

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

稲葉 真純 長野 伸一 佐々木 寛 山崎 智弘
Masumi INABA Shinichi NAGANO Hiroshi SASAKI Tomohiro YAMASAKI

溝口 祐美子 川村 隆浩
Yumiko MIZOGUCHI Takahiro KAWAMURA

株式会社 東芝 研究開発センター
Corporate Research & Development Center, Toshiba Corp.

Abstract: オントロジーを活用したアプリケーションの持続的運用において、オントロジーのメンテナンスコストの低減は重要な課題である。半自動化メンテナンスツールの構築に向けて、商品情報に関するオントロジーの洗練化や新語の獲得方法を検討する。

1. はじめに

近年、様々な分野のアプリケーションでオントロジー活用事例が見られるようになってきた。これはオントロジー研究が実用の段階へと移行する兆しといえる。従来から知識処理技術を用いるアプリケーションでは、精度を高めるために辞書が利用されてきた。軽量かつ大規模なライトウェイト・オントロジーを辞書的に用いるアプリケーションが生まれた背景には、多岐に渡る分野を一つの知識体系として表現可能であることが理由として挙げられる。オントロジーを知識体系として利用するにあたっては、正確性や豊富なデータ量が求められる。それゆえ、既存のオントロジーを他言語に変換する方法や、複数のオントロジーをマージングする手法が盛んに研究されている。オントロジー構築ツールについても、利便性の高いアプリケーションが開発されている。オントロジーの整合性を保つには人手による確認が不可欠である。従って、持続的に運用するアプリケーションにおいては、オントロジーの構築コスト、さらにメンテナンスコストを低減する必要がある。

一方、近年では EC サイトや情報共有サイトが充実し、データ集合として利用できるまでに成長している。本稿では、これらのデータ集合から半自動的にオントロジーを構築する手法について述べる。特に、ライトウェイト・オントロジーの効率的な構築方法を検討し、大

規模なオントロジーを利用したアプリケーションの構築、運用コストの低減を目指す。

2 章では、従来のオントロジー構築方法について述べる。続いて 3 章では、固有名詞抽出技術を用いた新語の獲得方法を考察し、オントロジーのメンテナンスツールへの適用を検討する。最後に、4 章で今後の課題を述べてまとめとする。

2. オントロジー構築方法

従来、オントロジーの構築には次の方法が採られてきた[1]。

- オントロジーエディタを利用した手動構築
- 複数オントロジーの比較・統合
- 既存オントロジーの変換
- 語彙表現辞書からの整形
- 電子的な技術文書からの抽出
- Web リソースからの抽出

オントロジーエディタは、人手によるオントロジー作成を支援するアプリケーションである。オントロジーの構造をグラフで表示し、概念の属性名と属性値を入力する手順が一般的である。オントロジーの整合性に留意しながら作業を進めることができる一方、ゼロからオントロジーを構築するため、作成者への負担が大きい。そこで、既存のリソースを利用してオントロジーを構築する方法が研究されている。既存のオントロジーをリソ

ースとするものには、WordNet[2]などを利用して複数のオントロジーをマージングする方法や、他言語へ変換する方法が研究されてきた。オントロジー以外にも、語彙表現辞書からオントロジー形式への整形も一つの手段である。しかし、構築可能なオントロジーは汎用的な語彙に止まり、分野に特化したオントロジーは構築しにくい。逆に、技術文書では、得られるデータが特殊な分野に限られる。一般的な新語をオントロジーに取り込む方法として、Web 上のリソースを活用する研究が活発に行われてきた。様々な分野に関して大量のデータを取得できるが、様式が自由であることから、データの抽出精度が課題である。

大規模なオントロジーを構築するためには、自動化技術が欠かせない。しかし、構築したオントロジーの整合性にも留意しなければならない。そこで、システムに任せられる処理と、人手による確認が必要な処理を分けた半自動化ツールの構築を検討する。

