様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成　25　年　5　月　21　日現在

機関番号：32502
研究種目：基盤研究（C）
研究期間：2010～2012
課題番号：22530516
研究課題名（和文）　社会調査の基盤を提供する自由回答の自動コーディングシステムの開発と公開
研究課題名（英文）　Development and Release of an Automatic Coding System of Answers to Open-ended Questions in Social Surveys
研究代表者
高橋　和子（TAKAHASHI KAZUKO）
敬愛大学・国際学部・教授
研究者番号：30211337

研究成果の概要（和文）：
社会調査では回答者の職業や産業は重要で、正確さを期するために自由回答で収集する場合が多い。しかし、統計処理のために収集後にコード化する作業が必須で、最近では国内標準コードに加えて国際標準コードの要請も生じており、コーダの負担が増大している。本研究では、自然言語処理や機械学習など人工知能における最新の成果を適用してコーディング作業を自動化し、結果をWebにより入手できるシステムを開発した。その際、各コードには人間による見直しが必要か否かを3段階の確信度で付与するため、作業の大幅な軽減が見込める。
研究成果の概要（英文）：
When we collect occupation data by open-ended questions in social surveys, we need to categorize those data for statistics processing. Conducting the occupation coding manually is time-consuming and complicated. Therefore, we have developed a Web-based automatic occupation coding system for both Japanese standard categories and international standard categories with Natural Language Processing and Machine Learning. When researchers upload an occupation data file on a website, they can obtain desired occupation codes files with confidence. The system can also be applied to industry data.
交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	1,700,000	510,000	2,210,000
２０１１年度	500,000	150,000	650,000
２０１２年度	1,100,000	330,000	1,430,000
年度			
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：自然言語処理
科研費の分科・細目：社会学・社会学
キーワード：社会調査法・自由回答・自動コーディング・SSM職業コーディング・ISCOコーディング・Webシステム・機械学習・確信度

１．研究開始当初の背景

近年、SSM（Social Stratification and Social Mobility）調査やJGSS（Japanese General Social Surveys：日本版総合的社会調査）、東京大学社会科学研究所パネル調査など多くの大規模社会調査が実施されている。さらに、COE各種プログラムによる調査も行われており、今後もますます増える傾向にあることが予想される。これらの調査は貴重な情報を提供し、学問の進歩に多大な貢献をしている。社会現象を分析する上で、社会学者が最も重視してきた変数は職業であり、上記の社会調査でも職業の情報が含まれている。

日本でも海外でも、職業は多くのコード（日本の場合は約200個､国際標準では約400個）を基礎データとして用いることが多いが、調査時にすべてのコードを回答者に提示するのは現実的ではない。また、仮に提示できたとしても、より本質的な問題として、回答者のもつ自己認識と客観的内容に違いがある場合が多いため、得られた情報の信頼性が低くなる可能性がある。したがって、職業については、大規模調査であっても自由記述による回答（自由回答）でデータを収集し、研究者側の判断でいずれかのコードに分類する（コーディング）ことが推奨されている。

しかし、実際には自由回答のコーディングは大変な手間がかかり、とりわけ、職業のコーディングは非常に複雑で熟練を要する。コーダはコーディングのためのマニュアル（1995年SSM調査委員会1995など）が用意され十分な訓練を受けるが、特に大規模調査で顕著であるが、多数のコーダが長期間作業するために結果の一貫性を保つことが困難で、調査データの質の低下は避けられない。また、コーディングに費やされる労力や費用も膨大である。

