1G5-OS-19b-3

複数自治体のデータの統合によるゴミ分別オントロジの構築と検索システムの構築

Integrating Garbage Segregation Ontologies of Multiple Municipalities

菊井玄一郎[*1] Genichiro KIKUI
甲斐拓斗[*1] Takuto KAI
渡辺謙一[*1] Kenichi WATANABE
内田三勝[*1] Kazuto UCHIDA
中田幸季[*1] Koki NAKATA
黒木さとみ[*2] Satomi KUROKI
但馬康宏[*1] Yasuhiro TAJIMA
齋藤美絵子[*2] Mieko SAITO

[*1]岡山県立大学 情報工学部
Faculty of Computer Science and Systems Engineering, Okayama Prefectural University

[*2]岡山県立大学 デザイン学部
Faculty of Design, Okayama Prefectural University

This paper describes a method of generating a garbage segregation dictionary for a given segregation category with examples. The method is based on ontology mapping from existing segragation dictionaries provided by different municipalities into the target ontology (dictionary). In order to find similar categories for a given target category, we tried two similarity measures, the kappa coefficient and the dice coefficient. Moreover, we introduced new criteria to eliminate inconsistent catetories. Experimental results show that the created dictionary contains 1340 entries with 88.5% precision.

1. はじめに

社会における様々な規則は当該分野における一定のオントロジ（カテゴリ）体系を人々が共有していることを前提として作成・運用されている．多くの規則は体系の上位レベルの概念を利用することにより一般的な形で記述されているため，背後にあるオントロジ体系を理解していないと具体的な状況に対して適切な規則が適用できない．

本稿で対象とする「ごみ分別」とは，基本的には全ての物理的な実体（以下では「モノ」と呼ぶ）に対して，自治体で定められた規則，および，モノのオントロジを使って，どの分別区分（例：可燃ごみ，粗大ごみ等）に属するかを決定することである．通常，モノのオントロジはモノをそれらの機能や用途によって分類したものであるのに対して，ゴミの分別においてはそのモノを構成する素材，および，大きさを主たる特徴とするため通常のオントロジ（例：ワードネット [Isahara 2008] など）とは必ずしも一致しない．また，多くのモノは複数の素材から構成されているため，どの素材に注目すれば良いか自明ではない．自治体ごとに，焼却炉の性能や利用可能な埋立地（廃棄場所）の状況などが異なるため同一名称の分別区分でも，自治体によってそれぞれに含まれるモノの集合が異なる場合がある．さらに，自治体によっては「商品を販売する際に使用される容器・包装」といった，素材や大きさとは異なる基準が用いられることもある．

これらのため，ごみの分別はその自治体に長く生活している人でも間違うことが多い．さらに，転居等で別の自治体に行くと分別方法に違いがあるため，間違いを誘発しがちである．

分別誤りを軽減する方法として，個々のモノに対してそれらがどういう分別区分かを提示することが有用である．実際，いくつかの自治体はモノ分別区分をモノの名前の五十音順に配列した，ゴミ分別のための「辞書」（以下「分別辞書」と呼ぶ）を各家庭に配布し，的確な分別がなされるように努力している．全ての自治体において十分な被覆率（カバレージ）の分別辞書があれば分別の誤りも少なくなることが期待できる．

しかしながら，実際の分別辞書のカバレージには限界があり掲載されていないモノにしばしば遭遇する上，辞書自体を作成していない自治体も多い．

この問題に対して，オントロジ・マッピングの考え方を用いて分別辞書のカバレージの向上や分別辞書の作成を行うことが考えられる．本稿ではこのうち，分別辞書を持たない自治体に対して他の自治体の分別辞書を活用することにより，分類辞書を自動作成することを試みる．

連絡先: 菊井玄一郎，岡山県立大学，岡山県総社市窪木 111，kikui'at'cse.oka-pu.ac.jp

2. 問題設定

2.1 分別表と分別辞書

自治体からのごみ分別情報の提供形態は「分別表」と「分別辞書」の２つに分けられる．

分別表とは分別区分（例：可燃ゴミ，不燃ゴミなど）ごとにどういうモノのが該当するかをモノのカテゴリのリストあるいは階層構造で整理している．カテゴリ体系は完全に示されているわけではなく，特に小分類や具体物はしばしば文字や絵による例示である．

