地方自治体の子育て支援事業比較表作成ツールの開発

金田 大海　多田 紘佳　中村 誠
新潟工科大学 工学部
mnakamur@niit.ac.jp

概要

近年，少子高齢化が深刻化しており，それに伴い総人口が減少傾向にある．本研究の目的は，地方自治体（以降，自治体という.）の子育て支援事業の類似事業をまとめ、比較が行なえる表を出力するツールの開発である．これにより，自治体の独自事業の発見が期待できる．そのために本研究では，子育て支援事業の内容を自治体のウェブサイトから取得し，TF-IDF を用いた SVM による文書分類により，必要項目と不要項目を分類した後，必要項目についてクラスタリングを行い，類似する事業をまとめ表を作成し出力するツールを開発する．

1 はじめに

近年，少子高齢化が深刻化しており，2015 年には 65 歳以上が約 26.6%、15 歳未満が約 12.5%であったのが、2045 年には 65 歳以上が約 36.8%に増加、15 歳未満が約 10.7%と減少するとされている。それに伴い総人口が減少傾向にある [1]. 特に, 少子高齢化による人口減少は小規模な自治体ほど顕著に表れている．また，少子高齢化により「社会保障分野の負担の増大と手取り収入の減少」や「生産年齢人口減少による経済成長の低下」などの問題が発生している [2]. これらの問題に共通する原因は, 生産年齢人口の減少である．このことから，少子高齢化社会を改善していくには、人口減少の多い自治体で生産年齢人口の減少を止めていかなければならない．そこで現在，少子高齢化対策としてそれぞれ子育て支援事業などを用意して出生率増加を図っており，ウェブサイトで確認できる．ところが，子育て支援事業は各自治体のウェブサイトで独立して提示されているため，自治体間の比較が難しい．そこで，子育て支援事業を 1 つの表にまとめることで比較しやすくなると考えられる．

類似している事業の対応付けを行う手法の研究はすでに行われている [3]. この研究では, 自治体のウェブサイトの HTML 内にあるアンカーテキストに着目し，編集距離を用いて対応付けを行っている．ここから，比較結果として精度は高くないが類似事業の対応付けは有効であると分かる．

したがって，本研究の目的は，各自治体の子育て支援事業の類似事業の比較表（以降，「比較表」という。）を出力するツールの開発である. 利用者は，このツールを開発することで，各自治体が子育て支援事業にどれだけ力を入れているか可視化ができ，他の自治体にはない独自の事業を発見できる. また，各自治体は自身の自治体に足りない事業を発見できることから，事業の充実が期待できる．

本研究では，子育て支援事業の内容を自治体のウェブサイトの HTML から取得し，TF-IDF ベクトルを用いた SVM による文書分類で必要事業と不要事業を分類し，必要事業についてクラスタリングを行った結果を表にして出力するツールの開発を行う．

2 背景

ここでは，本研究の目的である比較表を出力するツールの必要性について示す．また，本研究で使用しているクローリングと web スクレイピング，SVM，クラスタリングについて説明する．

2.1 子育て支援事業

本研究で対象としている子育て支援事業とは，子育てに関係する行政事業の一群を指す．これらについて，全国の自治体でウェブサイトを用意し提示している．図 1 は，柏崎市のウェブサイトである．図 1 のように，多くの自治体では事業ごとにウェブページを用意しており，そのウェブページに事業の内容などが掲載されている．図 2 は，柏崎市の児童手当のウェブページである．中には，1 つのウェブページに複数の事業をまとめて掲載している自治体もある．このように，ウェブページに記載されているため，クローリングを用いることで事業の内容を収集することが可能である．

図 1　柏崎市の事業が提示されているウェブサイト

図 2　柏崎市の児童手当のウェブページ

2.2 比較表

本システムにおける比較用の例を表 1 に示す．新潟県新潟市の「安産教室」や岡山県奈義町の「出産祝金」は独自の事業であることと，岡山県奈義町には里親制度の事業が足りないことが一目で分かる．比較表作成における問題の本質は，自治体間において類似事業を発見することにある．

表 1　子育て支援事業比較表の例

新潟県柏崎市	新潟県新潟市	岡山県奈義町
妊産婦の医療費を助成します	妊産婦医療費助成	乳幼児及び児童生徒医療費助成
未熟児養育医療給付制度のお知らせ	不育症治療費助成事業	不育治療支援事業
	安産教室	
里親制度をご存知ですか	里親制度の推進について	
		出産祝金

