

部分構造を用いた類似例規の検索

藤岡 和弥[†] 駒水 孝裕[†] 小川 泰弘[†] 外山 勝彦[†]

[†] 名古屋大学 〒464-8603 名古屋市千種区不老町

E-mail: fujioka@k1.i.is.nagoya-u.ac.jp, taka-coma@acm.org, {yasuhiro,toyama}@is.nagoya-u.ac.jp

あらまし 例規とは、地方自治体が定める条例・規則の総称である。各自治体が例規を作成する際には、一般に、他の自治体の類似例規を参照する。しかし、従来の例規検索手法は、例規の条文テキストを入力として、それと類似した条文テキストからなる例規を検索するものである。そのため、例規の一部分だけが類似している場合に、それを検索することは容易ではない。このような例規は、類似した部分を特定することにより検索できる。そこで、本稿では、例規が半構造文書であることに着目し、例規のうち類似した部分の探索を例規の部分木探索タスクとして定義した。また、テキスト類似度を用いた例規の部分木間類似度 2 種類を提案し、これらを用いた例規の類似部分探索手法を提案した。実験により、提案手法の有用性を示した。

キーワード XML, 部分構造検索, 類似検索, 例規検索

1 はじめに

2019 年 2 月現在、日本には 47 都道府県、1,741 市区町村の合計 1,788 の地方自治体が存在する。各自治体は条例や規則を自ら制定しており、これらは例規と呼ばれる。このうち条例は地方自治体の議会が、また、規則は都道府県知事や市区町村長がそれぞれ定めるものである。例規を定める権利は日本国憲法第 94 条および地方自治法第 14 条と第 15 条によって保証されており、各自治体は例規を解釈・執行することを通じて、行政を執り行っている。

複数の自治体の例規の中には、それらの制度の目的の全部または一部が共通しているものがある。それらを類似例規と呼ぶことにする。例えば、青少年の保護や健全な育成を目的として愛知県が制定している「愛知県青少年保護育成条例」の類似例規には、石川県の「いしかわ子ども総合条例」や福岡県太宰府市の「太宰府市青少年に有害な図書等の自動販売機の設置規制に関する条例」などがある。これらは一般に青少年保護育成条例と呼ばれており、47 都道府県すべてが定めている。多くの自治体に共通している例規は、この他にも、景観の保全や創生について定めた「景観条例」や、暴力団の排除について定めた「暴力団排除条例」などがある。

類似例規は、例規作成の際に参照資料として用いられるほか、自治体を対象とした研究において収集・分析されることがある。どちらの場合も、類似例規を収集・比較し、その類似部分の間の差分を明らかにすることが重要である。しかし、すべての自治体の例規を集めると 100 万件以上にもなるので、その中から類似例規を手で収集することは容易ではない。そのため、例規名が類似した例規を検索する研究 [1] や、条文のテキストの類似度を用いて、例規の検索 [2] や分類 [3] をする研究が行われてきた。

しかし、類似例規の目的どうしを比較すると、一方の目的が他方の目的に対する上位の目的であったり、あるいは部分的に

第一章 総則

第一節 通則

第二節 基本理念等

第三節 施策の基本的な考え方

第二章 子どもの健全な育成

第一節 乳幼児の出生及び発達の保障

第二節 青少年の健全な育成

第三節 青少年の健全な育成を 阻害する行為の規制

第三章 若者の自立に向けた支援

第四章 子育て支援

第一節 地域社会全体による子育て支援

第二節 子育てをする雇用労働者への配慮

第五章 食育の推進

第六章 子どもの権利擁護

第七章 石川県子ども政策審議会

第八章 雑則

第九章 罰則

図 1 いしかわ子ども総合条例の章と節の構成

重なっていたりするなど、全体では一致しない場合がある。例えば、愛知県青少年保護育成条例（全 31 条）は、そのすべての内容が青少年の保護や健全な育成の促進に関するものであるが、いしかわ子ども総合条例は、図 1 に示すように、青少年だけでなく、乳幼児や若者を対象とした内容も定めている。愛知県青少年保護育成条例と対応する部分は、青字で示した第 2 章の第 2 節（全 8 条）と第 3 節（全 22 条）に該当し、残りの 69 条（特に、赤字で示した食育や子育て支援など）は、そのほとんどの内容が愛知県青少年保護育成条例と関係のない内容である。

このことから、類似例規検索とは、制定の目的が共通する部分をもつ例規の検索といえる。この部分を例規の類似部分と呼

ぶことにする。例規の類似部分の探索は、(1) 例規のテキストから意味的なまとまりとなっている部分を抽出し、(2) 抽出した部分の間の類似度をそれぞれ計算することによって実現できる。このタスクを例規の類似部分探索タスクと呼ぶことにする。

例規文書は条や項などのテキストの単位が入れ子状になった半構造文書である。この構造は人間が意味的なまとまりを階層的に与えたものであり、このまとまりが例規の類似部分の候補となる。つまり、例規の類似部分探索タスクは、二つの木それぞれから互いに類似する部分木を探索するタスクであるとみなせる。

本稿では、例規の部分木の類似度として、MTR (Matching Text-node Rate) と ILINS (Inversed Longest Isolated Node Sequence) の二つと、それらを用いた例規の類似部分探索手法を提案する。MTR は、部分木のテキストの類似性を、マッチするテキストノードの割合によって評価する類似度である。一方、ILINS は、マッチするテキストノードが偏っていないことを評価する類似度である。MTR はテキストノード列の順序を考慮しない類似度であるが、ILINS は、それを補うことができる類似度である。これらを組み合わせることで、マッチしたテキストノードを十分に、かつまんべんなく含む部分木を検索可能にする。提案手法は、MTR と ILINS を用いて、可能なすべての部分木ペアのうち、類似度が 0 にならないものの類似度を総当たりで計算し、冗長なものを除いて出力する。