一方，**分別辞書**とは先に述べたようにモノの名前ごとにそれの分別区分を示したものであるが，適用条件（例：「30cm 以下のもの」）や捨て方の注意（例：「透明な袋に入れる」）などが付記されている．

分別表はほとんどの自治体において提供されているが，分別辞書は比較的人口規模の大きい自治体に限られていると思われる．実際，岡山県の 27 の市町村の全てについて分別表が提供されていたのに対して，分別辞書を提供しているのは 6 市にとどまっている．

そこで本研究では，分類辞書の提供されていない自治体に対して，当該自治体の分類表を手がかり（種）として，分類辞書を自動推定することを試みる．

2.2 分別辞書の推定

分別表における階層の節点（多くの場合　は葉節点）には具体的なモノの名称（普通名詞）が記載されている[*1]ので分別表からも小規模な分別辞書を作成することができるが，被覆率が低すぎて実用にならない．従って，この小規模な辞書の被覆率を何らかの方法で向上させなければならない．

本研究では「オントロジ写像 (ontology mapping)」の考え方を用いてこれを行う．すなわち，別途存在するオントロジ（「元オントロジ (source ontology)」と呼ばれる）のあるカテゴリ C_s を小規模な分別辞書のある分類区分 C_t に対応づけることにより，C_s の要素（今回の場合はモノの（名前の）集合）を C_t の要素に加える．ごみ分別のためのモノの分類体系が一般的なモノのオントロジに基づいているならば，たとえば日本語 Wordnet[Isahara 2008] などを「元オントロジ」として利用できる．しかしながら，上述のようにごみの分類体系は，通常のモノのオントロジとは異なるためこの方法は使えない．そこで，本研究では我々の収集した日本国内の 20 自治体の分別辞書を元オントロジとして構築対象の辞書（分類）を推定することとする．なお，以下では分別のためのカテゴリ構造を 1 階層に縮退させて考える．

2.3 分別情報の電子化

分別辞書の自動推定処理に当たり，まず，既存のデータを電子化する必要がある[*2]．これらのデータは自治体によって HTML（table 構造）で提供されている場合と PDF（しばしば印刷物のスキャン画像由来）の場合がある．前者のうち分別辞書についてはスクレイピングによって csv 形式のデータに変換した．その他については表示イメージを見て手作業でデータを投入した．特に分別表でカテゴリや実例をイラストのみで記述している場合がについては，作業者の判断[*3]によってモノの名称（基本的には普通名詞）に変換してデータ化した．

分別条件が付与されているものについてはモノの名称に分別条件文字列を追加して新たな名称とした．たとえば,「ものさし」に「金属製」という条件が付いている場合は「ものさし（条件：金属製）」とした．条件記述はかなり自由であり同意味異表現というケースが存在するが，今回は特段の処理を行わず，元の表現をそのまま使用した[*4]．

なお，分別区分の名称についても「可燃ごみ」「燃えるごみ」「もえるごみ」などの多様性がある．これらは元テキストの表記のまま投入し，最終的にファイルをマージしてワープロソフト[*5]の「表記揺れ統一」の機能を用いて表記ゆれの統一を行い，マスターデータとした．

さらに，分別辞書のマスターデータの分別区分名に対して，パターンマッチによって表 1 のような「正規化」を行ったものを準備した．

3. 写像手法

オントロジー写像についてはこれまでに様々なアプローチが提案されている ([Agrawal 2001], [市瀬 2002], [市瀬 2007], [市瀬 2008] など).

