e-Statにおける検索漏れを抑止する情報システムの開発とその検証

Development of Front-End Search System improving Recall in e-Stat

芦澤颯太† 松田純一 大曽根匡†
Souta ASHIZAWA† Junichi MATSUDA Tadashi OSONE†

†専修大学大学院　経営学研究科
† Graduate School of Business Administration, Senshu University.

要旨

e-Stat とは，政府統計の総合窓口のことであり，各府省等が実施する統計調査の各種情報を 1 つにまとめ，統計データの検索をはじめとした，さまざまな機能を備えたポータルサイトである．そして，政府が発信する統計データの源泉として多くの研究者や個人・企業などに活用されている．しかし，実際に使用してみると，検索漏れなどの課題があることがわかった．本研究では，その課題である検索漏れを抑止するフロンドエンドシステムの開発と，その検証結果について報告する．

1. はじめに

近年，IT の発展により，データサイエンスの活用が叫ばれてきている．そこで注目を集めているのがオープンデータである．オープンデータとは，デジタル庁が公開しているオープンデータの基本指針[1]によれば，「国，地方公共団体及び事業者が保有する官民データのうち，国民誰もがインターネット等を通じて容易に利用（加工，編集，再配布等）できるよう，次のいずれの項目にも該当する形で公開されたデータをオープンデータと定義する」としている．いずれの項目として，①営利目的，非営利目的を問わず二次利用可能なルールが適用されたもの，②機械判読に適したもの，③無償で利用できるものの 3 点を挙げている．

日本におけるオープンデータとして e-Stat[2]という政府統計のまとめサイトが，国によって公開され広く利用されている．e-Stat は，2008 年より運用が開始された．それ以前は，各省庁が独自に統計データを管理し，情報提供を行ってきたので，利用者は省庁毎にアクセスが必要であった．それが，e-Stat のサービス提供により，各省庁の検索機能の重複排除や利用者への効率的な情報提供が可能になり，利用者の利便性が大幅に改善された．しかし，各省庁が別々に管理していたデータを 1 つにまとめたことによる影響や，日本経済新聞の記事[3]が指摘している省庁間の縦割り体質などによって，検索もれや検索ノイズなどの課題がある．そこで，本研究では，主な課題である検索漏れを抑止するフロントエンドシステムを開発した．手始めに，複数の省庁で広く使われており，表記ゆれも多く発生している「性別」に関する用語を対象とした．本論文では，そのシステム開発と検証結果について報告する．

2. e-Stat の管理構造

e-Stat は，現在約 30 の府省から集められた約 710 の調査を閲覧することが可能になっている．主な機能として，検索機能と活用機能がある．検索機能では，分野，組織，キーワードで検索が可能になっており，分野では，「人口・世帯」や「労働・賃金」など 17 の分野から，組織では「総務省」や「文部科学省」など約 30 の府省の統計データを利用できる．キーワード検索を用いて，利用者の目的に合ったキーワードで全文検索できる．活用機能は，グラフ，時系列表，地図，地域の 4 つの方法で統計データを表示できる機能である．

統計データは，階層的に管理されている．その統計データの管理構造を図 1 に示す．約 710 の調査のそれぞれの「統計調査と統計概要」の下位に調査年や都道府県別などの「提供分類」がある．そして，その下位に「統計表」がある．統計表の下位に「事項名」があり，事項名の下位にデータの名称を表す「項目名」がある．統計表の具体的な管理構造の実例を図 2 に示す．これは，令和 2 年の国勢調査にお

