法政大学学術機関リポジトリ
HOSEI UNIVERSITY REPOSITORY

PDF issue: 2025-07-01

オントロジー自動構築に関する研究

谷口, 祥太 / TANIGUCHI, Shota

(出版者 / Publisher)
法政大学大学院理工学・工学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 理工学・工学研究科編 / 法政大学大学院紀要. 理工学・工学研究科編
(巻 / Volume)
55
(開始ページ / Start Page)
1
(終了ページ / End Page)
4
(発行年 / Year)
2014-03-24
(URL)
https://doi.org/10.15002/00010483

オントロジー自動構築に関する研究

RESEARCH ON ONTOLOGY AUTOMATIC CONSTRUCTION

谷口祥太
Shota TANIGUCHI
指導教員　藤井章博

法政大学大学院工学研究科情報電子工学専攻修士課程

In this paper, we propose a method of reduce the "costs" needed to build an ontology which is a fundamental technology of semantic web. The "cost" means time and knowledge in this paper. By using its own algorithm and morphological analysis, build the ontology automatically. We focused on the manufacturing industry in this paper. Get in crawling introductory essay of each company present in the portal site of the manufacturing industry, and extracted the keywords from the text. By using the extracted keyword, we constructed the ontology. Throughout this proposal, we were able to significantly reduce the time and knowledge required to build the ontology. However, improving the accuracy of the ontology is built is the future issue.

Key Words *: semantic web, ontology, web crawler, TF-IDF, MeCab, automatic construction*

１．はじめに

近年，次世代 Web と呼ばれるセマンティック Web が注目を集めている．現在の Web は人間を情報参照の対象とし, 人間が理解可能なコンテンツを提供することを目的としているため，ソフトウェアによる理解は困難である．一方，セマンティック Web は，オントロジーに基づき，Web コンテンツにソフトウェア可読なメタデータを付与することによって, ソフトウェアが Web コンテンツを理解し，推論することを可能にしようという試みである. セマンティック Web の実現により, Web サービスの自動的な連携, Web 上の異種データの分析, アプリケーションを横断したデータ統合および再利用などが可能となることが期待されている．オントロジーはセマンティック Web の核となる技術であり，セマンティック Web 実現のためには，さまざまな領域ごとに多数のオントロジーが構築されることが必要となる．しかし現状では，既存のオントロジーの大部分が，多大なコストをかけて手動で構築されており，オントロジー構築はセマンティック Web 実現のボトルネックとなっている．

オントロジーを手動で構築することは, 膨大な時間と知識が必要であり，間違いを起こしやすく，オントロジーの保守や更新が困難という問題がある．そこで，こういった問題を解決するために, 自然言語処理と独自のアルゴリズムを併用して, 文章をもとにオントロジーを自動構築する研究に取り組んできた. これにより先に挙げた問題を解決した上で，旬なワード，実用的なワードで構成された柔軟かつ効果的なオントロジーを構築することが可能となる．本研究では特に製造業の領域を対象とする．

２．先行研究

現在, オントロジーの構築には次の方法が用いられている[1].

① オントロジーエディタを利用した手動構築

② 複数オントロジーの比較・統合

③ 語彙表現辞書からの整形

④ Web リソースからの抽出

方法①で使われるオントロジー構築ツールとして「Protégé」「OntoStudio」「OilEd」「SWOOP」「法造」などが存在する[2]．しかし，オントロジーを手動で構築するには膨大な時間とその領域に関する深い知識が必要になるという問題がある．よって大規模なオントロジーになればなるほどオントロジー構築の自動化が必須となってくる．自動構築で言えば方法③のように辞書と検索サイトを用いてオントロジーを構築するものや，Wikipedia の情報をもとにオントロジーを構築するものが存在する[3][4]．しかし柔軟かつ効果的なオントロジーを構築する手法としては方法④がふさわしいと言える．

