# 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

高橋 和子 †　　魏　大比 †
† 敬愛大学 国際学部
{takak, a-toyohara}@u-keiai.ac.jp

田辺　俊介 ‡　　吉田　崇 ‡
‡ 東京大学 社会科学研究所
{tanabe, tyoshida}@iss.u-tokyo.ac.jp

## 1　はじめに

社会調査においては, 個人の仕事内容である「職業」や, 従業先の事業内容である「産業」は重要な情報であり, 国勢調査と同様に, より詳細な情報を得るために自由回答で収集される場合がある. このとき, 統計処理を行うために, 自由回答をあらかじめ決められた分類コードに変換する作業（職業 · 産業コーディング）[1]が必須であるが, 職業や産業のコードは個数が多く [2], コード化のルールも複雑なために [3], コーダにとって多大な労力や時間を要するという問題が存在する [3, 7].

高橋らはコーダを支援する目的で, 職業 · 産業コーディングにルールベース手法や機械学習を適用して自動化を行い, その結果を候補として提示するシステムを開発した [8, 9, 10]. システムは SSM（Social Stratification and social Mobility）調査や JGSS（Japanese General Social Surveys）など大規模プロジェクトを中心に利用が高まっているが, いずれも公開システムではないために, 一般の研究者にとって利用しやすい環境にあるとはいえない状況にある.

誰もが自由に職業や産業情報の自動コーディング結果を得ることができるためには, システム自体を Web により公開することが有効であると考えられる. しかし, 大容量の辞書やシソーラスを所持し, 構成が単純ではないシステムを, PC 環境がさまざまに異なる環境のもとで, 利用者自身がダウンロードして実行するのは容易なことではない. したがって, 本稿では, 利用者が Web サイト [4]を通じて, 職業や産業情報を所定の形式のデータファイルとしてアップロードすれば, 自動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の構築に取り組む.

その際, 新たに次の 2 つの機能についての検討も行う. まず, 近年の国際比較研究の隆盛に伴い, 職業 · 産業コーディングにおいても, 従来, 社会学で用いられてきた国内標準のコード以外に, 国際標準のコードへの変換が要請されており, これに対応した自動化システムも開発された [11]. これらの自動化システムが現在は独立に存在するが, 本システムにおいて整理統合し, 利用者の要望に応じて, 国内標準の職業 · 産業コードおよび国際標準の職業 · 産業コードのいずれかまたはすべての結果を一度に提供できるようにすることである.

もう一つは, 利用者によっては, 自動コーディング結果の信頼性が高い事例に対しては, 人手によるコーディングを省略したい場合もあるために [5], これに対応できるようにすることである.

以下, 次節で, まず, 現在利用されている職業 · 産業のコード体系および, これに対応するコーディング自動化システムについて述べた後, 3 節で関連研究について述べる. 4 節で Web による公開システムについて説明し, 最後にまとめと今後の課題について述べる.

## 2　職業 · 産業のコード体系とコーディング自動化システム

### 2.1　職業 · 産業のコード体系

各国において用いられる職業 · 産業コードの体系はさまざまであるが, 我が国の社会調査においては, 国勢調査で用いられる日本職業標準分類 · 産業分類に基づいて独自に作成された SSM 職業小分類 ·SSM 産業大分類 [1] が用いられる場合が多い [6]. したがって, 本システムでは, 国内標準のコードとして, SSM 職業小分類 ·SSM 産業大分類を対象とする.

また, 国際標準のコードとしては, 国際労働機構により作成された ISCO（International Standard Classification of Occupation）·ISIC（International Standard Industrial Classification） [4] を対象とする. ISCO や ISIC のコード体系は SSM 職業 · 産業分類と異なり, 階

[1]たとえば, 仕事の内容が「市役所庶務課で事務員（内勤）」と回答された場合, 職業コード「５５４」（総務 · 企画事務員）にコーディングする.

[2]我が国では, 職業は約 200 個, 産業は約 20 個のコードに分類されることが多い.

[3]たとえば前述の例において, 役職が「係長、係長相当職」（選択肢）であれば「５５４」であるが, 「課長、課長相当職」（選択肢）であれば「５４５」（管理的公務員）と別の職業コードとなる [2]. 職業コーディングでは, 役職や従業上の地位, 従業先事業の規模など仕事の内容以外の情報も用いられて総合的に判断される [3].

[4]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（SSJDA）の HP（http://ssjda.iss.u-tokyo.ac.jp/）を Web サイトとして用いる.