ける統計表「男女別人口」を簡略化したものである．図 2 の事項名「男女」では，下位に項目名として「総数」と「男」と「女」がある．

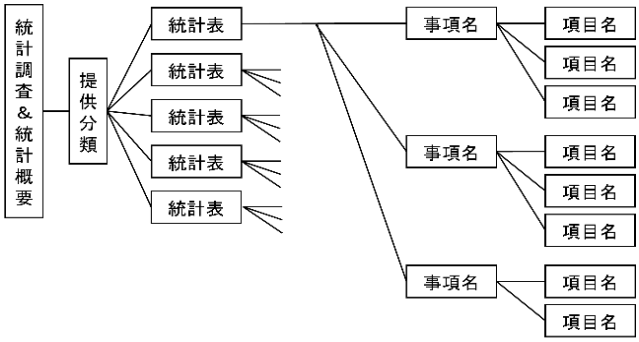

図 1　統計データの管理構造

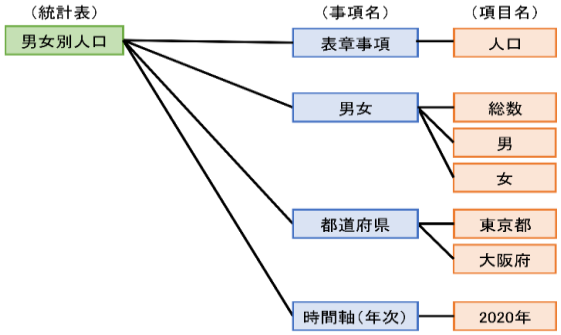

図 2　統計表の具体的な管理構造の例

3. 検索漏れと検索ノイズ

e-Stat 使用上の課題には，使い勝手の観点と，検索の正確性の観点からの課題がある．本研究では検索の正確性の観点の課題である検索漏れの抑止に焦点を当てた．

3.1. 検索漏れ

検索漏れとは，キーワード検索した場合，本来ヒットしなければならない統計表がヒットしない事象である．検索漏れと次に説明する検索ノイズの起こる原因として，ユーザが入力する検索キーワードが人によって異なることと，e-Stat 側のデータの表記不統一が考えられる．検索漏れの例として，図 3 の 2 つの事項名を考える．図 3 の左側の事項名は「性別」であり，右側の事項名は「男女別」である．そして，項目名はどちらも「男」と「女」である．したがって，性別に関するキーワードで検索した場合，どちらもヒットしなければならない．ところが，現在の e-Stat でユーザがキーワード「性別」で検索した場合，事項名が「男女別」の統計表はヒットしない．ユーザがキーワード「男女別」で検索した場合でも事項名が「性別」の統計表はヒットせず，検索漏れが発生する．

もう 1 つの検索漏れの例を図 4 に示す．事項名は「職種_128 職種区分」で右側の項目名は「プログラマー（男）」と「プログラマー（女）」である．この例では「性別」でキーワード検索しても，「男女別」でキーワード検索してもヒットしない．

性別	男
	女

男女別	男
	女

図 3　事項名不統一の検索漏れの例

職種_128職種区分	プログラマー（男）
	プログラマー（女）

図 4　検索漏れの例

3.2. 検索ノイズ

検索ノイズは，本来ヒットすべきでない統計表がヒットする事象である．検索ノイズの例を図 5 に示す．統計表 1 の事項名は「性別」，統計表 2 の事項名は「世帯属性別」である．ユーザが「性別」で検索した場合，統計表 1 と統計表 2 のどちらもヒットするが，統計表 2 は「属性別」の統計表であり，「性別」に関係しない統計表なので検索ノイズである．検索ノイズが多くあると，検索結果からユーザの所望する統計表だけを抽出するのに時間がかかってしまう．

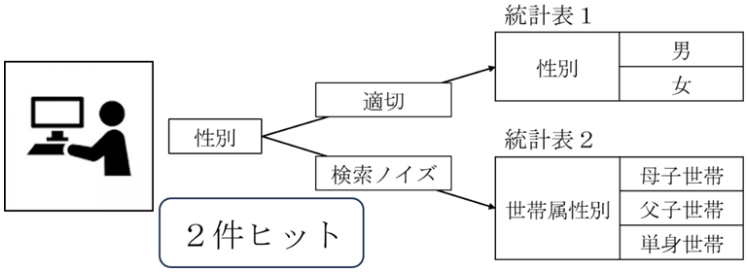

図 5　検索ノイズが起こる例

3.3. 検索漏れと検索ノイズの評価尺度

本研究では，検索の評価尺度として，2 種類の指標を用いる[4].