３．提案手法

（１）システム概要とフローチャート

本研究で提案するシステムは，文章をもとにオントロジーを自動構築するものである．提案するシステムのシステム概要を図１に，フローチャートを図２に示す．

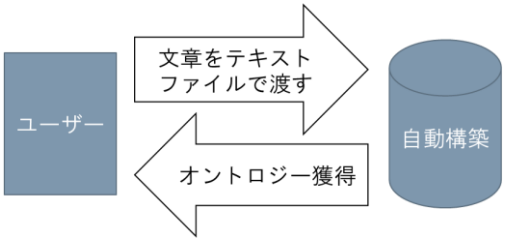

図１　システム概要

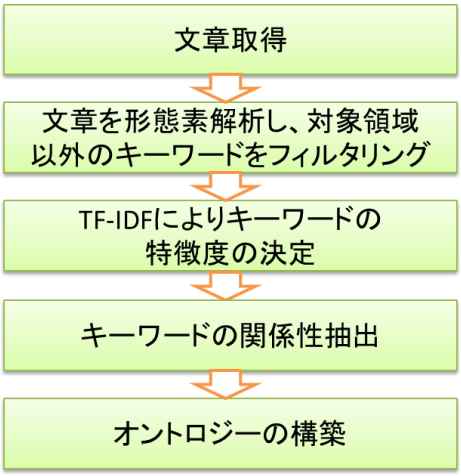

図２　フローチャート

（２）特徴度の決定

特徴度の決定をするための手法として TF-IDF がある．TF-IDF とは，ドキュメントに出現した特定のキーワードがどのくらい特徴的であるかを識別するための指標のことであり，情報検索等に使われるアルゴリズムである．ここでいう tf(term frequency)とは，あるキーワードが１つのドキュメントにどのくらい出現するかを表しており，idf(inverse document frequency)とは，あるキーワードがどのくらいの数のドキュメントに出現するかを表している．計算式は(1)，(2)，(3)のようになる．

$$w = tfidf = tf \cdot \log_2 \frac{N}{df} \tag{1}$$

$$tf = \frac{\text{キーワードの頻度}}{\text{ドキュメントで出現する総キーワード数}} \tag{2}$$

$$idf = \log_2 \frac{\text{総ドキュメント数}}{\text{ヒットページ数}} \tag{3}$$

TF-IDF を用いて得られる情報は，対象となるドキュメント，つまり一つのドキュメントにおける特定のキーワードの特徴度である．しかし，本研究で取得したい情報は全てのドキュメントにおける特定のキーワードの特徴度である．そのため，本研究では式(2)の計算をする前に変更を加える．複数存在するドキュメントを一つにまとめ，まとめたドキュメントに出現するキーワード数を「ドキュメントで出現する総キーワード数」，特定のキーワードの出現する回数を「キーワードの頻度」とする．このことにより，クローリングで取得した全てのドキュメントを対象とした特定のキーワードの特徴度の決定ができる．

（３）キーワードの関係性抽出アルゴリズム

キーワードとそのキーワードの特徴度を求めた後は，キーワード同士の関係性を調べる．そのためにまず特徴度の数値をもとにキーワードのグループ分けを行う．例としてグループ分けした表を表１に示す．また，表１の結果を用いてオントロジーを構築するために実行したプログラムの疑似コードを図３に示す．

表１　特徴度の値をもとにグループ分けしたキーワード

特徴度	キーワード	
グループ1 とても低い	組立	溶接
	試作	金型
	設計	加工
グループ2 低い	組立て	研削加工
	熱処理	修理
	試作品	デザイン
	印刷	研磨
	材料	メーカー
	塗装	
グループ3 高い	出張修理	印刷関連
	ボイラー溶接士	シーム溶接
	研磨剤	原材料
	印刷業	複合加工機
	研磨材	設計者
	静電塗装	印刷物
	半自動溶接	試作機
	簡易金型	TIG溶接
	レーザー加工機	印刷機
	放電加工機	研磨機
グループ4 とても高い	半自動アーク溶接	パーツフィーダメーカー
	スタンピングメーカー	材料費
	ロール曲げ加工	輸出加工区
	発泡材料	吹付塗装
	設計図	ﾎﾟﾝﾌﾟメーカー
	不燃材料	微細加工技術

```
FOR グループ1のリスト
  トップ要素に繋げる
ENDFOR

FOR グループ2のリスト
  FOR グループ1のリスト
    比較
    IF  部分一致したなら
      繋げる
    ELSE
      トップ要素に繋げる
  ENDFOR
ENDFOR
```

