意味を考慮した一般化階層自動生成技術

岡田 莉奈[1,a)]　長谷川 聡[1,b)]

概要：現在、国内では 2017 年に個人情報保護法が改正され、異種業界横断的なデータ分析の活発化に役立つ匿名加工情報に注目が集まっている。匿名加工情報を作成するための匿名化技法の一つに個人情報を汎化した表現に変える “一般化” と呼ばれる技法がある。一般化を行うためには、一般化階層と呼ばれる汎化ルールを定義する必要がある。現代のデータ分析で扱うデータ量は膨大であることから、一般化階層を人手で作成することは困難であるため、自動的に一般化階層を作成する従来技術がある。しかしながら、従来技術では、特に非数値データの一般化階層の作成において、一般化対象データの意味を考慮しないため、データ分析に適さない匿名加工情報を生成することがある。本稿では、Web 等から収集した情報を利用することにより、一般化対象データの意味を捉え、データ分析の幅を広げる一般化階層自動作成技術を提案する。提案技術の有効性については実データを用いた実験を通して報告する。

1. はじめに

現在、国内では 2017 年に個人情報保護法が改正され、異種業界横断的なデータ分析の活発化に役立つ匿名加工情報に注目が集まっている。匿名加工情報とは、収集した個人情報を個人が特定されないように加工したデータのことである。個人情報保護法では、匿名加工情報であれば個人に同意を得ずに第三者提供や目的外利用が可能になる。つまり、匿名加工情報を得た人や組織は分析者として、匿名加工情報に対して様々な分析を行うことができるようになる。

匿名加工情報を作成するための加工方法を匿名化と呼び、匿名化の一つに個人情報を汎化した表現に変える “一般化” と呼ばれる技法がある。一般化を行うためには、一般化階層と呼ばれる汎化ルールを定義する必要がある。

1.1 一般化階層生成の課題

一般化階層生成の課題として「自動生成」と「データ構造」の二点が挙げられる。

自動生成: 一般化対象のデータがカテゴリカルなデータの場合、一般化対象データ内の各値には単純には汎化関係が無い。各値の意味を考慮しない一般化階層による一般化は、一般化前の情報が損なわれ、分析の価値を低下させてしまう。それゆえ、各値の意味を解釈した上で概念間の関係性を自動で獲得し、一般化階層を定義する必要がある。

データ構造: 概念間の関係性を構造化したものとして、意味 (セマンティック) ネットワーク [4] やオントロジー [1] がある。これらは「is-a(の一つである)」や「has（を持つ)」などの概念間の関係を表すリンクから成るグラフ構造である。特に、「is-a」関係は汎化関係とも考えられ、一般化階層を含んだ情報であると理解することができる。しかしながら、意味ネットワークやオントロジーの is-a 関係は下位概念から上位概念にかけて N(多) 対 1 の関係ではなく N(多) 対 M(多) の関係を持っているため、複数の汎化 (一般化) 候補が出現してしまう (図 1)。それゆえ、一般化対象データ内の各値の一般化候補を一意に定めることができるよう、N 対 1 の関係で一般化階層を構成する必要がある。

1.2 関連研究

データ構造の課題を克服するためには、既存の意味ネットワーク等を用いて、N 対 1 関係で構成すれば良い。このため、本節では、自動生成に関連する既存研究を記す。

一般化階層を自動で生成する方法として、原田らによる方法 [9] がある。これは一般化対象データ内の各値の出現頻度に基づき、一般化による情報損失が小さくなるよう一般化階層を自動生成する。この手法は、値の出現頻度にしか着目しておらず、値の意味が考慮されない。それゆえ、意味が考慮されない一般化が行われてしまうことで、分析に適さない匿名加工情報が生成されてしまう可能性がある。

1.3 本稿の貢献

本研究の目的は、一般化対象データの意味を考慮した一

[1] NTT セキュアプラットフォーム研究所
3-9-11, Midori-cho Musashino-shi, Tokyo 180–8585, Japan
[a)] rina.okada.hg@hco.ntt.co.jp
[b)] satoshi.hasegawa.ks@hco.ntt.co.jp