本研究は、類似例規検索タスクに対して以下のような貢献がある。

(1) 類似例規検索のサブタスクである例規の類似部分探索タスクを定義した。

(2) 二つの類似度 MTR と ILINS, およびそれらを用いた例規の類似部分探索手法を提案した。

(3) 例規の類似部分探索手法の性能評価に有効な実験を設計・実施し、提案手法の有用性を確かめた。

本論文の構成は次の通りである。2 節では、例規の文書の構造や特性について説明する。3 節では、第 2 節の内容をもとに、例規の類似部分探索タスクを定義する。4 節では、例規の類似部分探索手法を提案する。5 節では、提案手法を評価するために行った実験の詳細とその結果について報告する。6 節では関連研究を紹介し、提案手法との違いについて説明する。7 節は本稿のまとめであり、今後の課題についても述べる。

2 例規文書の特性

本節では、例規文書の構造や例規文書に特有の性質について説明する。

2.1 階層構造

例規は国の法律や政令などの法令と同じ文書構造を有している。例規は基本的には、題名、その例規が定める内容を記述した本則、改正や施行日などの付随的な情報を集めた附則から構成される。本稿では、これらのうち本則のみを扱い、これ以降、たんに例規という場合、本則のことを指すこととする。

(深夜営業施設への入場の禁止等)	
第十七条の二 次に掲げる施設において営業を営む者及びその代理人、使用人その他の従業者は、深夜においては、当該施設に青少年を立ち入らせてはならない。	
一	個室を設けて当該個室において客に専用装置による伴奏音楽に合わせて歌唱をさせる施設
二	設備を設けて客に主に図書類の閲覧、視聴若しくは聴取又はインターネットの利用を行わせる施設(図書館法(昭和二十五年法律第百十八号)第二条第一項に規定する図書館を除く。)
2	前項各号に掲げる施設において営業を営む者は、深夜において当該営業を営む場合は、入場しようとする者の見やすい箇所に、深夜における青少年の入場を禁ずる旨を掲示しなければならない。
条	項 号

図 2 条・項・号の例 (愛知県青少年保護育成条例 第 17 条の 2)

例規は、階層を有する木構造の文書である。各階層は上位のものから編、章、節、款、目、条、項、号と呼ばれる。上位の構造は下位の構造をその内容ごとにまとめるものである。また、号はその細分を再帰的に持つことができる。

図 2 は条・項・号の例である。この図のように、例規のテキストは項のテキストと号のテキストから構成されている。項は例規文書を構成する基本単位の構造であると考えられ、項のテキストは図 2 のように基本的には通常の 1~2 文で記述される。また、号は項の中で複数の事柄を列挙するとき用いられる。号のテキストは、通常の文や体言止めの文で記述される。本稿では、項の各文と号の各文を例規文と定義し、その全体集合を $SENT \subseteq \Sigma^+$ (Σ はアルファベット全体の集合) とする。

本稿では、例規を順序木 $T = \langle V, S, E, r \rangle$ と形式化する。ここで、 V は内部ノード集合、 S はテキストノード集合、 $E \subseteq (V \cup S)^2$ はエッジ集合、 $r \in V$ はルートノードを表す。内部ノード集合 V は例規の全体を表すルートノードと、例規の部分(条や項など)を表すノードからなる集合である。テキストノード $s \in S$ は項と号のテキストを表すノードで、対応する項や号を表す内部ノード $v \in V$ の子ノードとなる。ここで、項と号のテキストは複数の例規文から構成される場合もあるため、テキストノードとテキストの間のマッピングを $text : S \rightarrow SENT^+$ とする。 $V \cup S$ の各ノードを行きがけ順で走査した場合の各ノードの出現順序を順序関係と考え、 $V \cup S, V, S$ はそれぞれ全順序集合となる。 $e \in E$ は T のエッジを表す。また、ある内部ノード $v \in V$ について、 v をルートとした部分木を $\tau(v) = \langle V(v), S(v), E(v), v \rangle$ とする。

T の各ノードに対して、以下の関数を定義する。

- 子ノードからなる集合 $children : V \rightarrow 2^{V \cup S}$
- 親ノード $parent : V \cup S \rightarrow V$
- 自身と祖先ノードからなる集合 $anc : V \cup S \rightarrow 2^{V \cup S}$
- 自身と子孫ノードからなる集合 $des : V \cup S \rightarrow 2^{V \cup S}$

2.2 例規文の類似性

例規文には、互いに類似したものが多く存在する。1,580 自治体の 678,322 例規から抽出した延べ 20,861,033 文を対象とした例規文の出現数の調査 [4] によると、数千回出現するような例規文が多数見られる。表 1 は、そのうち、出現数が多い上位 5 文である。第 1 位や第 5 位の文は、例規の種類に関わらず出現する定型文である。一方、第 3 位の文は、特定の種類の例

表 1 例規文のうち出現数上位 5 文 [4]

順位	出現数	例規文
1	34,512	この条例の施行に関し必要な事項は、規則で定める。
2	25,207	削除
3	17,334	基金に属する現金は、金融機関への預金その他最も確実かつ有利な方法により保管しなければならない。
4	14,870	委員の任期は、2 年とする。
5	14,574	この規則に定めるもののほか、必要な事項は、市長が別に定める。