このような中、研究代表者は、自由回答の処理・分析方法に関する研究を行ってきた（H3年度科研費およびH7年度科研費研究成果他）。最終的に、人工知能研究の一分野である自然言語処理や機械学習における最新の研究成果を適用して職業/産業データの自動コーディングを行い、その結果をコーダのヒントとなるよう画面に表示する「職業コーディング支援システム」（研究計画・方法 表１参照）を開発するに至った（H16～H17年度科研費研究成果他）。いずれもコーダの労力軽減だけでなく、正確性も高める効果があり、特に、最も性能の高い方法（ルールベース手法による自動コーディング結果を機械学習であるサポートベクターマシン（SVM）の素性に追加する方法）は、前述の大規模調査を始め多数の調査で利用されている。利用者からの評価が高く、国際調査実施担当の総務省統計センターからも注目され、代表的な自由回答である職業／産業データを基礎データに変換する有力な支援方法としての役割を果たしている。

しかし、最近になって、社会調査が直面する問題や利用者からの要望による状況の変化により、以下(1)～(3)に述べるような新たな課題が生じている。

(1)国際標準職業分類であるISCO（International Standard Classification of Occupations）の自動コーディング

社会学においては、国際比較研究が活発であるが、多くの日本の調査は独自の体系（SSM職業分類）を用いてきたため、直接、海外の調査と比較することができない。また、アーカイブによって調査票が英訳され、世界に開かれるようになったため、海外の研究者による日本の調査データ利用が増加しているが、日本独自の職業分類はそのまま海外で使えない。こうした理由から、新規調査だけでなく既存調査に対してもISCOコード（約400個）を付ける必要性が高まっているが、ISCOはSSM職業分類とは視点が異なる体系で、両者は単純に変換できないという大きな問題がある。

(2)システムの利用環境の整備

本支援システムはこれまで大規模調査での利用が多かったこともあり、伝統的な「コーディング合宿」を引き継ぐ形で、「多数のコーダが1カ所に集合し、コーディング作業管理者のもとで利用」されることが多い。しかし、支援システムの利用経験のある研究者を中心に、「自分の調査に対しても自由に利用できる」コンピュータ環境を求める声が多く、Web公開を検討する必要が出てきた。大規模調査における「同時集中処理」方式も、実際には、遠方からの参加者には負担が大きく、スケジュール調整も困難である。

(3)システムが処理できる自由回答の拡張

例えば、購入品名（統計センター「家計調査」）に対する自動コーディングが望まれているが、他の自由回答に対しても要望が多い。

以上より、社会調査においては、規模に関係なく、「自由回答を少ない労力で正確に基礎データに変換し、随時提供できるシステム」が必要で、これを実現するための新たな自動コーディングシステムの開発と公開が要請されている。

２．研究の目的

本研究の目的は、１．で述べた課題を解決するため、H7年度科研費およびH16～H17年度科研費研究成果である現システムを整理・統合した上で新たな機能を追加し、Webを通じて利用が可能な公開システムとして新たに再構築することである。具体的には、下記４項目の実現を目的とする。

(1)新規機能の追加

主な新規機能は次の2つである。

①ISCO自動コーディング機能の追加

ISCOに対しては、社会調査における国内標準であるSSM職業コードとは別の自動コーディング方法を提案する。

②機械学習による自動コーディング結果に3段階の確信度（信頼性）を付与

コーダがチェックする絶対量を減らすために、システムが出力するコードに、「人間がチェックをする必要がある」「チェックした方がよい」「チェックする必要がない（完全自動化）」の3段階に区別した確信度を付与する手法を提案する。判断の基準には、ク

ラス所属確率（H16～H17 年度科研費研究成果）の利用を予定している。

(2)分類精度の向上
①ISCO 訓練データセットの増強
機械学習では訓練データ（正解の付いたデータ）の量が増えるほど精度が向上するが、現時点で信頼できる ISCO コードが付いているのは 2005 年 SSM 調査データと JGSS-2006 のデータしかない。そこで、1995 年 SSM 調査データにも ISCO コードを付けて、ISCO 自動コーディング用の訓練データを増やす。
②エラー解析
新システムに活かすため、システムのエラー解析を行う。具体的には、コードごとにシステムの正解／不正解の状況を調査し、エラーの傾向を明らかにする。
③SVM における有効なアンサンブル学習の提案
SVM はもともと分類精度が高いため、分類精度をさらに向上させるのは困難が予想さえる。どのようなアンサンブル学習が有効であるかを検討し提案する。

