様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　１９　日現在

機関番号：　２７１０１
研究種目：　挑戦的萌芽研究
研究期間：　2014～2017
課題番号：　２６５９０１１３

研究課題名（和文）困窮者支援活動を効率化する情報システムに関する実証的研究

研究課題名（英文）An Empirical Study of an Information System to Reduce the Burden of Support Activities for Needy People

研究代表者

中尾　泰士（NAKAO, Yasushi）

北九州市立大学・基盤教育センター・教授

研究者番号：６０３０９５３１

交付決定額（研究期間全体）：（直接経費）　　1,400,000円

研究成果の概要（和文）：困窮者支援活動において，その日常的な支援事項を内容別に分類し，それぞれに文章化してデータベースで管理するシステムが運用されている。本研究では，支援員が記入した支援内容をシステム側でその内容ごとに自動的に分類するしくみの開発を目指した。使用した手法は機械学習の分野における「トピックモデルによる潜在意味解析」である。その結果，比較的高い精度で分類できる支援内容と，あまりうまく分類できない支援内容があることが分かった。支援の現場で実際に活用するには，システムの精度をもう少し改良する必要がある。

研究成果の概要（英文）：In the support activities for needy persons, a database system is operated for recording daily support contents. Support contents are classified into several items by contents manually in such system.In this research, we aimed to develop a system to automatically classify support contents entered by supporters.The method we used is "latent semantic analysis by topic model" in the field of machine learning. As a result, we found that there are support contents that can be classified with relatively high accuracy. However, support contents that can not be classified very well also exist. In order to actually use the system at the site of support, it is necessary to improve the accuracy of our system.

研究分野：　情報学

キーワード：　困窮者支援　データベースシステム　機械学習

１．研究開始当初の背景

ホームレス等の困窮した状態にある人々の多くは，単に住む家をなくした状況にあるだけでなく，それまで持っていた家族や友人関係，社会的所属等の関係性をも失った状態にあるといわれる。そのような状況においては，衣食住等の物質的な支援だけでは不十分であり，失われた関係性を回復し，また新たな関係を構築する支援も同時に必要となる（たとえば，[1]等）。困窮者に対して，物質面の支援に加え，関係性の再構築についての支援も行うためには，拡散してしまったその人に関する情報を「記録」として集積することから始める必要があり，その情報管理を行うためにデータベースシステムが活用できる。

我々は，NPO 法人「北九州ホームレス支援機構」（現在は NPO 法人「抱樸」）を中心として実施された，『福岡絆プロジェクト』[2]（内閣府『パーソナル・サポート事業』[3]のモデル事業，2010 年 11 月～2013 年 3 月）において，困窮者支援データベースシステムの構築に参加した[4][5]。『福岡絆プロジェクト』の具体的な中身と事業の検証などについては奥田他[6]に詳しい。

我々は，この経験を通して，データベースシステムを利用する現場の支援員のニーズを知るとともに，情報システムによって支援活動の効果的な補助ができないかと考えるようになった。特に，困窮者への日常的なサポートを複数の項目に分類し，内容を個別に文章で入力していく作業は，情報システムによる補助が可能な部分だと思われた。

２．研究の目的

本研究は，情報システムを利用した困窮者支援活動の効率化を目指したものである。

上述のように，困窮者支援活動において，日常的な対応記録を文章で入力し，その後の支援に利用するということが行われている。支援活動を行う支援員にとって，この対応記録入力が業務時間の多くを占めており，この業務の効率化が図れれば，デスクワークではない本来の支援活動に，より多くの時間を割くことが可能になるだろう。

具体的には，入力された支援内容の文章をシステムが自動で分類するしくみを構築することで，支援員の記録入力負担が軽減できるだろうと考えた。

３．研究の方法

（１）データの取得・整理・加工（自然言語処理）

NPO 法人「抱樸」から，困窮者支援活動に関するデータ提供を受けた。支援機構が運用するデータベースは全体で見ると個人情報の塊であるが，支援内容が記録されたテーブルのみを見れば，支援を受けている個人名等は符号化されているため，個人が特定できる情報にはなっていない。しかし，念のため，本研究に補助者として参加する作業者（後述）についても，情報管理の必要性を徹底し，情報管理等に関する遵守事項を記した誓約書を個別に取り交わして参加してもらった。

提供を受けたデータ，すなわち，日常的な困窮者支援の内容を記録したデータには，対応内容を分類したコード（大分類・小分類）とともに，対応内容記録が文章で記入されている。本研究では，とりあえず大分類のみを取り扱った。この大分類は，具体的には「就労」「生活」「福祉」「健康」「人間関係」「社会保障（法律）」「金銭」「現況」の 8 つである。