規にしか出現しない長い文であるが、多数出現している。なお、第 2 位の「削除」とはその部分が改正により削除されたことを表す特殊な表記である。

文献 [4] では、この調査結果から例規文の出現数 f とその順位 r の間に以下のべき乗則が成立することを報告している。

$$f \cdot r^\alpha = k \quad (\alpha \simeq 1, k \text{ は定数})$$

これは Zipf の法則であり、単語の出現数とその順位の間には一般に成り立つことが知られている。しかし、文において成り立つ例は例規文以外には報告されていない。なお、文献 [4] のコーパスでは、 $\alpha = 0.615, k = 4.14 \times 10^4$ の近似曲線が得られた。

このような性質は、例規の作成において類似例規を参照していることに起因すると考えられる。例規文を作成する際に、他の例規の条文が参照されると、それと類似した表現の例規文が作られる。さらに、参照によって作成された例規も再び参照の対象となりえる。このように、参照が自治体の間で繰り返されるうちに、多くの例規の間で表現が類似するようになる。その結果、類似した例規文が次々と作成され、多くの例規文が類似することになる。

また、項や号のテキストはその多くが単一の例規文であることから、項や号についてもテキストが一致するものが多いといえる。

2.3 類似部分間の差異

例規の類似部分どうしは目的が共通しているが、それぞれを含む例規どうしのテキストを比べると、テキスト全体が類似するとは限らない。例えば、景観条例の一種である宮城県七ヶ宿町街なみ景観条例と茨城県阿見町景観条例は、ともに景観に関する審議会について定めている。しかし、七ヶ宿町と阿見町の審議会には組織として異なる部分があり、それぞれの景観条例の当該箇所にも違いがある。具体的には、阿見町では審議会の会長や副会長の任命、専門部会の設置について定めているが、七ヶ宿町ではこれらの内容を定めていない。

また、類似部分の内容に差異がある場合、その部分の構造も異なることがある。前述の景観条例の例では、七ヶ宿町街なみ景観条例の第 12 条（審議会の組織）（全 3 項）で定められている内容は、阿見町景観条例の第 26 条（構成等）と第 27 条（委員の任期）に分けて定められている。

3 例規の類似部分探索タスク

本節では、例規の類似部分探索タスクを、二つの例規

$T_1 = \langle V_1, S_1, E_1, r_1 \rangle, T_2 = \langle V_2, S_2, E_2, r_2 \rangle$ から、例規の類似部分のペア $\langle v_1, v_2 \rangle \in V_1 \times V_2$ と対応する部分木のペア $\langle \tau(v_1), \tau(v_2) \rangle$ をすべて探索するタスクとして定義する。なお、本稿では簡略化のために部分木 $\tau(v_1)$ と $\tau(v_2)$ のペアをたんに $\langle v_1, v_2 \rangle$ と表す。

部分木のあるペアが互いに例規の類似部分どうかどうかを判定するために利用できる情報には、テキストと構造がある。例規の類似部分は表現が類似していることから、テキストの類似度は例規の制定目的の共通性と相関があるといえる。一方、例規の類似部分であってもその構造には多様性があるため、例規の類似部分であることと構造の類似性に関係があるとは必ずしもいえない。そこで、本稿では、テキストノードを用いた例規の類似部分探索を行う。

まず、例規 T_1 の部分木 $\tau(v_1)$ と例規 T_2 の部分木 $\tau(v_2)$ の間の類似度を、それぞれのテキストノード集合 $S(v_1) \subseteq S_1, S(v_2) \subseteq S_2$ を用いて計算する。ここで、この類似度を計算する関数を $sim_v : 2^{S_1} \times 2^{S_2} \rightarrow \mathbf{R}$ ($\mathbf{R}$ は実数全体の集合) と定義する。

sim_v を計算する際には、そのテキスト全体を結合して比較するのではなく、テキストノード単位で比較することが適当であると考えられる。これは、例規の構造の単位である項や号ごとに比較することにより、内容的な違いを類似度に反映しやすくなると考えられるからである。よって、文字列間類似度を $sim_s : \Sigma^+ \times \Sigma^+ \rightarrow \mathbf{R}$ とすると、 sim_v は、 sim_s をさまざまなテキストノードどうしで計算し、その結果を統合して部分木間の類似度を求める関数となる。なお、 sim_s には文字列間類似度を計算する任意の関数を用いることができる。

以上のことから、例規の類似部分探索タスクを形式化する。 sim_v の閾値を θ_v とすると、求める部分木ペアの集合を $M(T_1, T_2)$ は以下の式で表すことができる。

$$\begin{aligned} M(T_1, T_2) \\ = \{ \langle v_1, v_2 \rangle \mid v_1 \in V_1, v_2 \in V_2, sim_v(S(v_1), S(v_2)) \geq \theta_v \} \end{aligned}$$

ただし、この定義では部分木ペア $\langle v_1, v_2 \rangle$ に対して、任意の $d_1 \in des(v_1), d_2 \in des(v_2)$ をそれぞれルートノードとした部分木ペア $\langle d_1, d_2 \rangle$ も同時に得られる可能性がある。このペアは冗長であるため、出力に含まないようにする。そこで、まず、部分木ペアの間に、以下の半順序関係 $\preceq$ を定義する。

$$\langle v_1, v_2 \rangle \preceq \langle v'_1, v'_2 \rangle \iff v'_1 \in anc(v_1) \text{ かつ } v'_2 \in anc(v_2)$$

ただし、 $v_1 = v'_1$ かつ $v_2 = v'_2$ の場合を $\langle v_1, v_2 \rangle = \langle v'_1, v'_2 \rangle$ とする。また、 $\langle v_1, v_2 \rangle \preceq \langle v'_1, v'_2 \rangle$ かつ $\langle v_1, v_2 \rangle \neq \langle v'_1, v'_2 \rangle$ の場合を $\langle v_1, v_2 \rangle \prec \langle v'_1, v'_2 \rangle$ とする。