(3) Web 版システムとして公開
①システムの改変
現システムに(1)の機能を追加したシステムを、利用者が Web を通じて職業・産業データのファイルをアップロードすれば、コーディング結果のファイルをダウンロードできるようなシステムとして再構築する。
②システムの運用
公開を予定している東京大学社会科学研究所附属社会調査・データアーカイブセンター（SSJDA）の Web を介した利用手続き方法について検討する。

(4)処理が可能な自由回答の拡張
現システムでは、自動コーディングの対象が職業・産業情報に限定されているが、他の自由回答データに対しても適用できるようシステムの汎用性を高める。
①自由回答データの準備
②システムの汎用化

３．研究の方法
研究目的ごとに表にまとめる。
(1) 新規機能の追加

表 1　研究目的(1)の研究方法

	内容	担当者	予定年度
①	SSM 職業自動コーディングにおけるアルゴリズムを参考に、SVM において有効な素性を実験により検討し決定	研究代表者	2011
②	機械学習を用いる SSM 職業コードおよび ISCO の自動コーディング結果に対する確信度として、複数の分類スコアを用いて推定する「クラス所属確率」の利用を検討	研究代表者	2012
	確信度付与の画面表示形式を検討	研究代表者 共同研究者	

(2) 分類精度の向上

表 2　研究目的(2)の研究方法

	内容	担当者	予定年度
①	1995 年 SSM 調査データセットに対する人手による ISCO コーディング作業	学生アルバイト（のべ 60 名）	2010
	同上　作業管理・監督	共同研究者	
②	2005 年 SSM 調査 ISCO コーディング結果のコード別精査	共同研究者 1 名	2010
③	SVM において有効なアンサンブル学習を実験により検討し提案	研究代表者	2010

(3) Web 版システムとして公開

表 3　研究目的(3)の研究方法

	内容	担当者	予定年度
①	ルールベース手法による自動コーディングシステムを Web 公開用に再構築	研究代表者 研究協力者	2010
	機械学習による手法による自動コーディングシステムを Web 公開用に再構築	研究代表者 研究協力者	2011
	ルールベース手法と機械学習手法のハイブリッドによる自動コーディングシステムを Web 公開用に再構築	研究代表者 研究協力者	2012
②	新システム利用手続き（必要書類など）の検討	研究代表者 共同研究者	2011
	新システム利用のための Web ページ作成	共同研究者 1 名	2012

(4)処理が可能な自由回答への拡張

表 4　研究目的(4)の研究方法

	内容	担当者	予定年度
①	拡張が可能な自由回答データを準備	共同研究者	2012
②	システムの汎用化を検討	研究代表者	2012

なお本研究が真に社会調査に資するためには、情報処理分野の研究代表者だけでは不十分であるため、社会調査に精通し、システムの運用管理者・利用者としてともに本システムに関わってきた SSJDA の研究者 2 名を研究分担者とした協同研究を行う。

４．研究成果

研究目的ごとに述べる。

(1) 新規機能の追加

①ISCO 自動コーディング機能の追加

素性をさまざまに変化させた実験の結果、SSM 自動コーディングに適用した素性に、「ルールベース手法と SVM による手法のハイブリッドによる方法」により出力された「SSM コード（第 1 位のみ）」およびスキルの代用としての「学歴」の 2 つを素性として追加する方法が有効であることがわかった(図 1 参照)。