多くのオントロジー写像が特定の一つの元オントロジーをある程度のインスタンス（あるいはサブカテゴリ）数を持つ写像先オントロジに写像するのに対して，本稿でのオントロジ写像には次のような特徴がある．

1. 写像先オントロジの規模が小さい（インスタンス，カテゴリのノード数が少ない）
2. インスタンスの出現頻度が存在しない（「異なり語」のリストである）
3. 複数の元オントロジが存在する
4. 元オントロジの一部は写像先の体系に対して「ノイズ」になる可能性がある

これらの特徴を考慮して，本研究では，確率的な分類学習のアプローチ ([Agrawal 2001] など) は適用せず，写像先オントロジの各カテゴリについて，元オントロジーに存在するカテゴリの中から適切なものを選択することによって写像を行うこととした．

表 1: 区分名の正規化

分別区分名称の例	正規化名
燃えるこみ，可燃ごみ，燃やせるごみ	もえるごみ
燃えないこみ，不燃ごみ	もえないごみ
粗大ごみ，そだいごみ	粗大ごみ
危険ごみ，きけんごみ	危険ごみ
埋め立てごみ，うめたてごみ	埋立ごみ
出せないごみ，収集できないごみ	出せないごみ

3.1 手法 1：単一体系の選択

最も単純な方法は，他自治体の分別辞書の中から，写像先の分別カテゴリ名の集合と同じか「類似した」分別カテゴリ名の集合を有するものを 1 つ選んでそのまま利用するというものである．一つの分別辞書は基本的に無矛盾に作られているので一つのモノが複数の分別区分に入るようなことは起こらないというメリットがある．

実際にそのような選択が可能かどうかを調べるため，分類区分名の単純な文字列一致を用いて今回データ化した 20 自治体の分別辞書相互で分別区分の集合の一致度を DICE 係数によって評価すると最大 0.5 であり一致度は低い．一方，分別区分の名称を正規化した場合は最大で DICE 係数 0.91 の組が存在した．

正規化したものを用いる方が一致度は向上するが，例えば，「可燃ごみ」と「燃やすごみ」を同一視してよいかどうか自明ではない．どちらの場合でも単にどこかの自治体の分別辞書を利用するだけであるから被覆率は限定的であり，本稿では検討の対象外とする．

3.2 手法 2：カテゴリ名の類似性による写像

一つの自治体の分別辞書を丸ごと利用するのではなく，複数の自治体の辞書の中で使えそうな分別区分（カテゴリ）の情報を選択的に利用する方法である．選択の基準としてここでは，分別区分名の類似性を用いる．すなわち，構築対象辞書の各分別区分について，当該分別区分の名称と類似した名称のカテゴリを全て集めて利用する．手順は次の通りである．

1. 全ての辞書ファイル (csv 形式) をマージし，品名 (捨てるものの名称) でソートする
2. 分別区分を正規化する

∗1 イラストのみという場合も含む
∗2 ごく一部の都市ではボランティアによって RDF 化されている
∗3 通常は 1 名で, 判断に迷う場合は 2 名で判断した
∗4 条件を無視すると性能が下がる
∗5 Microsoft Word

3. 同一品名であるのに（自治体によって）異なる区分になっている場合，当該品名のレコードを全て削除する

4. 構築対象自治体の各分別区分について，完全一致あるいは類似した分別区分名をもつレコードを集めて辞書項目とする

上記において分別区分名の類似性の判定は２つの文字列の文字数の和に対する共通の文字数 (の 2 倍) の率が一定の閾値以上であるかどうかによって行う．

3.3 手法 3：カテゴリ内の要素の重なりによる写像

この手法も手法 2 と同様に，他自治体の分別辞書の中で構築対象自治体の分別情報として「使えそうな部分」を選択するという方法であるが，分別区分を選択する際に分別区分名（カテゴリ名）の類似性ではなく，カテゴリに含まれる要素の共通性を用いる所が異なる．

3.3.1 基本的な方法

分別区分の選択方法として，市瀬らによる HICAL[市瀬 2002]，および，比較のため dice 係数を用いる方法を試みた．

HICAL は異なる体系に属する２つのカテゴリについて，それらがどの程度インスタンスを共有しているかをカッパ統計量を用いて評価し，ある一定以上の有意水準で対応関係があるかどうか判定している．我々の問題においてはインスタンスというものは存在しないが，分類されるべきモノの名前をインスタンスとみなして同手法を適用した．但し，モノの名前にはゆらぎがあるなどインスタンスとは異なる部分がある．