この半順序を用いて、類似部分探索タスクの出力 $M_{max}(T_1, T_2)$ を以下のように定義する。

$$\begin{aligned} M_{max}(T_1, T_2) = \{ m \in M(T_1, T_2) \mid \\ \text{任意の } m' \in M(T_1, T_2) \text{ に対して, } m \not\prec m' \} \end{aligned}$$

4 提案手法

本節では、例規の部分木に対して、二つの類似度 MTR と

ILINS を提案し、また、それらを用いた類似部分探索手法も提案する。この手法では、MTR と ILINS の線形和を類似度 sim_v として $M_{max}(T_1, T_2)$ を求める。

4.1 MTR

MTR (Matching Text-node Rate) は、“テキストノードのマッチが多ければ類似部分である可能性が高い”という直観に基づいた類似度である。ここで、テキストノードのマッチとは、テキストノード間の類似度が閾値 θ_s 以上であることを指す。

二つのテキストノード集合 S_1, S_2 について、 S_1 のテキストノードのうち、 S_2 のテキストノードのどれかとマッチしたテキストノードの集合を以下のように定義する。

$$S_{match}(S_1, S_2) = \{s_1 \in S_1 \mid \text{ある } s_2 \in S_2 \text{ に対して, } sim_s(text(s_1), text(s_2)) \geq \theta_s\}$$

これを用いて、 $MTR: 2^{S_1} \times 2^{S_2} \rightarrow \mathbf{R}$ を以下のように定義する。

$$MTR(S_1, S_2) = \sqrt{\frac{|S_{match}(S_1, S_2)|}{|S_1|} \cdot \frac{|S_{match}(S_2, S_1)|}{|S_2|}}$$

上述の直観に従った類似度としては、可能なテキストノードのペアに対するマッチしたテキストノードのペアの割合も考えられる。しかし、この類似度には、図 3 の A, B のような場合を区別できないという問題がある。すなわち、二つの部分木 $\tau(a)$, $\tau(b)$ を比較するとき、A, B は、テキストノードの異なるマッチングを表している。図 3 は、二つの部分木 $\tau(a)$ と $\tau(x)$ の比較を、テキストノードのマッチングが異なる A, B の 2 パターンについて表した図である。橙色で塗りつぶされた四角形はマッチしたテキストノードであり、破線はテキストノードどうしのマッチを表している。A と B のマッチの数は同じであるが、A の場合は $\tau(a)$ と $\tau(x)$ とが類似した部分木どうしであると考えられる。一方、B の場合は $\tau(a)$ と類似する部分木は $\tau(x)$ ではなく $\tau(z)$ であると考えられる。そこで、B の場合の $\tau(a)$ と $\tau(x)$ の類似度が低くなるように、テキストノードのペアではなく、テキストノードを単位とした割合を用いた。

また、比較する二つの木の間のテキストマッチの割合に対して、それらの相乗平均を求めている理由は、この割合がどちらか一方でも低い場合にペナルティを与えるためである。

4.2 ILINS

ILINS (Inversed Longest Isolated Node Sequence) は、マッチしていないテキストノードが連続して多く存在するとき、値が低くなる類似度である。図 4 に示す例では、 $\tau(d)$ のテキストノードのうち上から 3 番目から 9 番目は $\tau(p)$ と無関係であると考えられる。そこで、 $\tau(d)$ と $\tau(p)$ の類似度を低くすることにより、 $\tau(e)$ と $\tau(q)$, $\tau(g)$ と $\tau(r)$ のように分けた部分木ペアをそれぞれ出力できるようにする。

ILINS の計算では、まず LINS (Longest Isolated Node Sequence) をアルゴリズム 1 によって計算する。このアルゴリズムでは、入力した木のテキストノードのうち、マッチして

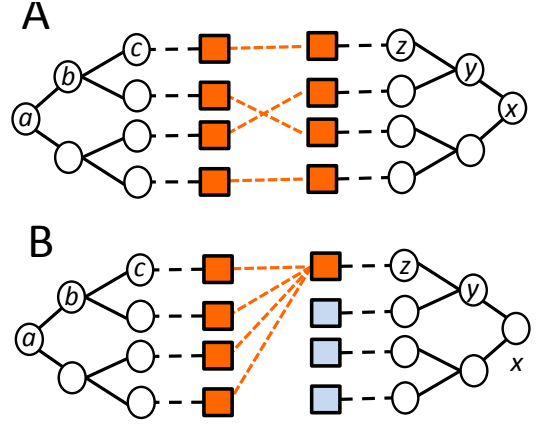


図 3 同じ数のテキストマッチでも MTR の値が異なる例

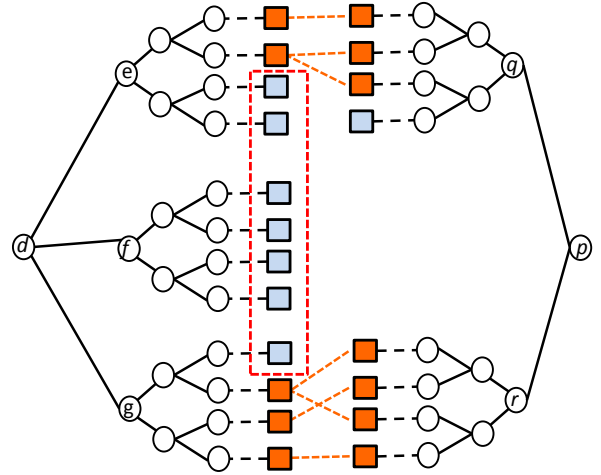


図 4 ILINS によって類似度を下げたい例

いないものが連続する系列の長さをカウントする。なお、マッチしているテキストノードが一つもない場合は、LINS の値は ∞ となる。次に、これらの関数を用いて、以下のスコア関数 $ILINS: 2^{S_1} \times 2^{S_2} \rightarrow \mathbf{R}$ を定義する。

$$ILINS(S_1, S_2) = \sqrt{\frac{1}{LINS(S_1, S_2) + 1} \cdot \frac{1}{LINS(S_2, S_1) + 1}}$$