[5]このような要請は, 国勢調査のように事例が膨大に存在する場合だけでなく, 事例がそれほど多くなくても, 熟練したコーダがいなかったり, 時間的コストがかけられないような場合にも生じる.

[6]職業が小分類であるのに対して, 産業が大分類であるのは, 社会学においては, 産業は職業ほど詳しい情報を必要としないためである. この傾向は, 国際標準においても同様である.

層的な構造をとる. ISCO は小分類である 4 桁まで（約 400 個）を必要とするが, 先述の SSM 調査（2005 年）では ISIC は主に亜大分類の 2 桁まで（約 60 個）を使用していたため, 本システムにおいてもこれに従う.

### 2.2 職業 · 産業コーディング自動化システム

SSM 職業 · 産業コーディングに対して, [8] は, 職業 · 産業に関する回答（仕事の内容に出現する単語, 従業先の事業内容に出現する単語, 従業上の地位, 役職, 従業先事業の規模 [7]）を用いて, 格フレームの考え方に基づいたルールベース手法による自動化システム（ROCCO システム）を提案した. この結果, SSM 産業コードは精度 90% 以上, 再現率（以後は, 正解率 [8] とよぶ）は 75% であったため, 現在も ROCCO システムによる結果が候補として利用される. 一方, SSM 職業コードは精度は 80% であったが, 正解率が 70% に満たなかったため, 改善が必要とされた. この問題を解決するため, [9] では, サポートベクターマシン（SVM）を適用し, 素性として, 前述の職業 · 産業に関する回答に加えて, ROCCO システムによる結果を追加する手法を提案し, 正解率を 80% まで向上させることができた. この値は, 熟練したコーダには及ばないものの一般のコーダより高く, また実際には, 第 1 位に予測された結果だけでなく複数個を候補として提示するために（たとえば第 3 位に予測された結果まで）, 正解率は 85% 以上を示すことができた. SVM と ROCCO システムの組み合わせ手法は利用者から評価されている.

ISCO や ISIC は, SSM 職業小分類や SSM 産業大分類と単純な変換が行えないため [13], 別途, 自動化システムの開発が必要であると考えられた. ISCO コーディングに対してはルールベース手法が構築されていないため, SVM による方法が開発されたが, 正解率が 70% 未満であった. [11] では, [8] を参考にして, SSM 職業小分類も素性として追加する方法を提案し, 正解率を 75% まで向上させることができた. ISCO コーディングは SSM 職業コーディングに比べるとコードの数が多く, 訓練事例のサイズも小さいため, SSM 職業コーディングより正解率が約 5% 低いが, 今後, ISCO コーディングが普及するにつれて正解付きの事例が増えれば, 訓練事例が増大するために, 正解率の向上が見込める. ISIC コーディングに対しても, ISCO コーディングと同様に, SSM 産業大分類を素性として追加する方法が有効であると考えられる.

[7] 従業上の地位, 役職, 従業先事業の規模は選択回答である.

[8] 本稿における正解とは, 職業 · 産業コーディングの実施により最終的に決定されたコードをいい, 正解率とは正解事例を全事例で割った値をいう.

## 3 関連研究

ここでは, 大韓民国統計庁において Web 公開が検討されている自動化システム（Web-based AIOCS） [5] について述べる.

[5] によれば, 韓国における国内標準の職業 · 産業コードは大韓民国統計庁により作成され, いずれもレベル 4 までの階層構造をもつ. 職業コードと産業コードの数は, それぞれ 442 個, 450 個である. [5] では, 自動化のアルゴリズムとして, ルールベース手法, 最大エントロピー法（MEM）, 情報検索技術（IRT）の 3 種類を用意し, 単独またはルールベース手法と他の 2 つの方法のいずれかまたは両方を組み合わせた計 6 種類の方法が存在する. ただし, 組み合わせの意味が本システムとは異なり, 手法自体は独立のままで, ルールベース手法によりルールがマッチしなかった場合に, 別の手法を実行するだけである. 単独の方法より複数の方法を組み合わせた方が性能がよく, 特に, ルールベース手法の後に MEM を実行する方法は, 非常に高い精度を示す（98.4%）. しかし, いずれの手法もコードの決定率が低いために正解率は高くなく, もっとも高い正解率を示すルールベース手法, MEM, IRT を順に実行する方法においても 76.3% で, [9] を約 4% 下回る.

Web 画面を通じたユーザインターフェイスは, ファイルによる入出力を想定する本システムと異なり, 一問一答の伝票形式画面上に, 会社名, ビジネスカテゴリ, 部門, 役職, 仕事の内容（自由回答）を入力すると, 同一画面に結果が表示される. 利用できる権利により, 利用者を anonymous user, authorized user, administrator の 3 種類に区別する点も, 研究者を対象とし, anonymous user は想定していない本システムとは異なる.

