総務省統計オープンデータを利用した パーソナライゼーションサービスへの適用可能性

深澤佑介　太田順

東京大学　人工物工学研究センター

yusuke.fukazawa@gmail.com

概要　本稿では総務省が公開するオープンデータを利用したパーソナライゼーションサービスの利用可能性について検討する。

キーワード　オープンデータ、サービスの個人化、情報推薦

1 はじめに

2013 年 6 月総務省が過去に調査した統計データを API 化して提供を開始した。API 化することにより、統計データの民間利用の活性化が期待される。民間では、自社データと突き合わせることにより、市場規模の把握や、市場開拓につなげることが可能になる。一方、統計データはユーザ個人の情報は提供されない。そのため、一見パーソナライズサービスには利用するのが難しいように見える。本稿ではオープンデータをパーソナライゼーションサービスという観点でどういった利用方法があり得るか検討する。

2 オープンデータの API 化

総務省オープンデータ［1］は、政府統計の総合窓口（e-Stat）で提供している国勢調査、人口動態調査、国民生活基礎調査などの統計調査を提供している。

オープンデータの API では以下の 3 種類の API を提供している。統計表情報取得 API、メタ情報取得 API、統計データ取得 API である。手順として、上記の 3 つの API を順番に利用する。統計表取得 API により、統計表情報を取得する。統計表情報には、統計表 ID、調査名、統計表名、調査年月等が含まれる。次に、メタ情報取得は、統計表 ID をもとにして、統計表に含まれるメタ情報（時間軸、地域事項、分類事項）を取得する。さらに、実際の統計データは、統計データ取得 API を利用し、統計表に収録されている統計データ（数値データ）を取得する。必要に応じて、メタ情報による絞り込みを行うことができる。

著者らは 2013 年 6 月時点の全データをダウンロードした。統計表は、27,258 種類の統計データが収録されていた。統計データは 50GB 程度のデータ量となった。以下の章で著者らは統計表間のクロス解析を行うが統計表間を突き合わせるための共通要素として時間軸（年）に着目する。時間軸（年）のデータがあった統計表の個数は 27,258 個中 516 個であった。たとえば、労働力調査、基準消費者物価指数、住民基本台帳人口移動報告、個人企業経済調査、家計調査などが該当する。ただし、516 個の統計表において、それぞれすべての年のデータがあるわけではなく、統計表によって偏りがある。1980 年~2013 年の間の統計表の個数を調査した結果を図1に示す。2002 年から 2009 年の間は一定以上の個数の統計表が安定して格納されているが、2011 年から格納されている統計表の個数が減少している。この原因については、現時点では明らかではない。

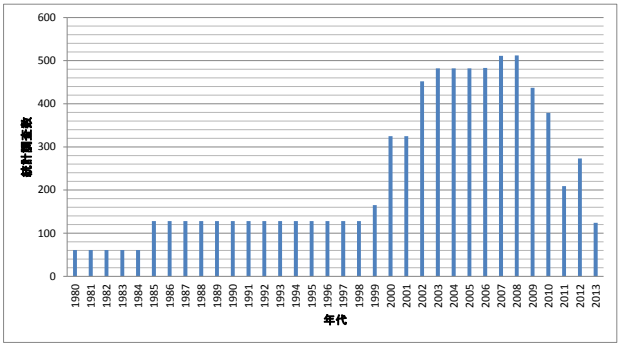

図 1　年代ごとの統計表個数の調査結果

3 パーソナライゼーションサービスについて

サービスの個人化や情報推薦で必要なことは、ユーザの属性情報や嗜好情報、行動情報を予測・推定することである。近年、ユーザの購買ログ、クリックログなどの Web 上の行動履歴や、Twitter などのソーシャルネットワーク上への書き込みなどからユーザの趣味や嗜好を推定する手法が提案されている。しかしながら、ほとんどのサービスでは、ユーザのすべてのデータを把握するのは難しく、未知の情報については既知の情報から推定するしかない。そこ