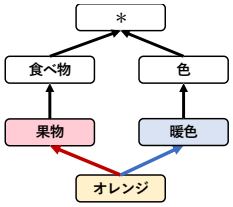

図 1 1 対 M 関係の一般化階層の例 (“オレンジ” に着目した場合、$M = 2$)。なお、「*」は最上位概念を表す。

般化階層を自動生成することである。本稿の提案手法はこの目的を達成し、下記三点の特徴を持つ。

(1) 概念関係の事実の抽出と推測の両方を利用することにより、正確性が高く、量の多い一般化階層を生成する。

(2) 概念関係の事実の抽出と推測の処理において、常に N 対 1 の関係でデータ構成することで、一般化対象データ内の各値の一般化候補を一意に定めることができる。

(3) 一般化対象データ内の各値に対して、複数の一般化階層候補が出現した場合、できる限り他の値と上位概念の値が同様な一般化階層になるよう、一般化階層を分割する。これにより、特定のデータだけが他と異なる一般化になることを防ぐ。

2. 提案手法

提案手法の流れを図 2 に示す。入力は、一般化対象データと汎化関係辞書 (下位概念から上位概念の関係を保持するリストの集合) とする。処理は、入力の一般化対象データと汎化関係辞書から一般化階層を構成するために必要な情報を抽出、追加するフェーズ 1 と、フェーズ 1 にて抽出した情報をもとに分析価値 (= 有用性) の高い一般化階層を分割するフェーズ 2 から成る。出力は、一般化対象データの概念を含む一般化階層の集合である。

2.1 入力

2.1.1 入力データ

匿名化前のデータは図 3 左のようなデータである。図 3 のデータには、「日時」、「場所」、「購買品」という属性が含まれている。このような各属性が本提案手法の入力の一つである一般化対象データ $\boldsymbol{x}$ になりうる。汎化関係辞書 D は、図 3 中央のような下位概念から上位概念の関係を表すリスト (以後、汎化関係リストと呼ぶ) の集合[*1]であり、公開されているデータベース (例えば、分類語彙表 [8]) や Web ページのスクレイピングによって入手する。

例えば、図 3 の属性「購買品」$\boldsymbol{x}$ = (苺, オレンジ, 柿) を図中の汎化関係辞書 D = {(*, 食べ物, 果物, オレンジ), ⋯ , (*, 食べ物, 果物, 柿)} を用いて一般化すると、一般化後は、図 3 右の $\boldsymbol{x}'$ = (果物, 果物, 果物) になる。

[*1] 説明の都合上、リストの集合で記載しているが、同じ概念をまとめ上げることにより、グラフ構造や木構造となることは明白である

2.1.2 汎化関係辞書の問題点

一見、汎化関係辞書 D さえあれば目的が達成されたように思える。しかし、一般化対象データ $\boldsymbol{x}$ 内の値がすべて D に含まれているとは限らない。さらに、D をそのまま利用すると、「一般化後データの統一感」と「一般化の度合い」といった有用性の問題が生じる。

一般化後データの統一感: 一般化には方向がある。例えば、“オレンジ” という概念を一般化する際に、“食べ物” という方向と “色” という方向への一般化が考えられる。このように一般化には方向があるため、D には単一の概念に対して複数の方向を持つ汎化関係が含まれうる。複数の方向を持つ汎化関係を用いた一般化の場合、一般化後のデータに統一感が生まれず、データ分析に差し支えのある状態となってしまう可能性がある。例えば、図 4 の場合、いかにも “食品” であるような一般化対象データにも関わらず、“苺” は “果物”、“オレンジ” は “暖色” という一般化が行われることは分析上差し支えがある。

一般化の度合い: どこまで一般化すれば匿名加工情報に近づくのか、に対して匿名加工ガイドライン [7] では、k-匿名性 [5] と呼ばれるプライバシ保護指標を満たすと好ましいと書かれている。k-匿名性を達成するためには、少なくとも k レコード以上同じ値を持つことが条件となる[*2]。