図３　疑似コード

図３の疑似コードが表すアルゴリズムと同様の処理をグループ３，４に対しても行う．このアルゴリズムでは，特徴度の値が低いキーワードほどオントロジーの上位に位置すると判断する．まずグループ１のキーワードをトップ要素に繋げる．次にグループ２のキーワード一つに注目し，そのキーワードがグループ１のキーワードに部分一致するかどうかを調べていく．部分一致した場合，注目していたキーワードはそこに繋がる．しなかった場合は注目していたキーワードはトップ要素に繋がる．この処理をグループ２のキーワードすべてに行う．その後，グループ３のキーワードに注目し，グループ２に対して同じ動作を，部分一致しなかったならばグループ１に対して同じ動作を，それでも部分一致しなかった場合はトップ要素に繋げる．グループ４に対しても同様な処理を行っていく．これにより，キーワードの関連性を調べ，オントロジーの繋がり方を求めることができた．

４．結果

図２で示したフローチャートに沿って処理を行っていく．まず，「文章取得」の部分を行った．本研究で対象としたのは製造業の領域であり，日本の製造業約 5000 社の情報を掲載しているサイトを参考にした．企業ごとに紹介文があるので，HTML タグで紹介文の位置を判定し，今回はクローリングで取得した．取得した文章はテキストファイルとして保存した．（紹介文が記載されていない企業もあるため，総ドキュメント数は 5000 よりも少なくなる．）次に，「文章を形態素解析し，対象領域以外のキーワードをフィルタリング」の部分を行った．なお，MeCab には Wikipedia のタイトルリストを辞書として登録し，その辞書を用いて形態素解析を行った．次に，「TF-IDF によりキーワードの特徴度の決定」の部分を行った．これには「４．要素技術の（１）」で紹介した手法を用いた．次に「オリジナルアルゴリズム手法」の部分を行った．これには「４．要素技術の（２）」で紹介した手法を用いた．最後に「オントロジーの構築」の部分を行った．その結果を図４，図５に示す．図４では，トップ要素である製造業に表１のグループ１，もしくはグループ２に存在する特徴度の低いキーワードが繋がっていることを確認することができる．そして加工には特徴度の高い加工関連のキーワードが繋がっていることが確認できる．また，図５は図４の溶接の部分に着目した図である．

