とする. また，Wikipedia は URL に記事の項目名がそのまま割り当てられている為，その URL に対して URL デコードを行い,日本語の項目名を獲得した. これにより英語版，日本語版 Wikipedia の URL とその対応関係が得られた. 上下関係のある他項目に対しても同様の処理を行い，

```
music%1:04:00:: 00543233 3 2
music%1:04:02:: 01162529 5 0
music%1:09:00:: 05718556 2 12
music%1:09:01:: 05718935 4 0
music%1:10:00:: 07020895 1 51
music_box%1:06:00:: 03801353 1 0
music_critic%1:18:00:: 10339856 1 0
music_department%1:14:00:: 08117540 1 0
music_director%1:18:00:: 09952539 1 0
music_genre%1:10:00:: 07071942 1 0
music_hall%1:06:00:: 03801533 1 1
music_hall%1:10:00:: 07020423 2 0
music_lesson%1:04:00:: 00889760 1 0
music_lover%1:18:00:: 09951616 1 1
```

図 5　WordNet インデックスファイル

```
00543233 04 n 01 music 0 058 @ 00407535 ...
00544441 04 n 03 bell_ringing 0 carillon...
00544605 04 n 01 change_ringing 0 001 @ ...
00544731 04 n 01 instrumental_music 0 00...
00544842 04 n 01 intonation 2 004 @ 0054...
00545059 04 n 01 percussion 1 003 @ 0054...
00545194 04 n 01 drumming 0 002 @ 005450...
00545344 04 n 01 vocal_music 0 003 @ 005...
00545501 04 n 02 singing 0 vocalizing 0 ...
00546070 04 n 02 a_cappella_singing 0 a_...
00546216 04 n 01 bel_canto 0 001 @ 00545...
00546299 04 n 01 coloratura 0 001 @ 0054...
00546389 04 n 02 song 0 strain 1 006 @ 0...
00546613 04 n 01 carol 0 002 @ 00546389 ...
```

図 6　WordNet データベースファイル

各項目を置き換える．得られた Wikipedia の記事については skos:note とし，参照先とする．これにより WordNet の項目を日本語と置き換えることが可能となり，また Wikipedia とも関連付けられることによりシソーラスができる．システムは以下のサーバおよびモジュールから構成される．

1）WordNet データベースファイル

WordNet の Web サイトよりダウンロードしたものである．独自形式のテキストデータとなっている．本稿では最新版の WordNet3.0 を用いた．ファイル構造はインデックスファイルとして，index.sense が存在する．これは WordNet の単語が全てアルファベット順に格納されており，内部は図 5 のように品詞の種類と固有 ID により別個のデータファイルと関連付けられている．

またデータファイルとして data.noun, data.verb, data.adj, data.adv が存在し，それぞれ名詞，動詞，形容詞，副詞ごとに分かれている．内部は図 6 に示すようにインデックスに記述されている固有 ID で単語の対応付けを行い，他の語句との synset や hypernym,hyponym が ID と属性により関連付けされている．また，語句の簡単な意味も記述されている．

2）サーバ

プログラムを動かすための Perl，Web サーバ．プロトタイプの実装は Windows 上で行った為，それぞれ Active Perl，AN HTTPD を用いた．

3）Perl モジュール

HTTP 通信などを行うために必要なモジュール類である．Web クローラとして WWW::Mechanize を用いた．Wikipedia サーバからクロールを行い，言語間リンクを取得する．

4）クライアント

クライアントは Web ブラウザがインストールされていれば Web 経由でデータを検索する事ができる．

出力画面は図 7 に示す．ヘッダ部分にリクエストした単語とその解説，その下に日本語版 Wikipedia へのリンク(Wikipedia:Ja)となっている．

リンクの下には hyponym, hypernym をそれぞれ表記し，リンクをクリックする事によりその単語のページを参照することが可能となっている．

5. 実験手法

本提案手法の評価手法について述べる．

予め正解データとして単語レベルで対応が取れる文書構造を持つ日英のデータを用意する．初めにシステムが生成した SKOS を用いて日本語正解データのシソーラス構造を調べる．次に，WordNet を用いて英語の正解データのシソーラス構造を調べる．日英双方のシソーラス構造を比較し，それぞれの対応箇所の確認を行う．この処理を一定量の語彙に対して行い，統計的にシステムが生成した SKOS の評価とする．

6. 考察

本稿において WordNet と Wikipedia の対応付け，日本語との対応を行いプロトタイプの作成を試みた．プロトタイプでは多義語を判別する構造を持たず，たとえば music を検索した場合 music#n#1 ，music#n#2 ，music#n#3，music#n#4，music#n#5 と