図 3 と図 4 の二つの例を考える。図 3 では、一つの方向のみで構成された D を利用して一般化対象データ $\boldsymbol{x}$ が 2-匿名性を満たすようにしており、$\boldsymbol{x}$ 内の “苺”、“オレンジ”、“柿” を全て 1 つ上位の概念に一般化することで 2-匿名性が満たされる。一方、図 4 では、複数の方向が含まれる D を利用して一般化対象データ $\boldsymbol{x}$ が 2-匿名性を満たすようにしており、$\boldsymbol{x}$ 内の “苺”、“オレンジ”、“柿” を全て 3 つ上位の概念に一般化することで 2-匿名性が満たされる。図 3 と図 4 を比較すると、図 3 のほうが一般化の度合いが低く済んでおり、一般化前と一般化後の概念の階層の差が小さいため、一般化後のデータの有用性が高いといえる。このように、複数の方向を持つ汎化関係から適切な一般化階層を作成する必要がある。

2.2 処理

2.2.1 フェーズ 1: 抽出と追加

まず、一般化対象データ $\boldsymbol{x}$ 内の各値と汎化関係辞書 D 内の各値に表記ゆれがある場合は、表記を統一する。次に、表記統一された D を用いて、表記統一された一般化対象データ $\boldsymbol{x}$ に関係のある一般化階層 (汎化関係リスト) を抽出する (Algorithm 1)。

D に一般化対象データ内のすべての値が含まれているとは限らないため、$x \notin D$ の一般化階層を生成する必要がある (Algorithm1, 14 行目に相当)。そこで、Algorithm 1 内

[*2] 正確には、準識別子と呼ばれる属性の属性値が k レコード以上同じ値を持つ。

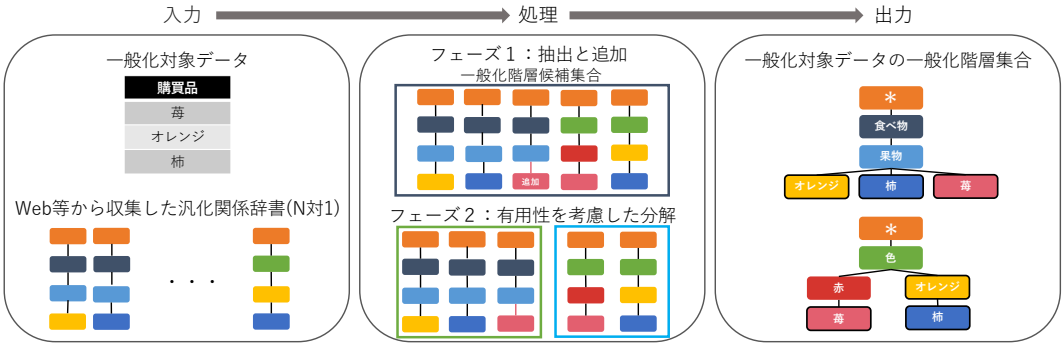

図 2　提案手法の流れ

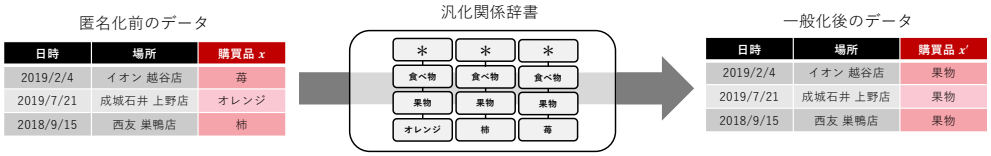

図 3　一般化の流れ (2-匿名性を満たす例)

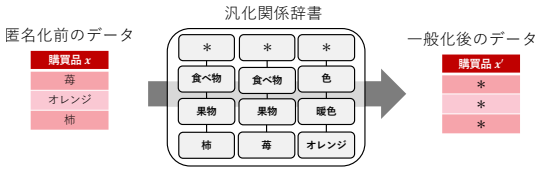

図 4　有用性の低い一般化の流れ (2-匿名性を満たす例)

Algorithm 1 フェーズ 1: 一般化階層の抽出と生成

Input: 表記統一された一般化対象データ $\boldsymbol{x}$、表記統一された汎化関係辞書 D
Output: 一般化階層候補集合 C

```
1: C = {}
2: for x ∈ x do
3:     existsFlg = False
4:     for d ∈ D do
5:         if x ∈ d then
6:             d の最上位概念から x までのリストを d' とする。
7:             C = C ∪ d'                          ▷ 抽出
8:             existsFlg = True
9:         end if
10:    end for
11:    if existsFlg = False then
12:        C = C ∪ generate(x, D)                  ▷ 生成
13:    end if
14: end for
15: return C
```