このうち，「就労」や「金銭」についての支援は特に説明はいらないだろう。「生活」は住居関係を中心にした支援内容，「健康」は病気対応など，「人間関係」は家族・友人などの対人関係に関する対応，「現況」は被支援者の現在の状況を記録するためのものである。区別が難しいのは「福祉」と「社会保障（法律）」に関する支援である。「社会保障（法律）」の方は法律に基づいた公的な支援内容という説明を受けてはいるものの，両者の違いは部外者にはそれほど明確ではない。

１つの支援内容記録を１つのファイルに格納し，それが分類されるべき大分類番号とともに管理する。処理したファイルの数を（表 1）にまとめる。

大分類	ファイル数
就労	2,304
生活	2,767
福祉	3,774
健康	12,187
人間関係	8,590
社会保障（法律）	2,068
金銭	26,729
現況	868
総数	59,287

（表 1）処理に用いたファイル数の大分類別集計

これらのファイル群に対して，多少の下処理（文字コード変換など）をほどこした後，各ファイル中の日本語文章を単語に分割し，その単語を品詞に分類するとともに，各単語の出現頻度を記録して，機械学習に利用する「辞書」を作成する。これらの一連の処理は，MeCab[7]，gensim[8]などのライブラリを活用し，主としてプログラミング言語 Python を使用した。

（２）分析アルゴリズム（機械学習）

上述の処理によって，ファイル群中にあら

われたすべての単語の一覧としての辞書と，各ファイル中にどの単語がどの程度出現するのかの情報が抽出される。これは各ファイルを「Bag of Words（BoW）」と呼ばれるものに変換したものである。すなわち，i 番目のファイル $\mathbf{F}_i$ に対して，辞書中の j 番目の単語が出現する頻度を d_{ij} として記録する：

$$\mathbf{F}_i = (d_{i1}, d_{i2}, \cdots, d_{iN}) ,$$

ここで，N は分析に使った辞書中にあらわれる総単語数であり，単語 j が出現しなかった場合は $d_{ij}=0$ とする。結果として，ファイル $\mathbf{F}_i$ が持つ単語出現情報は N 次元のベクトルとして表現できる。なお，あまりにも頻出する単語については，あらかじめ除外しておく。

このように，ファイル中の単語出現情報を数値化してしまうことにより，ファイル同士の類似度を計算できるようになる。なぜなら，同じような内容を記述しているファイルは同じような単語を使用していることが予想され，それらのベクトルは「似ている」と期待されるからである。

一般には，この N の値は非常に大きな値になる。我々が用いたファイル群においては，名詞の辞書のみで 15,000 程度であり，このような大きな次元を取り扱うためには大きな計算資源が必要になる。

しかし，実際には 1 つのファイルにおいて，ほとんどの要素 d_{ij} の値は 0 であり，このような「疎」なベクトルをまともに取り扱う必要はない。具体的には，次元削減と呼ばれる手法を使って，取り扱う次元を削減することができる。本研究では，次元削減の方法として，gensim[8]，および，scikit-learn[9]ライブラリ中の LSI（Latent Semantic Indexing）を用いて，300 次元まで次元を削減した。これは，ファイル群中の単語たちをグループ化して，300 種類の「トピック」にまとめてしまうことを意味する。この手法は「トピックモデルによる潜在意味解析」と呼ばれ，たとえば[10][11]などに詳しく記述されている。

こうして，M（=300）次元まで次元を減らした基底ベクトル $\mathbf{e}_j$ (j=1,…,M)を使って，ファイル $\mathbf{F}_i$ の特徴は，

$$\mathbf{F}_i = \Sigma_j \, w_{ij} \cdot \mathbf{e}_j$$

と表現できる。ここで，w_{ij} は，ファイル $\mathbf{F}_i$ のトピック $\mathbf{e}_j$ への重みづけであり，ファイル $\mathbf{F}_i$ は，この w_{ij} を要素としてもつ M（=300）次元のベクトルとして取り扱えることになる。

各大分類の支援内容を記録したファイル群の特徴は，トピックにまとめても「似ている」ことが予想される。そこで，同一大分類に属するファイル群について，ベクトル $\mathbf{F}_i$ の和をとり，その結果のベクトルを大きさ 1 に規格化しておく（単位ベクトル化）。これが，その大分類の特徴をもっともよく表す特徴ベクトルである。

こうして，「就労」から「現況」にいたる 8 つの大分類ごとに，その特徴ベクトル $\mathbf{v}_k$ (k=1,2,…,8)が用意できた。これが，およそ 60,000 件の支援内容をもとに，機械が学習した結果である。