①　再現率

再現率は，検索漏れの評価尺度である．この指標は網羅性の指標で，適切な統計表のうち，どのくらいの数がヒットしたかを見るものである．例えば，適切な統計表が 10 で，検索にヒットした適切な統計表が 8 の場合は，0.8 になる．再現率の式は以下の通りである．再現率が 1 に近いほど検索漏れが少ない．

$$再現率＝\frac{検索にヒットした適切な統計表}{適切な統計表}$$

②　適合率

適合率は，検索ノイズの評価尺度である．正確性の指標で，検索にヒットした統計表のうち，適切であった統計表である確率の事である．例えば，検索にヒットした統計表が 10 で，そのうち，適切な統計表が 7 である場合は，0.7 になる．適合率の式は以下の通りである．適合率が 1 に近いほど検索ノイズが少ない．

$$適合率＝\frac{検索にヒットした適切な統計表}{検索にヒットした統計表}$$

4. 開発システムの概要

開発したフロントエンドシステムは，統計表検索システムと拡張事項名生成プログラムで構成した．その構成を図 6 に示す．統計表検索システムで使用されているキーワード変換テーブルは，ユーザ側の表記ゆれを吸収させるためのテーブルである．一方，統計表データテーブルは e-Stat 側の表記不統一を吸収するためのテーブルである．統計表データテーブルは，拡張事項名生成プログラムを使用してあらかじめ準備しておく．これらのテーブルについては後述する．

開発システムにおける検索の流れは以下の通りである．①ユーザが入力した検索キーワードがキーワード変換テーブルの「用語」に一致すれば「代表語」に変換する．②変換した「代表語」を用いて統計表データテーブルを参照する．③その「代表語」が拡張事項名にヒットした統計表 ID を用いて統計表 URL リストを作成する．最後に，④その URL リストからユーザが求めている統計表を閲覧する．

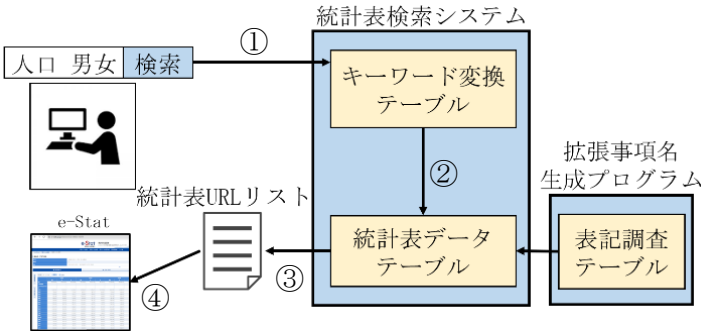

図 6　開発システムの構成

次に，開発システムで用いるキーワード変換テーブルと統計表データテーブルについて説明する．性別については，e-Stat を調査し，下記の図 7 のような同義語（横のつながり）と階層構造（縦のつながり）を考え，表記ゆれや表記不統一の問題を解決するために開発システムに利用した．

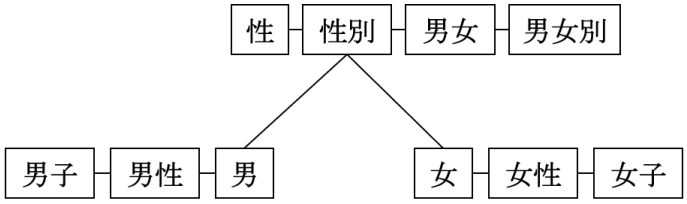

図 7　同義語と階層構造

（1）統計表検索システム

① キーワード変換テーブル

キーワード変換テーブルは，ユーザ側の表記ゆれを解消するために用いる．ユーザによって入力された入力キーワードを「代表語」に変換するテーブルである．性別に関する用語の場合，ユーザが入力した「性」や「男女」や「男女別」は，代表語の「性別」に変換される．