3. オントロジー・メンテナンスツール

3.1 オントロジー・メンテナンスツールの要件

オントロジー・メンテナンスツールの要件を以下に示す。

1. Webリソースから、商品オントロジーを構築
2. 冗長な商品名を整理
3. 定期的なメンテナンス(追加, 更新, 削除)
4. メンテナンスコストを低減(人的リソース)
5. 既存の商品オントロジーを活用

今回、我々は EC サイトや価格比較・クチコミ共有サイトなどから抽出した商品データをターゲットとして、オントロジーの構築を目指す(要件1)。これらのサイトから取得する商品データは、それぞれ一定のルールに従ってカテゴリ分けや、ネーム付けが行われている。特に商品名は、商品を表す汎用的な固有名詞と、色やサイズなどの特徴を表す固有名詞を組み合わせることが多い。このようなデータは、人間が商品を選ぶ際には有用であるが、アプリケーションで活用するには冗長である。従って、これらのデータから必要な部分を取り出し、整理する必要がある(要件2)。オントロジーを活用したアプリケーションにおいては、情報の精度を高めるためにオントロジーのメンテナンスが不可欠である。特に、商品オントロジーは、新商品が発売される度にデータを更新する必要がある(要件3)。しかし、オントロジーのメンテナンスには技術的なノウハウが必要で

ある。特別な知識を必要としないツールを構築することで、技術者でなくてもオントロジーのメンテナンスが行えるようにしたい(要件4)。既に構築した数十万件の商品オントロジーを活用することで、作業者を支援する機能を開発する(要件5)。

3.2 オントロジー・メンテナンスツールの設計

商品名を構成する固有名詞の並び順には、ある程度の規則性がある。本稿では、固有名詞の並びに規則性が見られるデータ集合を対象として、アプリケーションでの活用を目的とした商品オントロジーの構築手法を提案する。

前述の要件を満たすメンテナンスツールを設計する。メンテナンスツールのアーキテクチャを図1に示す。メンテナンスツールに Web リソースから取得した商品名を入力すると、過去の履歴を参照して、冗長な表現を整理した商品名の案を提示する。商品名から固有名詞を抽出するために、接続頻度判定、固有名詞オントロジー判定、時系列判定を用いる。これらの機能は、メンテナンス作業を繰り返すことでデータを蓄積する。各機能の詳細を述べる。

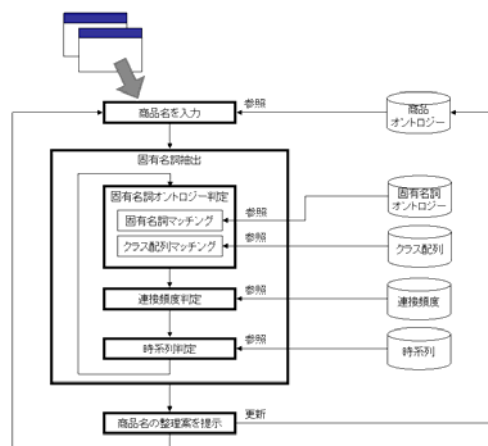


図1 メンテナンスツールのアーキテクチャ

3.3 固有名詞オントロジー判定

目的のオントロジーを構築する過程で作成される辞書オントロジーを利用して、オントロジー構築を支援する研究がある[3]。我々は既に商品オントロジーを構築しており、支援機能に利用することが可能である。固有名詞オントロジーの抽出例を図2に示す。“TOSHIBA dynabook SS RX1”という商品名を入力すると、固有名詞の単位に分解され、“TOSHIBA”, “dynabook”, “SS”, “RX1”となる。固有名詞が未知語の場合は、意味を表すクラスを付与する[4-6]。TOSHIBA は“メー

カ”であり、dynabook は“ブランド”を表し、SS は“シリーズ名”で、RX1 はその“モデル名”である。クラスの付与が完了すると、“メーカー ブランド シリーズ名 モデル名”のクラス配列 A が、ノートパソコンのデータとして登録される。この作業を繰り返すことで拡充した固有名称オントロジーとクラス配列を利用して、新語の獲得を目指す。例えば、新たな商品名“TOSHIBA dynabook Qosmio F40”を入力すると、“TOSHIBA”、“dynabook”、“Qosmio”、“F40”の4つの固有名称が抽出される。次に、固有名称オントロジーを参照して固有名称マッチングを行う。“TOSHIBA”と“dynabook”は既存のデータと一致する。一方、該当データが無い“Qosmio”と“F40”は未知語と判定される。ここで、クラス配列マッチングを行う。クラス配列データを参照し、クラス配列 A が、入力した商品名と同じ長さ4の配列であることがわかる。さらに、配列の項目を比較すると、先頭2つのクラス“メーカー ブランド”の並びが一致することがわかる。よって、残り2つのクラスは“シリーズ名 モデル名”と続く可能性が高いと判断できる。Qosmio を“シリーズ名”、F40 を“モデル名”の新語として獲得する。