## 4 Web 公開システム

### 4.1 システムの機能

第 1 節で述べたように, Web による公開システムでは, 利用者に対する利便性を考慮した次の 2 つの機能を検討する. 一つは, 複数種類であっても利用者の希望する職業 · 産業コードを一度に提供することで, もう一つは, 自動コーディングの結果がどの程度信頼できるかを利用者にわかりやすく提示することである.

#### 複数種類の職業 · 産業コードの提供

本システムで処理できる職業 · 産業コードの種類と用いられる手法を表 1 に示す. 表 1 において, ROCCO* は, SSM 職業小分類または SSM 産業大分類が存在しな

表 1: 本システムが処理する職業 · 産業コードと手法

| 職業 · 産業コード | 分類レベル | 手法 |
|---|---|---|
| SSM 職業コード | 小分類 | ROCCO, SVM |
| SSM 産業コード | 大分類 | ROCCO |
| ISCO | 小分類 | ROCCO*, SVM |
| ISIC | 亜大分類 | ROCCO*, SVM |

表 2: SVM における訓練事例　　単位：事例数

| 職業 · 産業コード | データセット | サイズ |
|---|---|---|
| SSM 職業コード | JGSS-2000, -2001 -2002, -2003 -2005 | 39,120 |
| ISCO | 2005SSM | 16,089 |
| ISIC | 2005SSM | 16,089 |

い場合に, ROCCO システムを実行して入手しておく必要があることを示す. たとえば, 新規の調査データに対して ISCO の希望がある場合には, ROCCO システムから開始する必要がある. いずれの場合も, 本システムは利用者の要求に応じて, 表 1 の任意のコードに対する該当プログラム群を選択し, 複数種類の結果を同時に提供することができる.

#### 結果に対する信頼性の提示

自動化システムによる結果に対する信頼性として, 機械学習の場合には, 予測されたクラスに対して推定するクラス所属確率 [12] を利用することができる. 本システムでは, 利用者にわかりやすくするために, クラス所属確率の値を次の 3 つのレベルに区別し, 候補となる各コードごとに提示する.

A．人手によるコーディングは不要
B．できれば人手によるコーディングを行う方がよい
C．人手によるコーディングが必要

利用者は, この記号を参考にしてコーダの作業量を減らすことができる. ただし, 各レベルの閾値をどのように決めるかについては今後の検討課題である.

### 4.2　システム構成

本システムは, これまでに多様な言語により開発された種々の自動化システムを新たに C 言語で統一する. ただし, 一部は Java や ruby のライブラリを利用している [9]. 以下では, ROCCO システムにおけるルール辞書とシソーラスおよび, SVM で用いられる訓練事例について述べる.

ROCCO システムは, 3 種類のルール辞書と 2 種類のシソーラスをもつ [8]. ルール辞書 α は, 仕事の内容に出現する単語から抽出された述語相当語と名詞との関係を格フレームにより表現して SSM 職業コードと関連づけたルール（3,524 個）から構成される. ルール辞書 β は, ルール辞書 α により決定された仮の職業コードに対して, 役職や従業上の地位, 従業先事業の規模の情報によるチェックを行って, （自動化システムとして）最終決定するためのルール（27 個）から構成される. ルール辞書 γ は, ルール辞書 α の SSM 産業コード版で, 512 個のルールから構成される. ここで, ルールを適用するには, 回答に出現する述語相当語と名詞における表記の揺れや同意語をグループ化しておく必要があるが, このためのシソーラスとして, 述語シソーラス（見出し語 2,900 個）と名詞シソーラス（見出し語 316 個）をもつ.

SVM において訓練事例として用いられているデータセットとサイズを, 職業 · 産業コード別に表 2 に示す. 未整理の状態で存在する正解付きのデータセット [10] もあるために, 今後, 訓練事例のサイズをより拡大することが可能である. また, 将来, 新規の職業 · 産業コードが要請された場合も, 訓練事例を準備できれば, 適宜対応は可能である.

### 4.3　利用方法

利用者は, まず, SSJDA のＨＰ [11] に掲載された利用に関する説明を読み, 所定の形式の入力用データファイルを作成する. その後, 以下に示す（1）から（4）の手続きにしたがって自動コーディングの結果を得る（図 1 参照）.

（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, パスワードの発行およびアップロード（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

本システムは手続き（3）と（4）の間で稼働するが, SSJDA 側の担当者が容易に操作できるように, オペレータ用画面の工夫を行っている（図 2 参照）.