の 15 行目の generate 関数では、類似概念検索を用い x に最も類似した概念 $d \in \boldsymbol{d} \in D$ を取得し、d を x に置換した $\boldsymbol{d}$ の集合を出力する[*3]。

類似概念検索：generate 関数で利用する類似概念検索として 2 段階の方法を用いる．1 段階目は、分散表現 [3] による類似度評価を利用した方法である。分散表現はベクトルとして表されるため、ユークリッド距離やコサイン類似度を用いて概念間の類似度を評価することができる。類似度を評価する概念が固有名詞である場合を考慮し、対象の概念を形態素に分解してから分散表現によって類似度評価を行うこととする（分散表現は形態素解析を行った文書を用いて学習される。固有名詞の分散表現は存在しないことが多々あるため)。

1 段階目において類似度評価ができない場合、すなわち分散表現側に評価したい概念が含まれていない場合、2 段階目を用いる。2 段階目では、対象の概念を形態素に分解し、共通する形態素を多く含む概念ほど類似度が高いと評価する方法を用いる。

2.2.2　フェーズ 2: 有用性を考慮した分割

Algorithm 1 によって出力される一般化階層候補集合 C は、汎化関係リスト集合 (図 2 のフェーズ 1 で得られるリスト集合) となっており、2.1 節にて述べた複数の方向を持つ一般化階層が含まれる。したがって、C をそのまま利用すると有用性の観点で問題があるため、フェーズ 2 では C から方向を考慮した一般化階層を作成する (Algorithm 2)。

フェーズ 2 では、複数の方向を持たないよう、一般化階層候補集合 C の中から、一般化対象データを含む汎化関係リストを取得し、一般化階層を作成する。その際、一般化の度合いを考慮し、上位概念が最も共通するような一般化階層を抽出する。上位概念が共通することにより、一般化対象のデータが同じ方向の概念に一般化される (Algorithm 2: 2-6 行目)。よって、k-匿名性を満たす際必要な一般化の度合いが小さくなる。

3.　評価実験

本提案手法を計算機上に実装し、実データを用いて意味を考慮した一般化階層を抽出できるかを確認する。

[*3] このとき正確には、当該 $\boldsymbol{d}$ の d より具体的な概念部分は削除する。

Algorithm 2 フェーズ 2: 有用性を考慮した分割

Input: 表記統一された一般化対象データ $\boldsymbol{x}$, フェーズ 1 の出力結果である一般化階層候補集合 C
Output: 本提案手法の出力である一般化階層の集合 $\mathcal{C}$

1: $U = \{C\}$ とする。
2: $h = \max_{\boldsymbol{d} \in C} \boldsymbol{d}$.length ▷ $\boldsymbol{d}$.length はリスト $\boldsymbol{d}$ の長さとする。
3: **for** $i \in [0, h]$ **do**
4: **for** $T \in U$ **do**
5: T 中の汎化関係リストのうち, i 番目の位の概念が共通する汎化関係リストで新たなグループ $T_1, \cdots, T_n$ を作成し ($T_1, \cdots, T_n$ のそれぞれは汎化関係リストの集合である, n はできたグループ数に依存する)、U から T を取り除き $U = U \cup \{T_1, \cdots, T_n\}$ とする。
6: **end for**
7: **end for**
8: グループ化した汎化構造リスト集合 U からそれぞれのグループごとに木構造 (一般化階層)L を作成し、順次 $\mathcal{C} = \mathcal{C} \cup L$ に格納する。
9: **return** $\mathcal{C}$