MTR と同様に、部分木の間で LINS の引数を入れ替えた値の逆数の相乗平均を ILINS とした。

4.3 探索アルゴリズム

探索では、比較している二つの例規のそれぞれを走査し、得られた部分木のペアについて、類似部分かどうかを以下の sim_v により判定する。

$$sim_v(S_1, S_2) = w_M \cdot MTR(S_1, S_2) + (1 - w_M) \cdot ILINS(S_1, S_2)$$

ただし、 $w_M (0 \leq w_M \leq 1)$ は MTR の重みである。

探索アルゴリズムをアルゴリズム 2 に示す。このアルゴリズムは、マッチするテキストノード $\langle s_1, s_2 \rangle \in S_1 \times S_2$ を起点として、すべてのノードのペア $\langle v_1, v_2 \rangle \in anc(s_1) \times anc(s_2)$ に

アルゴリズム 1 LINS の計算アルゴリズム

Input: テキストノード列 S_1, S_2 **Output:** $n \in \mathbb{Z}$

```
1: procedure LINS( $S_1, S_2$ )
2:    $counter \leftarrow 0$ 
3:    $n \leftarrow -\infty$ 
4:   for each  $s_1 \in sort(S_1)$  do
5:     if ある  $s_2 \in S_2$  に対して  $sim_s(s_1, s_2) \geq \theta_s$  then
6:        $n \leftarrow max(n, counter)$ 
7:        $counter \leftarrow 0$ 
8:     else
9:        $counter \leftarrow counter + 1$ 
10:  return  $|n|$ 
```

アルゴリズム 2 例規の類似部分探索アルゴリズム

Input: $T_1 = \langle V_1, S_1, E_1, r_1 \rangle, T_2 = \langle V_2, S_2, E_2, r_2 \rangle$ **Output:** $M_{max} = \{ \langle v_1, v_2 \rangle \mid v_1 \in V_1, v_2 \in V_2, \text{sim}_v(S(v_1), S(v_2)) \geq \theta_v \}$

```
1: procedure MATCHING( $T_1, T_2$ )
2:    $M_{max} \leftarrow \emptyset$ 
3:    $M_s \leftarrow \{ (s_1, s_2) \in S_1 \times S_2, \mid \text{sim}_s(text(s_1), text(s_2)) \geq \theta_s \}$ 
4:   for each  $(s_1, s_2) \in M_s$  do
5:      $A \leftarrow \{ \langle v_1, v_2 \rangle \mid v_1 \in anc(s_1), v_2 \in anc(s_2) \}$ 
6:     for each  $\langle v_1, v_2 \rangle \in A$  do
7:       if ある  $\langle v'_1, v'_2 \rangle \in M_{max}$  に対して,
          $\langle v_1, v_2 \rangle \preceq \langle v'_1, v'_2 \rangle$  then
8:         continue
9:       if  $\text{sim}_v(S(v_1), S(v_2)) < \theta_v$  then
10:        continue
11:        $M \leftarrow M_{max} \cup \{ \langle v_1, v_2 \rangle \}$ 
12:        $M_r \leftarrow \{ \langle v'_1, v'_2 \rangle \in M \mid \langle v'_1, v'_2 \rangle \prec \langle v_1, v_2 \rangle \}$ 
13:        $M_{max} \leftarrow M \setminus M_r$ 
14:  return  $M_{max}$ 
```

対して、対応する部分木ペアの類似度を計算する。これにより、類似度が 0 となる部分木ペアの類似度の計算を省略する。

5 実験

提案手法を用いたクラス推定実験を行い、類似例規のペアから類似部分が検索できることを確認した。本節では、実験の設定と、その結果について報告する。

5.1 実験設定

ここでは、実験の設定について説明する。

5.1.1 実験タスク

本実験では、例規の類似部分を用いた条のクラス推定を行った。ここで、クラスとは、例規が定める事項の分類と定義する。地方自治体の政策革新についての研究 [5] の中では、景観条例

の定める内容を景観保護地区や審議会、立ち入り調査などに分類しており、これらがクラスに該当する。文献 [5] では、各景観条例に対して、各クラスと対応する内容を含む/含まないの 2 値からなる特徴ベクトルを作成し、日本全国の景観条例に対する各種の統計的な分析を行った。

これらのクラスは、例規の制定目的を実現するための例規の部分で構成すると考えられる。つまり、類似部分が獲得できれば、それらは同じクラスに属するとみなすことができる。このことを利用し、その各条のクラスが既知である例規（クエリとする）を用いて、その各条のクラスが未知である例規（ターゲットとする）の各条のクラス推定を行う。

この実験では、クエリの部分木 $\tau(v_{query})$ とターゲットの部分木 $\tau(v_{target})$ のペアを提案手法により獲得し、それにより、 v_{target} のクラスが v_{query} のクラスであると推定する。ただし、複数のクラスと対応する v_{target} が存在するため、この実験では木のクラスの集合を推定する。また、出力が条どうしのペアでなかった場合、以下のような例外処理を行う。

- v_{query} と v_{target} の一方または両方が条よりも下位の階層であった場合、そのペアを出力から除く。
 - v_{query} が条よりも上位の階層であった場合、 $\tau(v_{query})$ のすべての条のクラス集合の和集合を v_{target} の予測クラス集合とする。
 - v_{target} が条よりも上位の階層であった場合、 $\tau(v_{target})$ のそれぞれの条の予測クラス集合を v_{query} のクラス集合とする。
- ある v_{target} について、得られる部分木のペア $\langle v_{query}, v_{target} \rangle$ は一つとは限らないため、部分木ペアそれぞれについてクラス推定を行い、それらすべての和集合を v_{target} の最終的な予測クラス集合とする。