② 統計表データテーブル

統計表データテーブルは e-Stat 側の表記不統一を吸収するためのテーブルである．このテーブルは，統計表名，事項名，拡張事項名，項目名，統計表 ID などのフィールドで構成した．統計表のデータは，e-Stat の API を用いて作成した．拡張事項名は，e-Stat 側の表記を統一するためのフィールドである．詳しくは，（2）の拡張事項名生成プログラムで説明する．

表 1　キーワード変換テーブル

用語	代表語
性	性別
男女	性別
男女別	性別

表 2　統計表データテーブル

統計表名	事項名	拡張事項名	項目名	統計表 ID
男女別人口	男女	性別	総数	0001
男女別人口	男女	性別	男	0001
男女別人口	男女	性別	女	0001

（2）拡張事項名生成プログラム

このプログラムは, 統計表データテーブルに拡張事項名を追加するプログラムであり, 事項名や項目名に性別に関する用語があれば, 拡張事項名に代表語の「性別」を追加する．このプログラムでは, 表 3 のような表記調査テーブルを用いており, 統計表データテーブルの項目名か事項名に調査用語があり，かつ，一致種別に該当した場合，代表語を統計表データテーブルの拡張事項名に追加する．一致種別は，性別に関係のない用語を排除するために作成した．例えば，表 3 の 4 レコード目の「_男」を例に挙げると，「_男」に一致種別を設けなければ，「人口_男」も「秋田県_男鹿市_計」もヒットしてしまう．一致種別の「後方一致」を設けることにより，性別に関係のない「秋田県_男鹿市_計」はヒットせず，検索ノイズを減らすことができる．表記調査テーブルは，性別に関する用語を対象として作成し，調査用語は全部で 98 種類となった．表 3 では一部分を抜粋して記載している．このようにして， e-Stat 側の表記統一を行った．

表 3　表記調査テーブル（一部抜粋）

番	調査用語	追加用語	階層種別	一致種別
1	男	性別	項目名	完全一致
2	女	性別	項目名	完全一致
3	男-	性別	項目名	前方一致
4	_男	性別	項目名	後方一致
~	~	~	~	~
98	性ー	性別	事項名	前方一致

5. 検証結果

「性別」と「性」と「男女」と「男女別」の 4 つの検索キーワードを用いて，従来方式と提案方式で検証を行った結果を述べる．提案方式は，4 つの検索キーワードをすべて「性別」に変換しているため，検証結果は同じ値となる．また，検索漏れの指標が再現率で，検索ノイズの指標が適合率である．ここでは，社会教育調査と人口推計の 2 つの調査に対する検証結果について報告する．

（1）社会教育調査

社会教育調査は，文部科学省[5]によれば，社会教育行政に必要な社会教育に関する基本的な事項を明らかにすることを目的としており，将来の教育政策に関わってくる統計調査である．統計表数は 1,219 であり，そのうち，性別を区別している適切な統計表は 207 であった．

① 検索漏れ数と検索ノイズ数

社会教育調査での検索漏れ数と検索ノイズ数を表 4 に示す．検索漏れ数をみると，従来方式では 80 件から 200 件ほどであったが，提案方式では 6 件まで激減し，提案方式の有効性が確認できた．一方，検索ノイズ数については，「性」に関しては 148 件から 76 件に半減した．しかし，「性別」「男女」「男女別」については検索ノイズが増加してしまった．

表 4　社会教育調査における検索漏れ数と検索ノイズ

	検索漏れ数		検索ノイズ数	
	従来方式	提案方式	従来方式	提案方式
性	85	6	148	76
性別	110		5	
男女	192		15	
男女別	207		0	

② 再現率と適合率

次に，再現率と適合率の結果を表 5 に示す．従来方式では 0.6 以下であった再現率が，提案方式では 0.971 となり，検索漏れが極めて少なくなったことがわかる．

従来方式の適合率では，「性別」は 0.951 でそれ以外は半分以下である．それに対し，提案方式の適合率は 0.726 となり，「性別」については減少してしまったが，「性」「男女」「男女別」の適合率は改善した．

表 5　社会教育調査における再現率と適合率

	再現率		適合率	
	従来方式	提案方式	従来方式	提案方式
性	0.589	0.971	0.452	0.726
性別	0.469		0.951	
男女	0.072		0.500	
男女別	0.000		0.000	