で、未知の情報と既知の情報の間を埋める役割として、オープンデータ化された統計データの利用に着目する。

国勢調査では性別、年代だけでなく、家族構成など様々なデモグラフィック情報に関する統計データを掲載している。そこで、デモグラフィック情報間の相関関係を見ることにより、相関の高いデモグラフィック情報の組み合わせが分かる。このデータを利用することにより、デモグラフィック情報の一部が分かっている場合、ユーザのそのほかの属性情報の推定に役立つ可能性がある。また、労働力調査、就業構造基本調査、民間給与実態統計調査などの労働力調査からは、国民の就労状況に関する調査が行われている。そこでは、転職や就職、退職などの労働に関する統計データが格納されている。国勢調査と組み合わせることで、デモグラフィック情報と就労状況間の相関を観察することができる。このように単一の統計表のみでなく、複数の統計表を組み合わせて解析することにより、未知の情報と既知の情報の間を埋めることができる可能性が高まると考えられる。

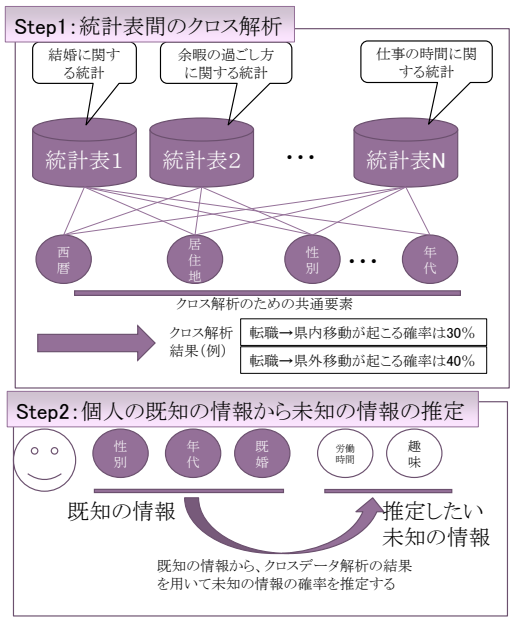

図 2　総務省オープンデータのパーソナライゼーションサービスへの活用方法の案概要

クロスの統計データの解析によるパーソナライゼーションサービスの活用方法について図 2 に示す。第一に、統計データ間を突き合わせ相関のある統計データを抽出する。異なる統計データを突き合わせるためには、突き合わせるための共通的な指標が必要である。共通的なパラメータとしては年代、居住地、性別などがありうる。たとえば、労働力調査の転職する人の動向と、国勢調査の県内、県外移動をする人の数の動向を時間軸という共通の軸で比較することで、相関の有無を調べる。相関があれば、転職というイベントが発生した時に県外に引っ越す可能性が高いか、県内の引っ越しで済む可能性が高いか、予測をすることができるようになる。

第二に、ユーザの既知の情報に基づき、未知の情報を推定する。クロス解析結果では、ある事象が発生したときに別の事象が発生する確率としてとらえることができるため、ユーザの未知の情報の発生確率を推定することが可能になる。

第一のステップについて、実際のデータから相関分析を行った。その結果を表 1 に示す。表 1 から転職というイベントの増減に対して、都道府県をまたがる引っ越しは 0.61 とある程度の相関をもって増減していることが分かる。一方、県内移動については −0.21 とあまり相関はない。このことから、転職というイベントが分かったユーザに対しては県をまたがる移動が起こる可能性が高いことが予測することができる。

表 1　転職と県内移動、県外移動の相関関係

			1	2	3
1	都道府県間移動者数の推移	住民基本台帳人口移動報告 平成24年住民基本台帳人口移動報告	1.00	0.85	0.61
2	都道府県内移動者数の推移	住民基本台帳人口移動報告 平成21年住民基本台帳人口移動報告	0.85	1.00	-0.27
3	離職期間別前職のある就業者数（転職者）	労働力調査 詳細集計　全都道府県 全国 年次	0.61	-0.27	1.00

しかしながら以下の課題を有する。

・統計表に格納されている統計のデータが、サービスが推定したいユーザの情報と一致するかどうか不明である。従来パーソナライズサービスで推定対象とされてきた情報とどの程度一致するか調査が必要である。

・クロス解析の結果について因果関係が不明確である。結果として相関があった場合でも、別の要因で両方の事象の増減、動向が一致する場合も考えられる。相関関係の信頼度を高めるため、統計データの背後の要因について分析を加えることが必要である。

4 おわりに

本稿では総務省が公開するオープンデータを利用したパーソナライゼーションサービスの利用可能性について検討した。今後、上記で述べた課題を解決し、実サービス上で総務省オープンデータを用いたパーソナライズ化の効果を検証する。

参考文献

[1] 総務省統計局, 次世代統計利用システム, API 機能の概要と活用事例, 2013.
http://www.soumu.go.jp/main_content/000230118.pdf