5.1.2 データセット

データセットとして、表 2 に示す 6 件の条例を用意し、 Q をクエリ、 $T_1, T_2, \dots, T_5$ をターゲットとした。用意した条例データは、eLen 条例データベース [6]¹ に格納されている XML 文書データである。なお、これらの条例は、[5] の分析が実施された時点（2004 年 3 月）から eLen 条例データベースに格納された時点（2013 年 7 月）までの期間に改正されていないものである。

クエリとターゲットの両方のすべての条に対して、文献 [5] を参照してクラスを付与した。なお、この文献には条とクラスの間に対応関係が明示されているわけではなく、ある条例に対して、それがクラスの内容を含んでいるかという情報しか記されていない。そこで、筆者が各条例の条とクラスの対応づけを行った。その際、どのクラスにも当てはまらない条や判断がつかない条が存在したため、本実験ではこれらの条を「その他」というクラスに分類した。

このようなクラス分類によって得られた各条のクラスを表 3 に示す。表中の数値は条例の条名である。ただし「23-30」のようにハイフンで結んだものは、第 23 条から第 30 条までのように条名の範囲を表す。 T_3 の第 16 条のように、複数のクラス

1 : <https://elensv.e-legislation.jp/>

表 2 実験用データセット

条例	自治体	題名	制定年
Q	茨城県 阿見町	景観条例	1999
T_1	宮城県 七ヶ宿町	街なみ景観条例	2000
T_2	山形県 山形市	美しい山形をつくる基本条例	1988
T_3	山形県 山形市	景観条例	1996
T_4	福島県 大玉村	ふるさと景観保護条例	1996
T_5	三重県 御浜町	熊野参詣道伊勢路景観保護条例	2002

表 3 実験用データセットのクラス分類

クラス	条例の条名					
	Q	T_1	T_2	T_3	T_4	T_5
景観地区		5	8	9		5, 6
地区基準	12	6		10		
届出指導	13-15	7-9	10	11-13, 16, 17, 19	6, 7	9, 12
許可承認						8
立入調査						11
禁止						7
違反公表	16				9	
大規模規制				16, 17		
大規模基準				15		
阻害指導						10
重要建築	11			18-20		
樹木指定	7					
協定認定				22		
首事協定			9			
助成制度	19-21	10	10	20, 25	10	
表彰制度	18			24		
審議会	23-30	11, 12	12		4	
その他	1-6, 8-10, 17, 22, 31	1-4, 13	1-7, 11, 13, 14	1-8, 14, 21, 23, 26, 27	1-3, 5, 8, 11	1-4, 14
条の合計数	31	12	14	27	11	14

に分類できる場合もあるので、各条は一つ以上のクラスに分類できる。

5.1.3 評価対象

本実験では、提案手法と article 手法の二つの手法を評価し、比較した。ここで、article 手法は、提案手法に対して類似部分の探索を条より上の階層では行わないという制約を与えた手法である。この手法は、提案手法と比較したとき、実験の評価基準について、「正解が条どうしの部分木ペアである」という事前知識を有しているため、すべての階層を対象としている提案手法と比較して有利な手法である。特に、推定の適合率が提案手法よりも高くなると予想できる。本実験では、この手法をベースラインとして提案手法を評価することにより、階層に依存しない手法が特定の階層に特化した手法に対してどの程度近づけることができるのかを評価する。

表 4 F2 が最大のときの結果とパラメータ (マクロ平均・brave)

手法	w_M	θ_v	θ_s	sim_s	適合率	再現率	F2
提案手法	0.8	0.4	0.45	tf	0.51	0.79	0.71
article 手法	0.8	0.4	0.35	tf	0.34	0.81	0.64

これら二つの手法それぞれに対し、各手法のパラメータ $sim_s, w_M, \theta_s, \theta_v$ を変更したうえで、比較を行った。まず、 sim_s には以下の三つを用いた。

- **jaccard** 比較する 2 テキストをそれぞれ文字 3-gram の集合に変換したとき、その間の Jaccard 係数。

- **tf** 比較する 2 テキストを tf 値 (単語の出現頻度) によってベクトル化したとき、その間の cosine 類似度。

- **tfidf** 比較する 2 テキストを tfidf 値 (単語の出現頻度と逆文書頻度の積) によってベクトル化したとき、その間の cosine 類似度。

テキストノードのマッチングについても、それぞれ別々の方法で獲得した。jaccard では、閾値以上の類似度のテキストを高速に検索することができる SimString [7] を利用した。tf と tfidf では、k-NN 法を用いて、類似する上位 500 テキストを検索し、その中から閾値 θ_s によるフィルタリングを行った。このときのテキストノードの探索空間は、ターゲットを定める 4 自治体の例規のうち 1,302 例規のもつ 42,905 テキストノードのテキストである。

数値パラメータ w_M, θ_v, θ_s については、以下の値を用いた。

- $w_M = \{0, 0.2, 0.4, \dots, 1.0\}$
- $\theta_v = \{0.1, 0.2, 0.3, \dots, 0.7\}$
- $\theta_s = \{0.25, 0.3, 0.35, \dots, 0.75\}$

5.1.4 評価指標

評価指標には適合率 (Precision), 再現率 (Recall), および再現率を重視した F2 値を用いた。この F2 値は、以下のよう計算する。

$$F2 = \frac{(1 + 2^2) \cdot Precision \cdot Recall}{2^2 \cdot Precision + Recall}$$