[9] 前処理である形態素解析は, これまでと同様に [6] を利用する.

[10] たとえば, SSM 職業コードの場合, 2005SSM, JGSS-2006, -2008, -2010（総計約 22,000 サンプル）が存在する.

[11] 現在は本システムの予告が掲載されている.

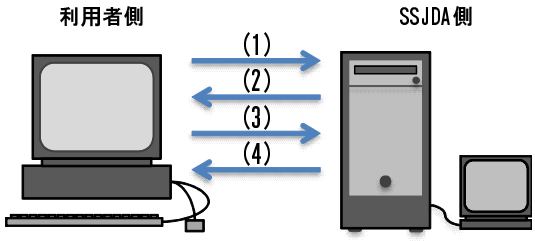


図 1: 利用手順


図 2: オペレータ用画面（SSJDA 側）

## 5　おわりに

本稿では, 多大な労力と時間を要する職業 · 産業コーディングにおいて, コーダを支援するために開発された自動化システムの Web 公開版の構築について述べた. Web 公開版システムでは, 利用者の希望により, 社会調査で用いられる職業や産業の国内 · 国際標準コードのいずれかまたはすべてを一度に提供でき, また, 現在は未完成であるが, 自動コーディングの結果に対する確信度を付与することができる.

今後の課題として早急に取り組むべきことは次の 3 つである. まず, 現在までに完成した部分について Web 公開を行い, 利用者や SSJDA の担当者による評価の結果, 改善点があれば対応する. 次に, 自動化システムによる結果に対する確信度付与の機能を追加する. さらに, SVM における分類精度を向上させるために, 未整理の状態にある正解付きデータセットを, 適宜, 訓練事例として活用する. 本システムは, 新規のコード体系をもつ職業 · 産業コーディングに対しても, 正解付きのデータセットがあれば, 拡張は容易である.

**謝辞**　2005 年 SSM 調査データの利用に関して, 2005 年 SSM 調査研究会の許可を得た. 日本版 General Social Surveys（JGSS）は, 大阪商業大学 JGSS 研究センター（文部科学大臣認定日本版総合的社会調査共同研究拠点）が, 東京大学社会科学研究所の協力を受けて実施している研究プロジェクトである.

本研究は科研費（22530516）の助成を受けたものである.

## 参考文献

[1] 1995 年 SSM 調査研究会. 2006. SSM 産業分類 · 産業分類（95 年版）.

[2] 1995 年 SSM 調査研究会. 2006. 1995 年 SSM 調査コード · ブック.

[3] 原純輔. 1984. 社会調査演習. 東京大学出版会.

[4] Bureau of Statistics; International Labour Office. 2001. Coding Occupation and Industry. Bureau of Statistics; International Labour Office.

[5] Y. Jung, J. Yoo, S-H. Myaeng and D-C. Han. 2008. A Web-based Automated System for Industry and Occupation Coding. In *Proceedings of the Ninth International Conference on Web Information Systems Engineering (WISE-08)*, LNCS, pp.443–457.

[6] 黒橋禎夫, 長尾真. 1998. 日本語形態素解析システム JUMAN version 3.61. 京都大学大学院情報学研究科.

[7] 盛山和夫. 2004. 社会調査入門. 有斐閣.

[8] 高橋和子. 2000. 自由回答のコーディング支援について ー格フレームによる SSM 職業コーディング自動化システムー. 理論と方法　Vol.15 No.1, pp. 149–164.

[9] 高橋和子, 高村大也, 奥村学. 2005. 機械学習とルールベース手法の組み合わせによる自動職業コーディング. 自然言語処理　Vol.12 No.2, pp. 3–24.

[10] 高橋和子, 須山敦, 村山紀文, 高村大也, 奥村学. 2005. 職業コーディング支援システム（NANACO）の開発と JGSS-2003 における適用. 日本版 General Social Surveys 研究論文集［4］　JGSS で見た日本人の意識と行動, pp. 225–242.

[11] 高橋和子. 2008. 機械学習による ISCO 自動コーディング. 2005 年 SSM 調査シリーズ１２　社会調査における測定と分析をめぐる諸問題, pp.47–68.

[12] K. Takahashi, H. Takamura, and M. Okumura. 2008. Direct estimation of class membership probabilities for multiclass classification using multiple scores. In *Knowl Inf Syst* 19(2), pp.185–210. Springer London.

[13] 田辺俊介. 2006. ISCO と SSM 職業分類の相違点の検討ー国際比較調査における職業データに関する研究ノートー. 社会学論考　Vol.27, pp. 53–78.