2.3 クローリングと web スクレイピング

クローリングとは，ウェブサイトから HTML などの情報を収集することであり，web スクレイピングとは，クローリングにより web サイトから取得した HTML データを解析し，特定のデータを抽出することである．

本研究では，新潟県内の 23 の自治体と岡山県奈義町を対象に，子育て支援事業に関係すると思われる「妊娠出産」と「子育て」の 2 つのカテゴリーの事業を対象にそれらが書かれているウェブページのクローリングを行う．また，内容が書かれているウェブページをスクレイピングによってテキストファイルに書き出す．

2.4 SVM

SVM は，2 クラスのパターン識別器を構成する手法である．学習モデルから，「マージン最大化」という基準で線形しきい素子のパラメータを学習し，計算によって 2 値のクラスをつくる [4]。本研究では分類方法として「ソフトマージン」を用いる．

本研究では，SVM により子育て支援事業に関係している事業としていない事業に分類するために用いている．スクレイピングにより取得したテキストには，給付金や支援などの事業のみならず，図 3 のようにイベント等の一覧しか書かれていないものや表彰関係，おたよりや報告などと子育て支援のための情報提供には不要な，ウェブページが含まれている．

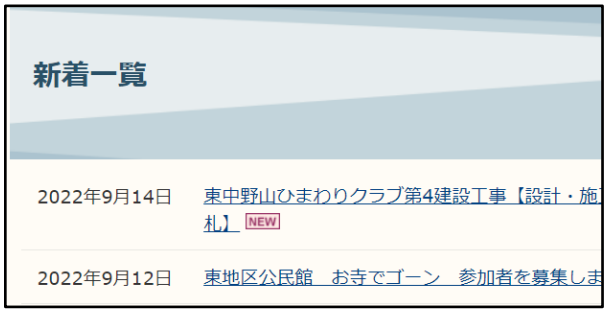


図 3　不要なウェブページの例

2.5 クラスタリング手法

本実験では階層的手法を用いる．階層的手法とは，最も類似度の高い組み合わせからまとめる手法である．階層的手法には，クラスタ間における距離の求め方が複数存在する．主な手法は，「群平均法」，「重心法」，「完全リンク法」，「メジアン法」，「単リンク法」，「ウォード法」，「重み付き平均法」である．

本研究では，各自治体におけるさまざまな子育て支援事業を対象に，TF-IDF でベクトル化し，それからコサイン類似度を計算する．その計算結果を用いてメジアン法でクラスタリングを行う．メジアン法とは，それぞれの重心にクラスタ内の個数に応じた重み付けを行い、それぞれのクラスタの重み付けした重心間距離の 2 乗をクラスタ間の距離とする手法である。

2.3 評価手法

比較表の精度については，purity と inverse purity の 2 つから求められる F 値で評価することが出来る[5]．図 4 に正解クラスタとクラスタリング結果の関係を示す．同じ図形は類似事業を示している．

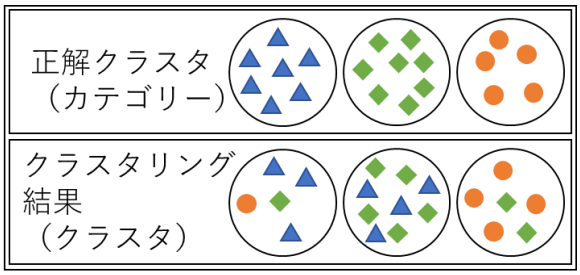


図 4　正解クラスタとクラスタリング結果の関係

Purity とは，図 4 下部の各クラスタにおいて最もよく現れるカテゴリーの出現頻度に注目し，クラスタ内においてそのカテゴリーが占める割合が大きいものを高く評価する評価尺度であり，式（2）で求まる．

$$Purity = \sum_i \frac{|C_i|}{n} maxPrecision(C_i, L_j) \quad (2)$$

このとき，C は評価対象とするクラスタ集合，L は人手で作成したカテゴリー集合，n はクラスタリング対象の文書数である．また，あるカテゴリーL_jに対するクラスタC_iの適合率は式（3）によって求められる．