（3）分類判定アルゴリズム

さて，いまここに新しい文書ファイルが来たとする。そのファイルに対して，上述と同様のことを行い，この新文書の特徴ベクトル $\mathbf{u}$（M=300 次元ベクトル）を算出する。この特徴ベクトル $\mathbf{u}$ が，既存のどの特徴ベクトル $\mathbf{v}_k$ (k=1,2,…,8)に「似ている」かを計算すれば新文書を分類することができる。

同じ次元をもつベクトル同士について，「似ている」度合いを判定するために使ったのは cos 距離である。2 つの単位ベクトル $\mathbf{a}$, $\mathbf{b}$ の内積とそれらがなす角度 θ の間に，

$$\mathbf{a} \cdot \mathbf{b} = \cos\theta$$

の関係があることを使う。$\mathbf{a}$ と $\mathbf{b}$ が向く方向が近い（θ が小さい）ならば，$\mathbf{a} \cdot \mathbf{b}$ の値は大きくなり，全く同じ方向を向いているとき（θ=0）に最大値 1 をとる。これを使って，ある文書の特徴ベクトル $\mathbf{u}$ とそれぞれの大分類が持つ特徴ベクトル $\mathbf{v}_k$ (k=1,2,…,8)との間の cos 距離を計算することで，もっとも近い大分類を判定した。

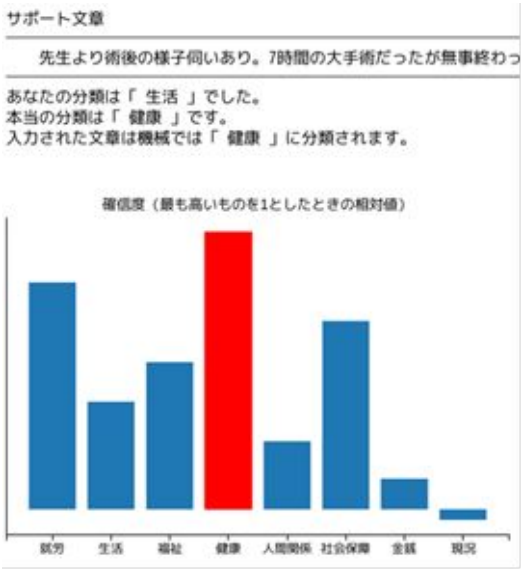
（4）性能評価と情報の可視化

構築したシステムの性能評価のために，与えられた文書をどの支援大分類に分類するかについて，プログラム（機械）と人間のどちらが正しく分類できるかを比較する実験を行った（実験 1）。

この実験に参加した作業者は 30 代から 40 代の年代に属する女性 3 名である。いずれも困窮者支援の経験はなかった。実験の前に，作業者には各大分類のおおまかな内容について説明をしておいた。

作業者はランダムに表示される文章を読んで，それをどの大分類に分類するかを判定する。各自の判定の結果が送信されたのち，画面には「正解」（もともとの支援内容が分類されていた大分類）と，同じ文章を機械が判定した分類結果，および，その「確信度」（上述の cos 距離をもとにしたもの）を表示させる。その画面サンプルを（図 1）示す。機械による判定の結果とその確信度をグラフ化して表示している。

こうして，作業者自身もこの作業を積み重ねることによって，次第に分類を「学習」していく過程になっている。結果的にこの作業は各作業員が 1,000 件程度ずつ行い，合計では 3,630 件のデータが集まった。



（図 1）システム性能実験の画面サンプル

また，作業者が特定の支援分類を念頭に置いて入力した文章を機械がどの程度正しく分類できるかの実験も行った（実験 2）。前述のように，作業者はいずれも困窮者支援の経験はないが，それぞれが 1,000 件程度の実際の支援内容と分類を判読する過程を経験した後でこの実験を行ったため，ある程度は架空の支援内容を作文する作業も可能であった。この実験では，3 名の作業者で合計 610 件のデータを収集した。この実験は，文章入力から機械による判定が帰ってくるまでのシステムの反応時間が実用的なものであるかの性能評価も兼ねていたが，おおむね良好な反応時間だったと考えている。