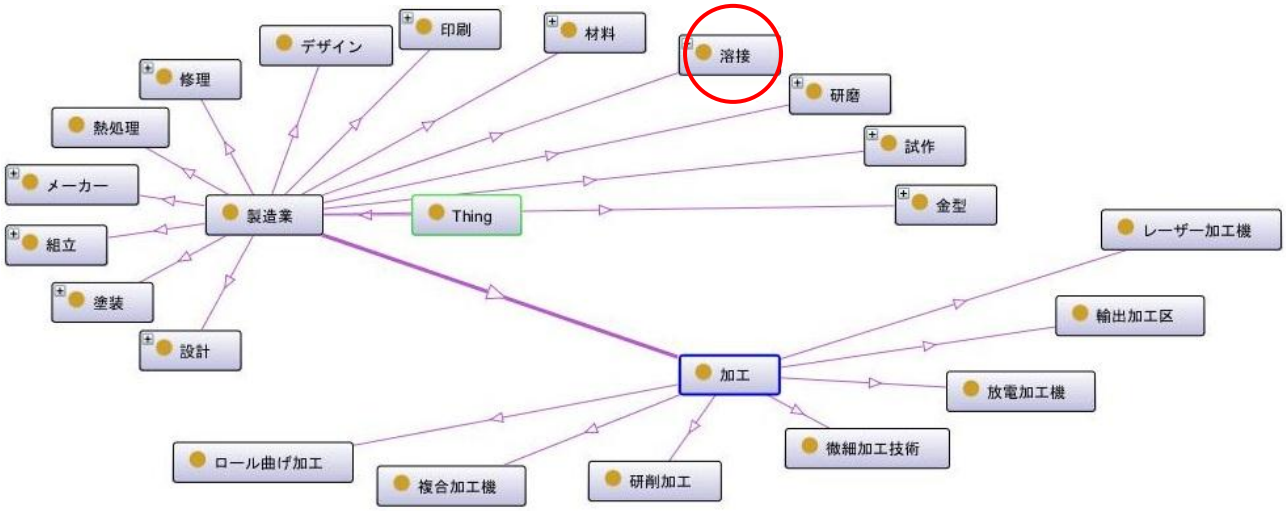

図４　構築したオントロジーグラフ

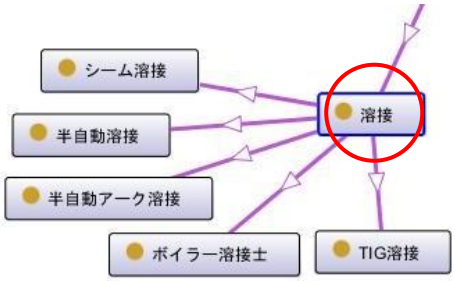

図５　溶接に関するオントロジーグラフ

５．考察

本研究によって，オントロジーを構築する上で必要となっていた知識，時間などのコストを削減することができた．しかし，精度は若干手作業よりも劣る場合が考えられる．図５では，溶接の下位概念として「半自動アーク溶接」と「半自動溶接」が存在している．本来ならば「半自動アーク溶接」は「半自動溶接」の下位概念と考えられるが，本研究で用いた手法では本来の意図とは異なり，図５のような結果になる．また，製造業以外のキーワードをフィルタリングする部分において，対象としたサイト上の企業検索用タグ一覧にあるキーワードリストを利用した．本研究ではそのキーワードリストをプログラム内に組み込んだが，キーワードリストをテキストファイルとして保存し，そのファイルを参照して利用できる形にすることにより，サービスとしての有用性が向上すると考えられる．

６．まとめ

本研究の目的であったオントロジーを構築するためのコスト削減は達成できた．しかし，精度の問題や，実際にサービスとして利用するための問題など，今後の課題は残っている．

謝辞：本研究を進めるにあたり，ご指導頂いた藤井章博准教授に感謝致します．また，日常の議論を通じて多くの知識や示唆を頂いた研究室の皆様に感謝します．

参考文献

1) 稲葉真純，飯田貴之，山崎智弘，布目光生，溝口裕美子，長野伸一，川村隆浩：オントロジーメンテナンスのための固有名詞抽出技術の検討，電子情報通信学会技術研究報告．AI，人工知能と知識処理，Vol.107，No.428，pp.83-86，2008

2) 古崎晃司，溝口理一郎：オントロジー構築ツールの現状，人工知能学会誌，Vol.20，No.6，pp.707-714，2005

3) 内田英里，　石野武志：オントロジーの自動構築に関する基礎的研究，人工知能学会，セマンティックウェブとオントロジー研究会，2003

4) 白川真澄，中山浩太郎，荒牧英治，原隆浩，西尾章治郎：Wikipedia と Web の情報を組み合わせたオントロジー構築の試み，電子情報通信学会誌．D，情報・システム，Vol.J94-D(3)，pp.525-539，2011

5) Toby Segaran, Colin Evans, Jamie Taylor：セマンティック Web プログラミング，株式会社オライリー・ジャパン，2010