$$Precision(C_i, L_j) = \frac{|C_I \cap L_j|}{|C_i|} \quad (3)$$

Inverse purity とは，図 4 上部の各カテゴリーに対して最大の再現率となるクラスタに注目し，クラスタ内において各カテゴリーで定められた要素を多く含むクラスタを高く評価する評価尺度であり，式(4)で求まる．

$$InversePurity = \sum_j \frac{|L_j|}{n} maxRecall(C_i, L_j) \quad (4)$$

このとき，あるカテゴリーに対するクラスタの再現率は式（5）によって求められる．

$$Recall(C_i, L_J) = \frac{|C_i \cap L_j|}{|L_j|} \quad (5)$$

また，purity と inverse purity の調和平均 F は式（6）により定義される．

$$F = \frac{1}{\alpha \frac{1}{Purity} + (1-\alpha)\frac{1}{InversePurity}} \quad (6)$$

本研究では$\alpha = 0.5$として評価を行った．

なお、図 4 における purity は 0.6，inverse purity は 0.65，F 値は 0.62 となる

3 提案手法

本研究の各自治体の子育て支援事業の比較表を作成するツールの流れは以下の通りである．

1. あらかじめ各自治体のウェブサイトから，事業内容が書かれている本文を抽出し，テキストファイルにする．
2. チェックボックスを用いたユーザーインターフェースで比較する自治体を選択する
3. 選択した自治体のテキストファイルを参照する
4. 参照したテキストファイルから，子育て支援事業に必要な項目と不要な項目を SVM を用いて分類する．
5. 必要項目に分類した方を，TF-IDF でベクトル化し，cos 類似度を用いたメジアン法のクラスタリングによって比較表を出力する．

このときの、チェックボックスを用いたユーザーインターフェースを図 5 に，出力結果を図 6 に示す．

類似している事業の対応付けを行う手法の研究[3]では，「編集距離」を用いているが比較表の精度は高くなかった．そこで，本研究では TF-IDF ベクトルを用いた cos 類似度によりクラスタリングを行うことで精度の向上を目指す．

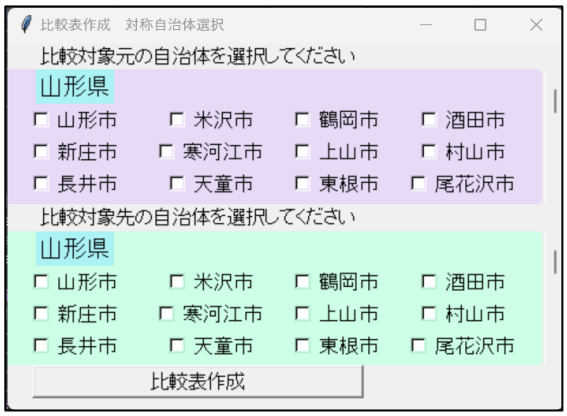
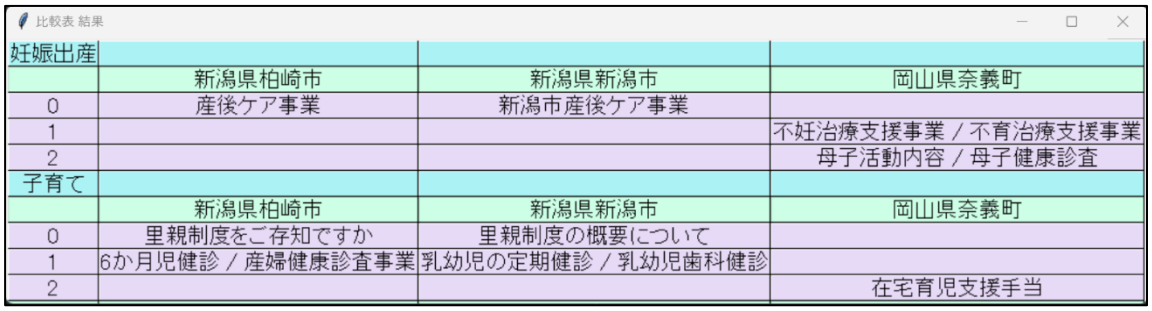