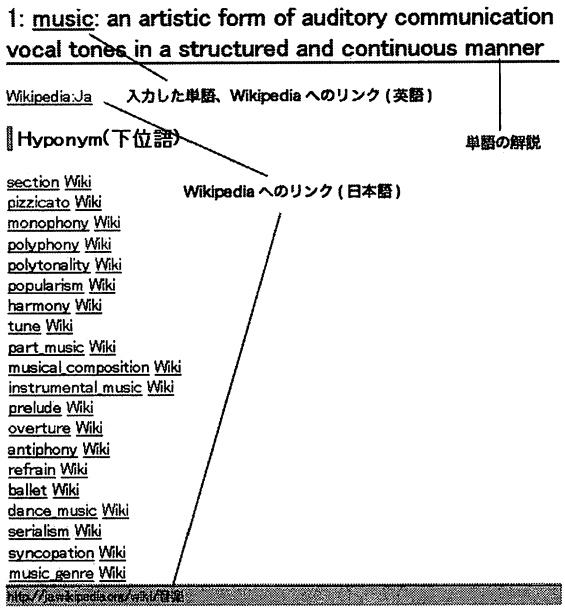


図 7 システム画面

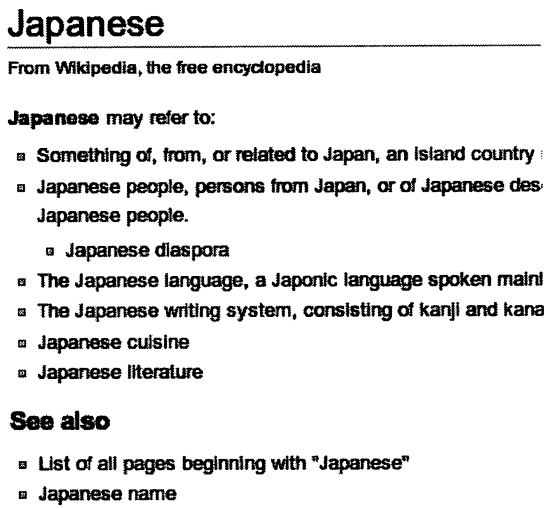
5 つの項目が得られるが，本システムで訳した場合全て「音楽」であった．基本的に一番目が最も用いられる用例であり，それ以降も似た語源であるものが多い傾向にあったが，まったく意味の違う多義語も存在した．

例として「soul」という単語では，システムでの日本語訳は「霊魂」であったが，他の多義語として「ソウルミュージック」などが存在する．

また，単語そのものでは成り立たないが成句でのみ成り立つ単語などが含まれた．例として「music#n#5」の項目では medicine (punishment for one's actions) "you have to face the music"; "take your medicine"という解説が存在し，music 単体のみではその意味を持ち得ないが「face the music」という成句にすることにより「punishment for one's actions」という意味を持ちえるという事例があった．

また，曖昧さ回避（Disambiguation）のページにヒットした場合，日本語を取得することが出来なかった．例として「Japanese」の意味としては「Japanese language:the language (usually considered to be Altaic) spoken by the Japanese」「japanese people:a native or inhabitant of Japan:」などが存在するが Wikipedia では曖昧さ回避のリンクによって記事に項目が羅列されており，日本語との対応関係のリンクを得ることができなかった．(図 8)

また，hypernym, hyponym も全て日本語に置換する予定であったが，一つの項目に対してクロールからアクセス，解析の流れに時間かかるため，主項目のみに


Japanese

From Wikipedia, the free encyclopedia

Japanese may refer to:

- Something of, from, or related to Japan, an island country
- Japanese people, persons from Japan, or of Japanese des Japanese people.
  - Japanese diaspora
- The Japanese language, a Japonic language spoken maini
- The Japanese writing system, consisting of kanji and kana
- Japanese cuisine
- Japanese literature

See also

- List of all pages beginning with "Japanese"
- Japanese name

図 8 曖昧さ回避

対して行っている．この問題は Wikipedia の SQL ダンプデータをサーバ内部で用いるなどして対応する予定である．

7. まとめと今後の課題

本稿では，Web リソースにおけるシソーラスの構築についてプロトタイプの開発を行い，WordNet と Wikipedia を用いたスキーマ間の意味的な類似性の発見方式について検討を行った．

今後の課題は以下のとおりである．

1. 対訳データを用いたシステムの正確性の検証
2. 多言語間リンクシステムのサーバ内実装
   Wikipedia のサーバにも負荷をかけているためこれをサーバ内部での実装とする
3. 多義語の判別
   現在では全て同じ訳語で表示されてしまうため，多義語の判別を行う必要がある．

これらの課題を解決した上でアルゴリズムの改良を行い，シソーラスの精度の向上を目指すことを目的とする．

文　献

[1] October 2008 Web Server Survey
http://news.netcraft.com/ (2008-11-05)

[2] 時実 象一. “日本発行の科学技術雑誌の調査 (1) 電子ジャーナル化の状況”. 情報管理. Vol. 51, No. 8, pp. 571-579 , Oct. 2008.

[3] Dempsy et al.: “Metadata: A Current View of Practice and Issues”, J. of Documentation, Vol. 54, No. 2, pp. 145-172, Mar. 1998.

[4] WordNet
http://wordnet.princeton.edu/ (2008-11-05)
[5] Wikipedia
http://www.wikipedia.org/ (2008-11-05)
[6] Thinking XML: Querying WordNet as XML
http://www.ibm.com/developerworks/xml/library/x-think29.html
(2008-11-05)
[7] 日本語ウェブオントロジーの試み
http://www.kanzaki.com/docs/sw/jwebont.html
(2008-11-05)
[8] SKOS
http://www.w3.org/2004/02/skos/ (2008-11-05)
[9] 伊藤他: "Wikipedia のリンク共起性解析によるシソーラス辞書構築", 情処論, Vol.48, No. SIG-20, pp. 39-49, Dec. 2007.
[10] M.Erdmanm.et al.,:"Wikipedia Link Structure Analysis for Extracting Bilingual Terminology", 情処研報, Vol. 2007, No. 65, pp. 551-556, July. 2007.
[11] 内田他: "オントロジの自動構築に関する基礎的研究", SIG-SW&ONT-A301-05, June. 2003.
[12] 稲葉他: "固有名詞抽出技術を用いたオントロジー・メンテナンスツールの設計", SIG-SWO-A701-06, July. 2007.
[13] 上田他: "オントロジーエディタ Protege-OWL を使った OWL オントロジー構築", 人工知能学会誌, Vol. 21, No. 4, pp. 446-454, July. 2006.
[14] OWL
http://www.w3.org/2001/sw/WebOnt/ (2008-11-05)