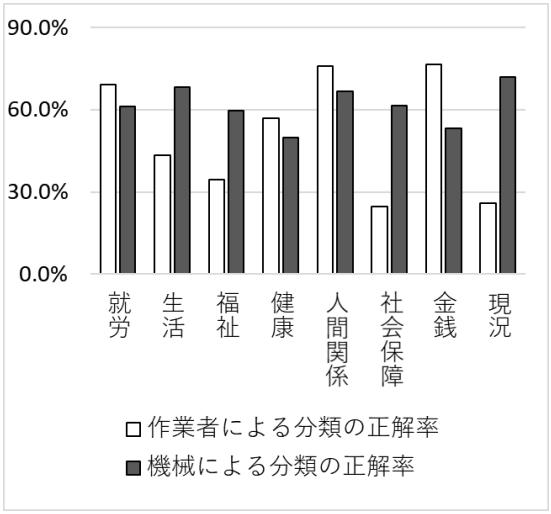
４．研究成果

まず，（実験 1）において得られたデータについて，実際の困窮者支援の現場で記録されたものを別の人間（作業者）が読んで分類した結果と，機械が判定した分類結果について，それぞれの正解率を大分類項目ごとに比較した結果を（図 2）に示す。

実験中，ランダムに表示された支援内容の内訳は，「就労」: 149 件，「生活」: 145 件，「福祉」: 263 件，「健康」: 797 件，「人間関係」: 528 件，「社会保障（法律）」: 117 件，「金銭」: 1581 件，「現況」: 50 件の合計 3,630 件である。これらの比率はもとのファイル分布（表 1）とほぼ同様の比率であった。

機械の方が人間よりも高い正解率を示している分類は「生活」「福祉」「社会保障」「現況」の分類である。特に「福祉」と「社会保障（法律）」と名付けられた分類においては，困窮者支援活動の経験のない作業者には分類が難しかった可能性がある。

逆に，「就労」「健康」「人間関係」「金銭」については人間の判断の方が高い正解率を示した。



（図 2）各分類において作業者の判定と機械による判定の正解率の比較

この実験においては，ある文書は，それを最初に記入した者が指定した分類（記入者の分類＝「正解」），それを読んだ作業者が下した分類（作業者の分類），システムによる機械分類の 3 つの分類結果を持つことになる。それらの一致状況について 3,630 件をケース分けすると以下のような結果になった。

（ケース 1）3 つが同じ分類になる場合：1,646 件，全体の 45.3%
（ケース 2）作業者とシステムは一致するが，記入者の分類（正解）とは異なる場合：335 件，同 9.2%
（ケース 3）記入者と作業者は一致するが，システムが異なる結果を出す場合：714 件，同 19.7%
（ケース 4）記入者とシステムは一致するが，作業者が異なる結果を出す場合：401 件，同 11.0%
（ケース 5）それぞれ全てが異なる場合：534 件，同 14.7%

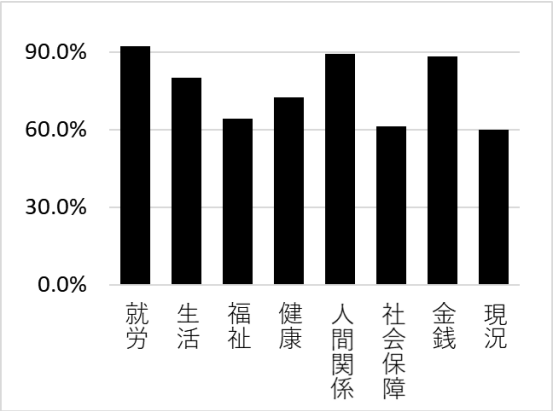
これらのうち，（ケース 1）がもっとも望ましい状況である。すなわち，記入者が適切な文書作成を行い，システムと判読する作業者もそれを正しく認識した結果といえる。

一方，（ケース 2）は，記入者の表現力に問題があったため，判読者と機械が誤った分類をした可能性がある。

そこで，判読した作業者とシステムが同じ分類を行った場合のうち，それが正解である場合の割合を大分類ごとに算出した結果を（図 3）に示す。

「就労」「人間関係」「金銭」については 90%程度の正解率を与えるのに対し，「福祉」「社会保障（法律）」「現況」の正解率は 60%程度にとどまっているのが分かる。支援の現場で，これらの大分類は明確な差異が認識されているのかも知れないが，対応内容として記録される際には支援内容入力者の間でも分類

に混乱があるのかもしれない。今後の検討が必要である。

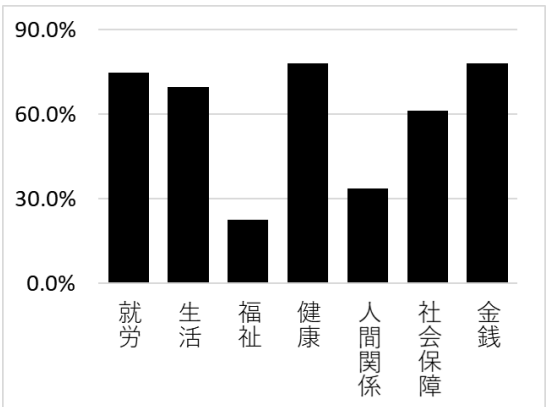


（図 3）判読者とシステムが同じ分類判定をしたものの正解率

その他では，（ケース 3）はシステムの性能の問題，（ケース 4）は判読した作業者の読解能力が至らなかった可能性，（ケース 5）は 3 者の間の複合的な原因によるものと考えられる。