③ 適合率低下の例

適合率低下の例を表 6 に示す．この統計表は，施設等別の職員数をまとめている表で，項目名に「女性」という性別に関する用語があるが，この場合の「女性」は性別を区分するものではなく，施設について述べており，適切な結果ではない．

表 6　適合率低下の例

分類	データ
統計表名	施設等別職員数（政令指定都市・市・町村別）
事項名	施設等別−職員数区分別
項目名	女性教育施設−計

（2）人口推計

人口推計とは，総務省統計局[6]によれば，国勢調査による人口を基準にし，出生や死亡，出入国，転出入などの人口動向から毎月の人口を求め，将来人口の推計に役立てている．人口推計の統計表数は 517 あり，そのうち，性別を含んでいる適切な統計表は 428 であった．

① 検索漏れ数と検索ノイズ数

人口推計における検索漏れ数と検索ノイズ数を表 7 に示す．検索漏れ数をみると，従来方式での「性」と「性別」では，適切な統計表 428 件のうち，360 件と 404 件で 8 割程度であったが，提案方式では 0 件になり，検索漏れを防ぐことができた. 検索ノイズは，従来方式と提案方式どちらも 0 であった.

表 7　人口推計における検索漏れ数と検索ノイズ数

	検索漏れ数		検索ノイズ数	
	従来方式	提案方式	従来方式	提案方式
性	360	0	0	0
性別	404		0	
男女	0		0	
男女別	43		0	

② 再現率と適合率

人口推計での再現率と適合率の結果を表 8 に示す. 提案方式では，再現率が 1.0 となり，検索漏れがなくなったことが検証できた．さらに，適合率も 1.0 となり，検索ノイズもないことがわかった，

表 8　人口推計における再現率と適合率

	再現率		適合率	
	従来方式	提案方式	従来方式	提案方式
性	0.159	1.000	1.000	1.000
性別	0.056		1.000	
男女	1.000		1.000	
男女別	0.900		1.000	

6. まとめ

本研究では，e-Stat における課題である検索漏れや検索ノイズを抑止するシステムを開発し検証を行った．検証を行った結果，一定の成果を得ることができた．しかし，検索漏れや検索ノイズがすべてなくなったわけではなく，より精度の高い方法を考える必要がある．また，今回は性別に関する用語で検証を行ったが，他の用語に関しても調査する必要がある.

参考文献

[1] デジタル庁, “オープンデータ，オープンデータ基本指針,” https://www.digital.go.jp/resources/open_data/, 2023.10.1 参照.

[2] 総務省統計局, “e-Stat 政府統計の総合窓口,” https://www.e-stat.go.jp, 2023.10.1 参照.

[3] 日本経済新聞, “政府統計，8 割がデータ検索できず　縦割りが浮き彫り,” 2021.9.1, https://www.nikkei.com/article/DGXZQOUA31AJD0R30C21A8000000/, 2023.10.1 参照.

[4] 絹川博之，田中和明，池上信男, “日本語情報検索システムにおけるキーワードの自動抽出,” 日立評論 5 月号, 1982, pp.75-78.

[5] 文部科学省, “社会教育調査 - 文部科学省,” https://www.mext.go.jp/b_menu/toukei/chousa02/shakai/index.htm, 2023.10.1 参照.

[6] 総務省統計局, “統計局ホームページ/人口推計,” https://www.stat.go.jp/data/jinsui/1.html, 2023.10.1 参照.

[7] 関西学院高等部数理科学部, “小中学生のための統計情報ポータルサイト「e-Stat Junior」の提案,” STAT DASH グランプリ, 2016, https://www.e-stat.go.jp/api/sites/default/files/uploads/Policy-3_Sasaki_DAIJIN-1.pdf, 2023.10.1 参照.

[8] 芦澤颯太，松田純一，大曽根匡, “e-Stat での統計データ検索におけるいくつかの課題抽出とその解決方法の提案,” 情報システム学会，第 18 回全国大会・研究発表大会, 2022.