また，今回のデータでは複数のカテゴリ体系（各自治体の分別辞書）に出現するモノの名前の集合が大きく異なる，すなわち，２つのカテゴリの一方にしか現れないインスタンスが多く存在している．kappa 値を取る際には原則としてデータ全体に対して双方の体系でどのカテゴリに所属するかが判断が付与されている必要がある．そこで，今回は対応を評価するカテゴリ（分別区分）の所属する２つの体系（２つの自治体の辞書全体）の一方にしか出現しないインスタンス（モノの名前）は計算から除外した．

dice 係数による方法は，２つのカテゴリについてそこに含まれるインスタンス（モノの名前）の一致の程度を dice 係数で評価し，一定の閾値を超えるものを「対応あり」とみなす方法である．この方法は単に 2 つの集合の共通部分とそれぞれの単独の部分の要素数から計算できるので上述のような問題はない．しかし閾値を設定する基準がない．

3.3.2 一貫性の検査

分別区分辞書は判別のために用いるものであるから，同一物が複数の分別区分に入っていてはいけない．このような状況を排除して一貫した (consistent な) 辞書を出力するために手法 2 では，作成された辞書において，

> 同一のモノ（分類要素）が複数の分別区分（カテゴリ）に含まれる場合，このモノに関するレコードを全て削除する

という最終結果をフィルタリングする方法をとった．ここでもその方法を採用する．

これに加えて，カテゴリを選択する段階でも次に示す手法で一貫性のチェックを試みた．

> 対応関係があると判定された分別区分のうち，構築対象の辞書において複数の分別区分に属するインスタンスを含むものは一貫性違反とする．

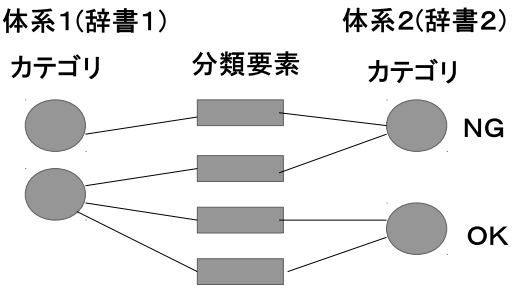

図 1: 一貫性の説明図

表 2: カテゴリ対応の例

係数値	写像先	写像元	
	カテゴリ	自治体	カテゴリ
0.67	出せないごみ	玉野	家電
0.62	出せないごみ	浜頓別	家電リサイクル
0.46	出せないごみ	各務原	家電リサイクル法対象品
0.28	資源ごみ	各務原	紙類、古着
0.20	出せないごみ	阿賀野	禁止
0.20	もえないごみ	阿賀野	もえないごみ
0.18	もえるごみ	川崎	普通ごみ

図 1 において体系 2 の NG と記された分別区分はこの区分に所属する分類要素（インスタンス）が体系 1 の側で複数の分類区分に所属するため一貫性違反となる．

一貫性違反したカテゴリを辞書に含めないと，辞書のサイズは小さくなるが，より一貫性が向上することが期待される．

4. 評価実験

上記の手法 2 および 3 について簡単な実験を行った．

まず，分別辞書を推定する自治体を選び，そこの分別表の電子化データをから初期分別辞書（カテゴリ体系）を構築する．これを写像先オントロジとして，他の自治体の分別辞書を先述の方法で追加（写像）する．

比較した手法は方法２と方法３であり，前者についてはカテゴリ名の完全一致と部分一致（一致率が 0.6 以上のもの）の２種類，後者については kappa 係数，dice 係数，それぞれについて，カテゴリ選択の際の一貫性検査をするものとしないものの４通りで実験した．なお，kappa 係数を使う際の危険率は 0.05, dice 係数の時の閾値は 0.1 とした．

評価尺度は，本来，適合率と再現率で行うべきであるが，理想的な辞書の構築が困難であったため，適合率と辞書サイズ（モノの名前の数）を用いた．

なお，推定対象自治体として岡山県総社市を用いた．

まず，抽出されたカテゴリの対応関係について dice 係数による一致度の高いものから 7 ペアを表 2 に示す．

表の写像先は今回辞書構築の対象である総社市のカテゴリである．例えば，表の４行目を見ると各務原市の「紙類，古着」が総社市の「資源ごみ」に対応づけられていることが分かる．このように対応の多くは整合的であり，カテゴリ名が大きく異なっていても対応づけられていることが分かる．