再現率を重視した理由は、本稿では既存の手法で獲得できなかった類似部分を網羅的に獲得することを目的としているためである。

これらのスコア計算において、「その他」クラスであるという推定が正しかった場合に、それを正解とする指標 (brave) と、正解とも不正解ともしないで取り除く指標 (skeptical) の両方をそれぞれ評価した。これは、「その他」クラスどうしのペアには、本来は正解とすべきペアとそうでないペアのどちらも出現しえるためである。本実験では、brave と skeptical の両方について、マイクロ平均とマクロ平均の両方の場合のスコアを計算し、この 4 通りそれぞれについて、F2 値が最大となった場合の提案手法と article 手法を比較した。

5.2 実験結果

実験の結果を表 4~7 に示す。brave では、マイクロ平均・マクロ平均の両方において、提案手法の F2 値が article 手法の

表 5 F2 が最大のときの結果とパラメータ（マイクロ平均・brave）

手法	w_M	θ_v	θ_s	sim_s	適合率	再現率	F2
提案手法	0.2	0.3	0.45	tf	0.41	0.74	0.64
article 手法	0.8	0.4	0.35	tf	0.29	0.77	0.58

表 6 F2 が最大のときの結果とパラメータ（マクロ平均・skeptical）

手法	w_M	θ_v	θ_s	sim_s	適合率	再現率	F2
提案手法	1.0	0.5	0.35	tfidf	0.37	0.83	0.67
article 手法	1.0	0.4	0.35	tfidf	0.39	0.83	0.68

表 7 F2 が最大のときの結果とパラメータ（マイクロ平均・skeptical）

手法	w_M	θ_v	θ_s	sim_s	適合率	再現率	F2
提案手法	0.2	0.3	0.5	tf	0.38	0.70	0.60
article 手法	0.6	0.3	0.5	tf	0.47	0.70	0.64

F2 値を上回った．一方，skeptical では逆に提案手法の F2 値が下回る結果となった．また，どの場合も w_M 以外の各数値パラメータの変動は 0.2 以下であり， sim_s も tf が $3/4$ を占めている．

skeptical の場合に提案手法の F2 値が article 手法のそれを下回るような結果となったのは，階層に制限が無いため，必要以上に大きな部分木どうしのペアを獲得してしまい，それによって適合率が低下したことが原因である．ただし，F2 値の差はそれぞれ 0.01, 0.04 と小さいため，提案手法は，事前知識なしでも事前知識ありの手法と近いスコアを出することができる手法といえる．

5.3 ケーススタディ

ここでは，提案手法により得られた阿見町景観条例（表 2 の条例 Q ）と七ヶ宿町街なみ景観条例（表 2 の条例 T_1 ）の審議会クラスの部分木ペアを対象とした分析の結果を報告する．

マイクロ平均・指標 skeptical にて F2 値が最大となった場合の提案手法の結果のうち，審議会のクラスに分類されたものを表 8 に示す．条例 Q の第 3 章には第 23 条～第 30 条が含まれるため，表 3 より，この結果は両方の条例の審議会クラスの条をすべて抽出できた結果といえる．

この結果を分析するため，表 9 を作成した．これは，条例 Q と条例 T_1 のうち，審議会クラスに属する条に含まれる項を対象として，同じ目的で定められた項どうしを手対で対応させて作成したものである．なお，項 ID は条名と項番号をピリオドで結んで便宜的に付けたものであり，例えば「23.1」は「第 23 条第 1 項」を表す．また，2 列目と 4 列目はその条の条見出しと項の内容を要約したものである．ただし，ある項について，その親の条に他の項が含まれない場合は要約を省略した．

提案手法により，条例 T_1 の第 11 条は 条例 Q の第 24 条と第 25 条の両方に対応するという出力が得られたが，表 9 から，同じ内容を別の構造によって記述した例であることがわかる．この結果より，構造的な類似性を見逃すというタスク設定が

表 8 提案手法により得られた審議会クラスの部分木ペア（ Q と T_1 ）

阿見町景観条例（条例 Q ）	七ヶ宿町街なみ景観条例（条例 T_1 ）
第 3 章第 24 条	第 11 条
第 3 章第 25 条	第 11 条
第 3 章	第 12 条

例規の類似部分探索において有効であることが確かめられた．

また，条例 T_1 の第 12 条が条例 Q の第 3 章に対応するという出力が得られた．正解では第 26 条・第 27 条に対応するものであり，これを包含する結果となっている．このとき同時に得られた条のうち，第 28 条～第 30 条の項や号は，第 26 条・第 27 条と同じ審議会クラスに属する．さらに，条例 T_1 の第 11 条・第 12 条のどちらの項や号とも類似しないため，テキスト情報だけでは容易に検索できない部分である．実際，article 手法の結果では，第 24 条～第 27 条しか獲得できていない．この結果から，特定の階層に依存しないような手法にすることにより，汎用性だけでなく，再現率も向上させることができるということが確かめられた．

一方，今回の実験の評価スコアには影響していないが，表 8 には提案手法の失敗例も見られる．条例 T_1 の第 11 条と条例 Q の第 23 条との部分木ペアは，表 9 では対応しているが，提案手法では獲得できていない．条例 Q の第 23 条第 1 項は条例 T_1 の第 11 条第 1 項のテキストとマッチしているが，それだけでは sim_v が 0.58 となり，閾値 $\theta_v (= 0.60)$ により切り捨てられる．一方，第 24 条・第 25 条の場合は，それらのテキストが類似していたため，表 9 にあるような項どうし以外にもテキストがマッチし，結果的に sim_v が θ_v を超えた．この結果は，一つの条と複数の条が対応するなどの構造的な多様性に対して，提案手法が対応しきれていないという問題を示唆している．