（ケース 3）では，判読した作業者が正解した 2,360 件のうち，システムは 714 件を間違えており，その割合は 30.3%にあたる。一方，（ケース 4）では，システムが正解した 2,047 件のうち，判読者が間違えた割合は 19.6%の 401 件である。これだけを見ると一見，人間の眼の方がシステムよりも正しい結果を導き出している印象がある。

しかし，細かく見ていくと，（ケース 3）において，システムはどの大分類でも誤分類率がほぼ同じ約 30%なのに対し，人間の眼の場合は，「福祉」「社会保障（法律）」「現況」の大分類において，機械が正解している場合の 60%～70%も間違える結果が出ている。分類に微妙な判定が必要な場合にはシステムの方がより正しい判定を下せるといえるかもしれない。



（図 4）作業者が入力した架空の支援内容をシステムが判定した正解率

次に（実験 2）の結果について述べる。作業者に入力してもらった文書の内訳は「就労」：99 件，「生活」：99 件，「福祉」：62 件，「健康」：100 件，「人間関係」：101 件，「社会保障（法律）」：49 件，「金銭」：100 件の合計 610 件である。（図 4）にそれぞれの大分類におけるシステムによる判定の正解率を示す。

（図 4）を見れば，「就労」「生活」「健康」「金銭」が比較的高い正解率を示す一方，「福祉」「人間関係」が極端に低い正解率になった。困窮者支援の経験のない作業者には，これらの分類の文章を作成することは難しかったことがうかがえる。

本研究期間を通じて，研究代表者の大学における管理運営業務が増大したこともあり，当初の計画通りになかなか研究が進まなかったことについては忸怩たる思いがある。今後は，これらの成果を論文等にまとめるとともに，必要に応じてシステムの改良に取り組みたいと考えている。

＜引用文献＞

[1] 奥田知志，「第三の困窮と犯罪」，犯罪社会学研究，Vol.35，pp. 21-37，2010 年
[2] 「福岡絆プロジェクト計画」，https://www.kantei.go.jp/jp/singi/kinkyukoyou/suisinteam/PSSmp2/siryou9.pdf，2018 年 6 月 10 日アクセス
[3] 内閣府，「パーソナル・サポート・サービス」について，https://www.kantei.go.jp/jp/singi/kinkyukoyou/suisinteam/SNdai5/sankou1.pdf，2018 年 6 月 10 日アクセス
[4] Asaba, N. and Nakao, Y.,“A Case of the Development of the Database System for Efficient Support to Needy Persons”, Proceedings of the International Conference on e-Commerce, e-Administration, e-Society, e-Education, and e-Technology, pp.1732-1733, 2013
[5] 中尾泰士，浅羽修丈，「困窮者支援のためのデータベースシステム構築－『福岡絆プロジェクト』における事例」，北九州市立大学基盤教育センター紀要，第 21 号，pp.17-35，2014 年
[6] 奥田知志，稲月正，垣田裕介，堤圭史郎，『生活困窮者への伴走型支援』，第 3 章，pp.100-169，明石書店，2014 年
[7] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://taku910.github.io/mecab/，2018 年 6 月 11 日アクセス
[8] ŘEHŮŘEK, Radim and Petr SOJKA. “Software Framework for Topic Modelling with Large Corpora.”, in Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks, pp.46-50, 2010

[9] Pedregosa, F. *et al.*, “Scikit-learn: Machine Learning in Python”, Journal of Machine Learning Research, vol.12, pp.2825-2830, 2011
[10] 佐藤一誠,『トピックモデルによる統計的潜在意味解析』, コロナ社, 2015 年
[11] 岩田具治,『トピックモデル』, 講談社, 2015 年

５．主な発表論文等

〔雑誌論文〕（計 0 件）

〔学会発表〕（計 0 件）

〔図書〕（計 0 件）

〔産業財産権〕
該当なし

〔その他〕
該当なし

６．研究組織

(1)研究代表者
中尾　泰士（NAKAO, Yasushi）
北九州市立大学・基盤教育センター・教授
研究者番号：60309531

(2)研究分担者
浅羽　修丈（ASABA, Nobutake）
北九州市立大学・基盤教育センター・准教授
研究者番号：50458105