次に定量的な評価結果を表 3 に示す．表の「尺度」において，「名称 A」および「名称 B」はそれぞれ方法 2 の完全一致と部分一致に対応する．

この結果を見ると，方法３のカテゴリ単位で選択する方法

表 3: 実験結果

手法	一貫性	適合率 j%(正解個数)	辞書サイズ
名称 A	-	80.0 (554)	692
名称 B	-	75.9 (727)	957
Kappa	no	88.5(1186)	1340
Kappa	yes	87.2(512)	587
dice	no	88.8 (384)	432
dice	yes	93.3 (153)	164

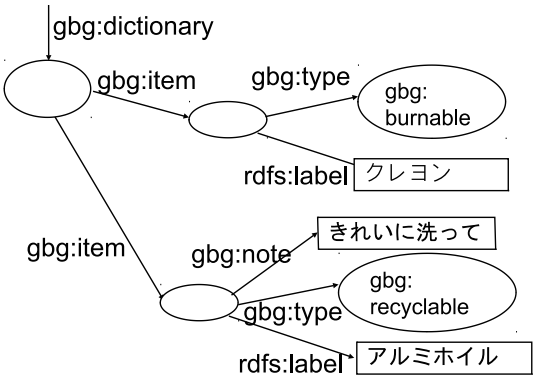

図 2: 辞書項目の RDF の例

が全データを使う方法より，優れているといえる．dice 係数について閾値 0.1 は決して高いとはいえないが，予想以上に制限的であったと思われる．また一貫性のチェックによる適合率の向上は必ずしも顕著ではなかった．

一方，kappa 値の場合は先述したように 2 つの体系の一方にしか現れないインスタンスが多く存在し，これらを排除することにより，恐らく対応性が過大評価されているように思われる．この問題に関する検証は今後の課題である．

また，表記ゆれの問題も大きく，本来，一つのモノであるにもかかわらず，統合されないケースが散見された．これについては自然言語処理等による解決が必要である．また，オントロジー辞書（wordnet など) を用いることも考えられる．

5. 検索処理

得られた辞書を用いて RDF を作成し，これに対する簡単な検索インタフェースを作成した．RDF 化にあたっては一つのモノを一つのリソースとし，gbg:type なる property 述語で分別区分を示すリソースと結びつけることとした (図 2) ．

RDF は vituoso(オープンソース版) で管理し，入力した文字列と部分一致する品名を分別区分と共にリスト表示するインタフェースを作成した (図 3)．部分文字列で入力できるものの，利便性を向上させるためにはさらなる改善が必要である．

図 3: 検索画面の一部

6. おわりに

ごみ分別のための分別辞書の推定方法について述べた．HICAL,dice 係数などを用いることで適合率約 90%程度で辞書を構築できることが分かった．

本稿での検討はまだ初期段階であり，たとえば複数の分類体系をクラスタリングしてマージするなど複数の体系が全体として持つ情報を更に利用する方法を検討する必要がある．また，広い意味での表記のゆらぎの解決が必要である．モノの名称については Wordnet 等のオントロジを用いることで同義語との照合が行える可能性がある．分別条件については一定のパターンがあるものの（例：「Xcm 以下のもの」）かなり自由に記述されているのでより深い言語処理，あるいは，人手のアノテーションが必要と思われる．

参考文献

[Agrawal 2001] Agrawal, R. and Srikant, R.: On Integrating Catalogs, WWW10, pp.603-612, (2001).

[Isahara 2008] Isahara, H. et al.: Development of Japanese WordNet, LREC-2008(2008),

[市瀬 2007] 市瀬龍太郎: 情報の意味的な統合とオントロジー写像, 人工知能学会誌, Vol.22, No.6, pp.818-825 (2007)

[市瀬 2002] 市瀬龍太郎，ほか: 階層的知識間の調整規則の学習, 人工知能学会論文誌, Vol.17, No.3, pp.230-238 (2002)

[市瀬 2008] 市瀬龍太郎: オントロジーマッピングに有効な特徴の抽出, 第 22 回人工知能学会全国大会, pp.2E1-1 (2008)