6 関連研究

本節では，本稿と関連した二つの研究について説明する．

6.1 例規の比較支援

竹中ら [8] は，2 例規の間で，テキストが類似した条のペアを探索する手法を提案した．この手法では，一方の例規の条のテキストに対して，他方の例規のすべての条のテキストとの間でそれぞれ類似度を計算し，類似度がもっとも高いペアを出力するという手順を両方の例規のすべての条に対して繰り返す．

本稿ではすべての階層を対象としたのに対して，竹中らは条に限定した手法を提案したという点で異なる．

6.2 極大類似部分木探索

類似部分の探索と近いタスクに取り組んだ研究として，寺島ら [9] の極大類似部分木探索の研究がある．極大類似部分木探索タスクは，二つの XML 文書データ中の類似した部分木ペアのうち極大なものを探索するタスクであり，例規の類似部分探索タスクと類似したタスクであるといえる．ただし，極大類似部分木であることの必要条件は，例規の類似部分のようにテキ

表 9 審議会クラスの項単位での対応表 (Q と T_1)

阿見町景観条例 (条例 Q)		七ヶ宿町街なみ景観条例 (条例 T_1)	
項 ID	条見出し (内容)	項 ID	条見出し (内容)
23.1	審議会の設置	11.1	街なみ景観審議会 (設置)
24.1	審議事項	11.2	街なみ景観審議会 (審議事項)
25.1	助言	11.3	街なみ景観審議会 (助言)
26.1	構成等 (委員の人数)	12.1	審議会の組織 (委員の人数)
26.2	構成等 (委員の要件)	12.3	審議会の組織 (委員の要件)
26.3	構成等 (臨時委員)		
27.1	委員の任期 (基本)	12.2	審議会の組織 (委員・補欠委員の任期)
27.2	委員の任期 (補欠)		
27.3	委員の任期 (臨時委員)		
28.1	会長及び副会長 (選出方法)		
28.2	会長及び副会長 (会長)		
28.3	会長及び副会長 (副会長)		
29.1	会議 (召集・議長)		
29.2	会議 (成立・議決)		
30.1	専門部会		

ストが類似していることだけではない。それに加えて、構造が類似していること、また、極大類似部分木ペアの両方の部分木の子どうしもまた類似部分木となる必要がある。この点で、本稿で定義した例規の類似部分探索タスクとは異なる。

7 ま と め

本稿では、類似例規検索のサブタスクとして例規の類似部分探索タスクを部分木探索タスクとして定義した。また、MTR と ILINS の 2 種類の類似度と、これらを用いた例規の類似部分探索手法を提案した。この手法では、例規の任意の階層の間での類似部分が探索可能である。実験では、提案手法を用いて類似部分木探索によるクラス分類タスクを実施し、マクロ平均で適合率 0.37 ・再現率 0.83 ・ F_2 値 0.67 、マイクロ平均で適合率 0.38 ・再現率 0.70 ・ F_2 値 0.60 という結果を得た。また、提案手法の出力例を分析することにより、自治体間の制度の違いによる類似部分間の差異に対して頑健な手法であることを示した。これらの結果より、提案手法の有効性を示した。

今後は、メインタスクである類似例規検索の実験を実施し、既存の例規検索手法との比較を行う。また、本手法は例規の構造的な多様性や表現の類似性などの性質を活用したものであり、同様の性質を持つほかのドメインへの利用が期待できる。そのようなドメインには、製品のマニュアルや利用規約、法人の定款、契約書などが考えられる。これらのドメインについての調査を行い、提案手法が例規以外を対象としたタスクに対して活用可能かどうかについても検討する。

謝 辞

本研究の一部は JSPS 科研費基盤研究 (B)No.18H03492 の助成を受けたものである。

文 献

- [1] 竹中要一, 若尾岳志. 都道府県の類似例規の抽出と応用. 情報処理学会論文誌数理モデル化と応用 (TOM), Vol. 8, No. 1, pp. 80–86, 2015.
- [2] Kazuya Fujioka, Makoto Nakamura, Yasuhiro Ogawa, Tomohiro Ohno, and Katsuhiko Toyama. Search method for ordinances and rules in japanese local governments based on distributed representation. In *9th International Conference on Knowledge and Systems Engineering, Vietnam*, pp. 185–190, 2017.
- [3] 角田篤泰, 齋藤大地, 大谷忠. スーパーコンピュータを利用した自治体例規の形式的分析. 名古屋大学法政論集, Vol. 247, pp. 322–298, 2012.
- [4] 藤岡和弥, 駒水孝裕, 小川泰弘, 外山勝彦. Zipf の法則は例規文の出現数においても成立する. 平成 30 年度電気・電子・情報関係学会東海支部連合大会, No. L4-1, 2018.
- [5] 伊藤修一郎. 自治体発の政策革新: 景観条例から景観法へ. 木鐸社, 2006.
- [6] 角田篤泰. 自治体向け立法支援のための条例データベース eLen 情報処理学会第 76 回全国大会講演論文集, pp.4–456–4–457, 2014.
- [7] Naoaki Okazaki, Tsujii Jun'ichi. A Simple and Fast Algorithm for Approximate String Matching with Set Similarity. *Journal of Natural Language Processing*, Vol. 18, No. 2, pp. 89–117, 2011.
- [8] 竹中要一, 若尾岳志. 地方自治体の例規比較に用いる条文対応表の自動生成. 言語処理学会第 17 回年次大会発表論文集, pp. 520–523, 2011.
- [9] 寺島慎太郎, 天笠俊之, 北川博之. XML データにおける類似極大部分木の効率的な探索. In *DEIM Forum 2011*, No. E1-2, 2011.