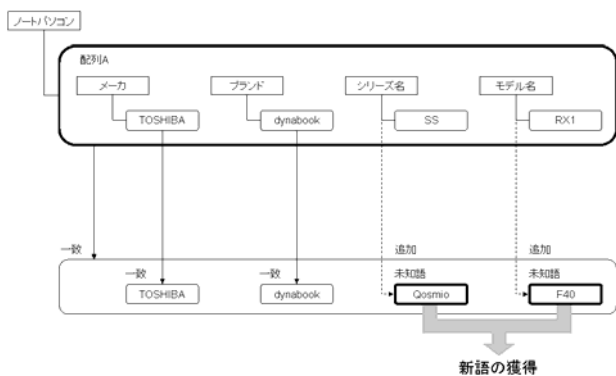


図2 固有名称オントロジーの抽出

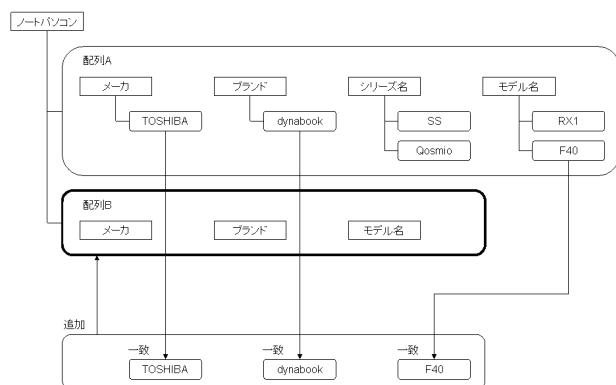


図3 繰り返しによる作業効率の向上

同じ商品分類であっても、クラス配列の長さやクラスの並びが異なる配列を得ることがある。図3のように、新たな商品名“TOSHIBA dynabook F40”から抽出した固有名称が、既に固有名称オントロジーに登録済である場合は、長さの異なる新しい配列 B が登録される。作業を繰り返すことで既知の情報が増え、それをを用いて自動処理が働き作業効率が向上する。

条件にマッチする候補が複数ある場合、最も相応しい情報を提示するために、候補の優先順位を決める必要がある。そこで我々は、固有名称およびクラス配列の出現頻度、固有名称の隣接頻度、時間の経過を加味した固有名称の出現頻度を用いることを検討する。

3.4 隣接頻度判定

コーパスから複合名称を抽出する技術として、古くから C-value[7]が知られている。これを拡張した MC-value[8]は、単名称と複合名称の区別なく固有名称を抽出する手法で、対象の固有名称に隣接する固有名称の出現頻度を算出するものである。MC-value は次式で定義される。

$$MC\text{-value}(a) = length(a) \times (n(a) - \frac{t(a)}{c(a)})$$

a: 固有名称

length(a): a の長さ

n(a): a の出現回数

t(a): a を含むより長い固有名称の出現回数

c(a): a を含むより長い固有名称の異なり数

この手法を用いることで、図4に示す携帯電話や音楽プレーヤ、デジタルカメラなどの機器で使われる“ブラックローズ”や“マラカイトグリーン”などの新しい色名を固有名称として獲得することができる。

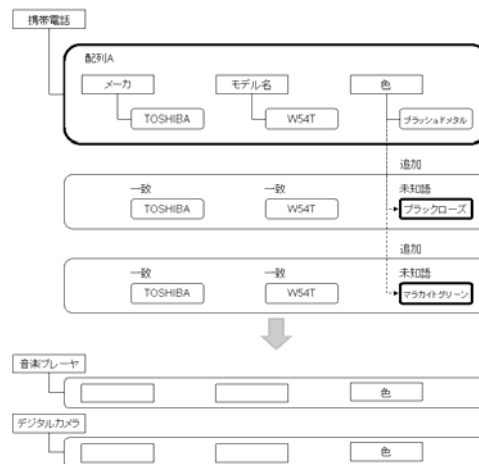


図4 隣接頻度判定

3.5 時系列判定

Web リソースからの情報抽出においては、データの時間的な増減を判定することが有効である[9]。最近になって頻繁に用いられるようになったデータを優先的に提示することや、かつては頻繁に使われていたが使用頻度が落ちてきたデータを提示しないようにすることができる(図5)。判定に用いるデータ範囲を、新規データ集合のみに絞る、または既存データを含めた全集合とする、分野別に切り替えるなどの操作により、状況に応じた精度の高い情報を推薦できる。