図 5　ユーザーインターフェース



妊娠出産			
	新潟県柏崎市	新潟県新潟市	岡山県奈義町
0	産後ケア事業	新潟市産後ケア事業	
1			不妊治療支援事業 / 不育治療支援事業
2			母子活動内容 / 母子健康診査
子育て			
	新潟県柏崎市	新潟県新潟市	岡山県奈義町
0	里親制度をご存知ですか	里親制度の概要について	
1	6か月児健診 / 産婦健康診査事業	乳幼児の定期健診 / 乳幼児歯科健診	
2			在宅育児支援手当

図 6　出力結果の一部

4 実験

4.1 実験の目的

文書分類を用いない比較表作成ツールにより，クラスタリングにメジアン法を用いて比較表を作成したところ，F 値は 0.371 となった．しかし，この比較表には子育て支援事業と関係のないものも含まれていた．そのため，不要項目を含むか否かで精度は変わると考えた．また，この比較表は異なる分野の事業が一緒になっていたため，分野別で比較表を作成することで見やすい表となり，精度も向上すると考えた．そこで，本実験の目的は，文書分類を用いて不要項目を分類し必要項目のみで比較表を作成することと，分野別で作成することで，クラスタリングの精度が向上するのか検証する．正解データは，不要な項目も考えたデータを新しく用いる．

4.2 実験方法

本実験では，新潟県柏崎市，新潟県新潟市，岡山県奈義町の 3 つの自治体を対象とし，不要項目の分類を行うか否かでモデルの性能を比較する．分野別については，「妊娠出産」と「子育て」の 2 つの分野を対象とし，それぞれで分類を行ったモデルの性能と文書分類を行ったモデルの性能で比較を行う．性能の比較には，評価手法を用いた精度の評価を行う．このとき，カテゴリー集合である正解データは，人手によって 3 つの自治体の子育て支援事業のみを対象に類似事業を対応付け，それ以外を不要な事業としてまとめた表を用いた．

4.3 実験結果

分類の有無について，3 つの自治体で作成した比較表の評価結果を表 2 に示す．1 行目は不要項目の分類を行っていない場合，2 行目は行った後のクラスタリングの結果を表している．

表 2　分類の有無による評価結果

	purity	inverse purity	F値
分類無し	0.467	0.357	0.371
分類有り	0.574	0.467	0.515
理想の表	0.725	0.647	0.684

表 3　分野別による評価結果

	purity	inverse purity	F値
分野別	0.472	0.642	0.544

また，分野別で作成した比較表の精度の評価結果を表 3 に示す．

比較表の精度について不要項目の分類を行うか否かで比較すると，F 値が約 0.144 向上した．これらにより，不要な項目がなくなることで，比較表の精度が向上したと考えられる．また，このとき作成した表は，不要な項目がなくなったため類似事業が分かりやすい表となった．参考までに，クローリングやスクレイピングが完全で，不要項目が存在しない場合の性能を表 2 の 3 行目に示す．

分野別の精度について，表 2 と表 3 の F 値から，分野別にすることで約 0.029 向上した．この値より，分野別にすることで SVM とクラスタリングの精度が向上したのではないかと考える．このとき作成された表は，分野別になっているため，分野ごとの事業が分かりやすい表となった．

5 おわりに

実験から，SVM を用いて分野別に作成することで，不要な項目を無くし，比較表の精度を向上させることが出来た．また，分野別にしたことで見やすい表となった．これらにより，子育て支援事業の比較表作成ツールを開発することが出来た．

今後は，SVM の教師データを増やすことや，比較表の精度のさらなる向上を目指す．

謝辞

本研究は，科学研究費補助金（19H04427，代表：中村 誠）の助成を受けたものである．

参考文献

[1] 河合雅司，未来の年表　人口減少日本でこれから起きること，講談社現代新書，2017
[2] 厚生労働省，少子化の影響と主な対策に関する整理, (引用日: 2023 年 1 月 5 日.)https://www.mhlw.go.jp/shingi/2002/06/s0614-3a.html
[3] 多田紘佳,中村誠，地方自治体ウェブサイトから得られる子育て支援に関する行政サービスの比較, 電子情報通信学会信越支部大会, P-30 (2019)
[4]サポートベクターマシン入門，栗田多喜夫，産業技術総合研究所 脳神経情報研究部門 , 2002
[5] 杉山一成，奥村学, 半教師有りクラスタリングを用いた Web 検索結果における人名の曖昧性解消, 自然言語処理，16 巻(2009) 5 号