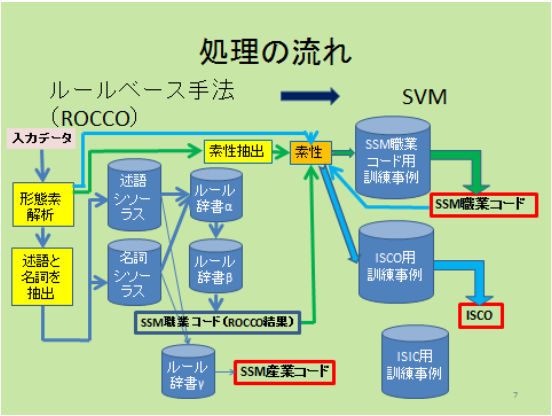


図 1　新規自動コーディング処理の流れ

[実験結果] 2005SSM データセットを訓練データとし（16,089 サンプル）、JGSS-2006、EASS-2008、JGSS-2010 を評価データとした実験の結果、予測第 3 位までの分類精度（正解率）は順に 72%、72%、69%であった。

[評価および今後の課題] 分類精度のさらなる向上が必要である。

[主な成果発表] 雑誌論文①、学会発表①②、その他①（学会発表資料の公開）

②機械学習による自動コーディング結果に 3 段階の確信度を付与

当初予定していた「クラス所属確率」を利用する方法は処理が複雑過ぎたため、確率までは求めず、SVM における分離平面からのスコアを複数個利用するアイディアを利用して、第 1 位のスコアと第 2 位のスコアの差により 3 段階（A「人間がチェックする必要がない（完全自動化）」、B「チェックした方がよい」、C「人間がチェックをする必要がある」）に区別する方法を提案した。

[実験結果] 2005SSM データセットを訓練データとし、JGSS-2006、EASS-2008、JGSS-2010 を評価データとした実験結果は下表の通り。

表 5　確信度別分類精度（正解率）（単位：%）

コード	A	B	C
SSM 職業コード	９５	７２	３６
ISCO	９４	６８	２８

表 6　確信度別カバー率（単位：%）

コード	A	B	C
SSM 職業コード	２９	４３	２８
ISCO	7	６７	２６

[評価および今後の課題] SSM 職業コード、ISCO のいずれにおいても、確信度Ａによる分類精度の高さは評価できる。しかし、作業量軽減のためには、特に ISCO における確信度Ａのカバー率を向上させる必要がある。

[主な成果発表] 学会発表①②、その他①（学会発表資料の公開）

(2) 分類精度の向上

①ISCO 訓練データセットの増強（実施せず）

②エラー解析（実施せず）

[実験結果] ルールベース手法におけるシソーラスやルール辞書の更新は行わず、訓練データを従来の JGSS-2000、-2001、-2002、-2003 に JGSS-2005 を追加した（計 39,120 サンプル）。SSM 職業コードと SSM 産業コードの自動コーディング結果は下表の通り。

表 7　コード別分類精度（正解率）（単位：%）

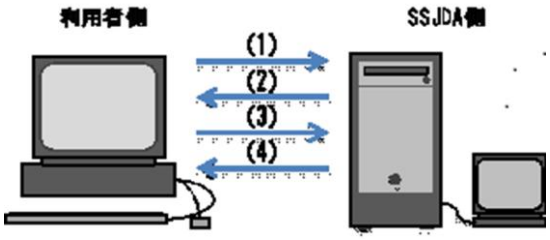
コード	JGSS-2006	EASS-2006	JGSS-2010	2005 SSM
SSM 職業	７９	７９	７８	８１
SSM 産業	７１	７８	７４	７０

③SVM における有効なアンサンブル学習の提案

複数の分類器を生成するために、まず素性を変化させる方法を提案したが、訓練データ構築の手間がかかるという問題があったため、次には訓練データを変化させて未知の事例に対するクラスを予測させ、各事例においてクラス所属確率の高い分類器の予測を最終結果とする方法を提案した。

［評価および今後の課題］提案手法は従来手法と比較すると、分類精度の低い困難なタスクや分類器数が少ない場合において有効性を示した。ただし、処理手続きが複雑なため、本システムには採用しないことにした。
［主な成果発表］学会発表⑤⑥⑦⑧

(3) Web 版システムとして公開
①システムの改変
本システムの利用方法を図 2 に示す。SSJDA 担当者用の操作画面を図 3、モニタ画面を図 4、Result View 画面を図 5 に示す。