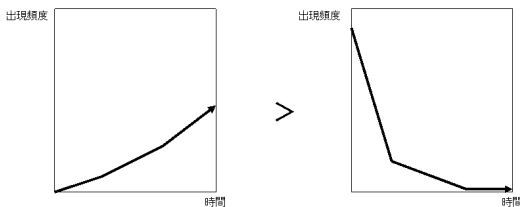


図5 時系列で見た出現頻度

4. 考察

オントロジーのメンテナンスを行う上で、自動化が可能な要件1, 要件5について検討を進めてきた。各検討手法については、メンテナンスツール支援機能としての有効性を検証し、スコアの重み付けを配分する必要がある。

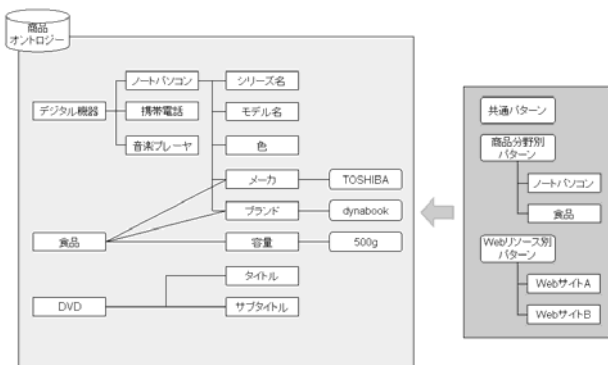


図6 メンテナンス作業のパターン分類

作業者の使い心地に関わる要件2, 3, 4については検討を開始した段階である。図6の例のように、デジタル家電と食品では、同じ“メーカー”や“ブランド”クラスでも、出現する固有名詞が異なるためにノイズが入り易い。そこで、メンテナンス作業のパターン分類を導入し、商品分類やデータを取得した Web リソースなどのパターンを作業者が指定する方法が考えられる。インターフェースについても、十分に検討の余地があるだろう。オントロジーのメンテナンスという観点から、特に、俯瞰性に乏しいグラフや表に替わる表示方法を検討していきたい。

5. まとめ

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について議論した。膨大なデータを入手できる Web リソースからオントロジーを構築することは魅力的であるが、整理されたデータを自動的に取得することは困難である。今回は、半自動化メンテナンスツールの構築に向けて、既存オントロジーを利用した語彙獲得への取り組みを紹介した。

参考文献

- [1] 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)
- [2] WordNet, <http://wordnet.princeton.edu/>
- [3] 廣田 啓一, 佐々木 裕, 加藤 恒昭: オントロジー主導による情報抽出, 人工知能学会誌, Vol.14, No.6, pp. 78-86 (1999)
- [4] 藤井 敦, 石川 徹也: World Wide Webを用いた事典知識情報の抽出と組織化, 電子情報通信学会論文誌, Vol.J85-D-II, No.2, pp. 300-307 (2002)
- [5] Sakai, T., Saito, Y., Ichimura, Y., Koyama, M., Kokubu, T. and Manabe, T.: “ASKMi: A Japanese Question Answering System based on Semantic Role Analysis”, RIAO 2004 Proceedings, pp.215-231, (2004)
- [6] 石谷 康人, 鈴木 優, 布目 光生: 意味クラス解析と意図推定に基づくインタラクティブな情報検索インターフェース, 電子情報通信学会技術研究報告, Vol.106, No.605, pp. 7-12 (2007)
- [7] Frantzi, K. and Ananiadou, S.: “Extracting Nested Collocations”, COLING-96, pp.41-46 (1996)
- [8] 湯本 紘彰, 森 辰則, 中川 裕志: 出現頻度と接続頻度に基づく専門用語抽出, 情報処理学会研究報告, Vol.2001, No.86, pp. 111-118 (2001)
- [9] 藤木 稔明, 南野 朋之, 鈴木 泰裕, 奥村 学, document streamにおけるburstの発見: 情報処理学会研究報告, Vol.2004, No.23, pp. 85-92 (2004)