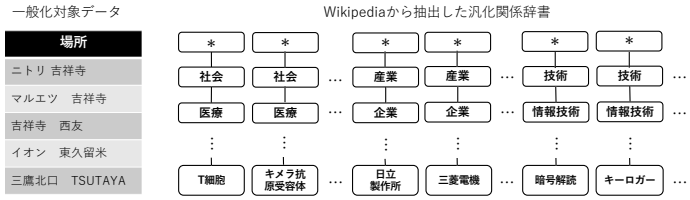

図 5 入力データ

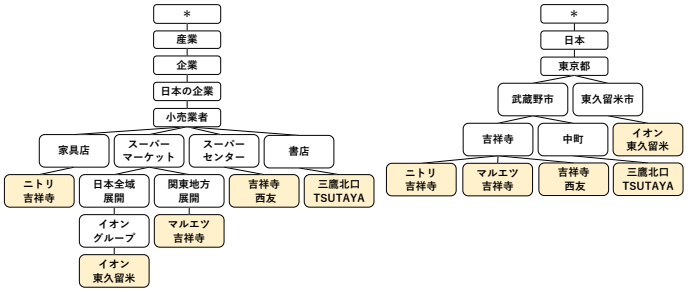

図 6 出力データ

3.1 実験設定

入力とする一般化対象データと一般化階層を図 5 とする。図 5 内の一般化階層は、Wikipedia:カテゴリ [*4] ページ内の「百のカテゴリ」内に含まれる項目のリンクを辿る [*5] ことによって得たデータである。各概念の形態素解析する方法として MeCab [2], [6] を、各概念を分散表現で表す方法として Word2Vec [3] [*6]を利用した。

[*4] https://ja.wikipedia.org/wiki/Wikipedia:Category
[*5] 注意点: Wikipedia はスクレピングを禁止しているため、Wikipedia データベースをダウンロードし、ローカル環境でリンクを辿った。
[*6] Word2Vec の学習に使用したデータは 2019 年 5 月 7 日時点で最新の Wikipedia 記事の dump データであり、モデル構築時に使用したパラメータは { 単語ベクトルの次元, 無視する単語を決定する最低頻度, 各文にて考慮する最大の単語幅 }={200, 20, 15} とした。

3.2 結果と考察

提案手法を用いて得られた一般化階層集合を図 6 に示す。Wikipedia 中の膨大な汎化関係の中から、今回対象となったデータに適した一般化階層が 2 つ抽出された。図 6 の結果より、1 つは「お店の種類」による一般化階層、もう一つは「地理情報」による一般化階層である。フェーズ 2 のアルゴリズムを通すことで、上位概念が共通になるように一般化階層が生成されている。特に地理情報による一般化階層では、比較的細かい粒度である町域での一般化も可能となっており、フェーズ 2 による効果があったと言える。このように提案手法の想定した意味が考慮された一般化階層を取得できたことが分かる。取得した一般化階層は 2 つ出力されているが、匿名加工者あるいは分析者が「お店の種類」あるいは「地理情報」による一般化階層のどちらかを選択し一般化処理を実施すれば良い。

4. まとめ

パーソナルデータを一般化する際に使用する一般化階層を自動的に生成する技術を提案した。正確かつ大量の一般化階層を生成するために、概念間の関係性の抽出と推定の処理を行った。また、有用性を高めるために、一般化の方向を考慮した一般化階層の分解方法を考えた。これにより、これまで分析に利用しやすい一般化階層の定義は匿名加工者の負担であったが、その負担の軽減を実現させた。

参考文献

[1] Thomas R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.*, Vol. 43, No. 5-6, pp. 907–928, December 1995.
[2] T. KUDO. Mecab : Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.net/*, 2005.
[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc., 2013.
[4] M. Ross Quillian. The teachable language comprehender: A simulation program and theory of language. *Commun. ACM*, Vol. 12, No. 8, pp. 459–476, August 1969.
[5] Latanya Sweeney. K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, Vol. 10, No. 5, pp. 557–570, October 2002.
[6] Sato Toshinori. Neologism dictionary based on the language resources on the web for mecab, 2015.
[7] 個人情報保護委員会. 個人情報の保護に関する法律についてのガイドライン (匿名加工情報編). November 2016.
[8] 国立国語研究所コーパス開発センター. 分類語彙表—増補改訂版データベース—.
[9] 原田邦彦, 佐藤嘉則. 一般化階層木の自動生成と情報エントロピーによる歪度評価を伴う k-匿名化手法. 研究報告コンピュータセキュリティ（CSEC）, Vol. 2010, No. 47, pp. 1–7, jun 2010.