(1)［利用者］利用申請書をメイルにより SSJDA に送信（希望する職業・産業コードの種類を明記）
(2)［SSJDA］ユーザ ID、パスワードの発行とアップロード（ダウンロード）場所の指定
(3)［利用者］入力用データファイルをアップロード
(4)［利用者］結果ファイルをダウンロード
図 2　システムの利用方法

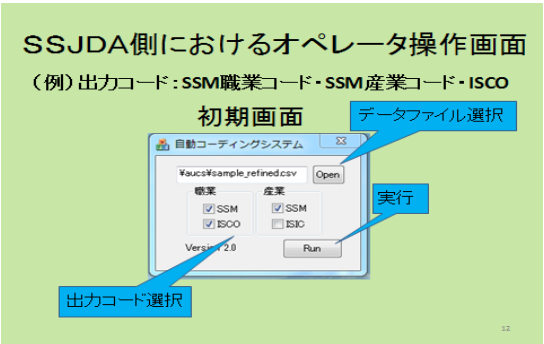


図 3　SSJDA 担当者操作画面（初期画面）

図 4　モニタ画面例（処理途中）

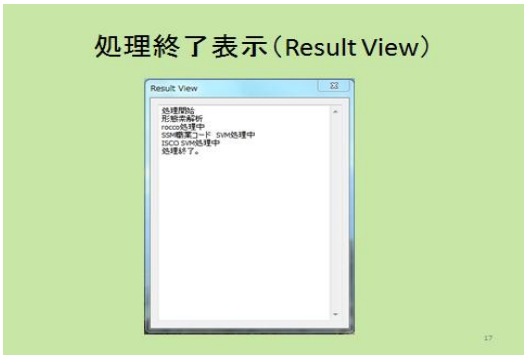


図 5　Result View 画面例（処理終了時）

［主な成果発表］学会発表①②③④、その他③

②システムの運用
本システムは、平成 25 年 9 月より SSJDA の Web サイトにより公開予定である。
［主な成果発表］その他②④⑤

(4)処理が可能な自由回答への拡張
①自由回答データを準備（実施せず）

②システムの汎用化
システムの拡張を考慮した設計を行った。

本研究は社会調査方法における大きな変革で、職業・産業データに限定されてはいるもののこれまで手作業で行っていた自由回答から基礎データへの変換を、人工知能分野の最新の研究成果を取り入れて迅速に正確に随時提供できるようにした。特に、a)これまで扱えなかったデータを用いた分析ができるようになり、特に ISCO コードは国際比較研究の推進や海外からの利用で社会学の進展に大きく貢献する。b)個々の研究者の調査だけでなく大規模調査のコーディングも、インターネットを利用することで、各自の居場所で各自の予定に合わせた実施ができる点は評価できる。さらに、本研究は、社会調査方法論と情報処理分野にまたがる学際的な研究で、社会調査における他分野との協同の発展可能性を広げる意義ももつ。

５．主な発表論文等
〔雑誌論文〕（計　１件）

①　高橋和子、ISCO 自動コーディングシステムの分類精度向上に向けて―SSM および JGSS データセットによる実験の結果―、大阪商業大学 JGSS 研究センター編 『JGSS Research Series No.8:日本版総合的社会調査共同研究拠点研究論文集[11]』、査読なし、2011、193－205、
http://jgss.daishodai.ac.jp/research/monographs/jgssm11/jgssm11_17.pdf

〔学会発表〕(計　８件)
① 高橋和子、田辺俊介、吉田崇、魏大比、李偉、確信度付き職業・産業コーディング自動化システムの開発と公開、数理社会学会第 55 回年大会報告要旨集、38-41、2013 年 3 月 19 日、東北学院大学土樋キャンパス

② 高橋和子、田辺俊介、吉田崇、魏大比、李偉、Web 版職業・産業コーディング自動化システムの開発、言語処理学会　第 19 回年次大会論文集、769-772、2013 年 3 月 15 日、名古屋大学東山キャンパス
http://www.anlp.jp/proceedings/annual_meeting/2013/pdf_dir/P5-8.pdf

③ 高橋和子、魏大比、田辺俊介、吉田崇、社会調査における職業・産業コーディング自動化システムの Web 公開、言語処理学会　第 18 回年次大会論文集、219-222、2012 年 3 月 14 日、広島市立大学
http://www.anlp.jp/proceedings/annual_meeting/2012/pdf_dir/P1-7.pdf

④ 高橋和子、魏大比、田辺俊介、吉田崇、職業・産業自動コーディングシス ムの Web 公開に向けて―機械学習による手法、数理社会学会　第 52 回大会・ネットワークが創発する知能研究会　第 7 回合同開催、2011 年 9 月 6 日、信州大学松本キャンパス

⑤ 高橋和子、多クラス SVM におけるクラス所属確率を用いたアンサンブル学習の提案、情報処理学会　第 201 回自然言語処理・第 86 回音声言語情報処理 合同研究発表会、2011 年 5 月 16 日、東京大学本郷キャンパス
https://ipsj.ixsq.nii.ac.jp/ej/index.php?active_action=repository_view_main_item_detail&item_id=74053&item_no=1&page_id=13&block_id=8

⑥ 高橋和子、魏大比、田辺俊介、吉田崇、職業・産業自動コーディングシステムの Web 公開に向けて、数理社会学会　第 51 回大会、2011 年 3 月 8 日、沖縄国際大学
http://www.u-keiai.ac.jp/international/teachers/staff-013/upimg/20110418113705965634487.pdf

⑦ 高橋和子、クラス所属確率を用いた多クラス SVM におけるアンサンブル学習、情報処理学会　第 73 回全国大会論文集、 2-25－2-26、2011 年 3 月 4 日、東工大大岡山キャンパス
https://ipsj.ixsq.nii.ac.jp/ej/index.php?active_action=repository_view_main_item_detail&item_id=76230&item_no=1&page_id=13&block_id=8

⑧ 高橋和子、クラス所属確率を利用したアンサンブル学習、人工知能学会　第 24 回大会発表論文集、2010 年 6 月 9 日、長崎ブリックホール
https://kaigi.org/jsai/webprogram/2010/pdf/260.pdf

〔その他〕(計　５件)
①敬愛大学＞国際学部＞教員紹介－国際学科－高橋和子 HP
http://www.u-keiai.ac.jp/international/teachers/staff-013/index.html

②東京大学社会科学研究所附属社会調査・データアーカイブ研究センター＞社会調査＞共同調査と共同研究＞自動コーディング（職業・産業）HP
http://ssjda.iss.u-tokyo.ac.jp/joint/autocode/

③高橋和子、魏大比、産業・職業コード自動化システムインストール＆実行マニュアル

④田辺俊介、自動コーディング（職業・産業）システム利用申請書（案）

⑤高橋和子、田辺俊介、入力ファイルの形式

６．研究組織
(1)研究代表者
高橋　和子（TAKAHASHI KAZUKO）
敬愛大学・国際学部・教授
研究者番号：30211337

(2)研究分担者
田辺　俊介（TANABE SHUNSUKE）
東京大学・社会科学研究所・准教授
研究者番号：30451876
吉田　崇（YOSHIDA TAKASHI）
東京大学・社会科学研究所・助教（H22→H23）
静岡大学・人文社会科学部・准教授（H24）
研究者番号：80455774

(3)連携研究者（なし）

(4)研究協力者
魏　大比（豊原　明）(GI DAIBI)(TOYOHARA AKIRA)
東京工業大学大学院・情報理工学研究科・博士研究員（H22→H23）
(株)名校教育グループ・代表（H24）
李　偉（RI I）
東京工業大学大学院・理工学研究科・博士課程在学（H24）