G-02

統計表データを用いた動向情報の根拠探索システムの検討

Study on Evidence Search System for Trend Information on the Web using Statistical Tables.

松井 侑祐†　　宮森 恒†
Yusuke Matsui　　Hisashi Miyamori

1.　はじめに

近年、インターネットの急速な発展・普及に伴い、膨大な量の情報が溢れている状況となっている。その中には、質が高くて役に立つ情報からデマや誤った情報など、質の高いものから低いものまでが混在し、どれが正しい情報なのかをユーザが選別することはますます困難となっている。Facebook や Twitter、ブログなど、情報発信手段は発達する一方、信頼度の高い情報を見分けるための手段は現状で十分に整っているとはいえない。

ここで、信頼度を見分ける方法の一つに、客観的根拠の有無に着目した方法が挙げられる。例えば、電子掲示板等に掲載された匿名発信者による真偽の不明確な情報について、何らかの情報源から適切な客観的根拠を見つけることができれば、ユーザ自身がその情報について信頼度を判断するのは格段に容易になると考えられる。しかし、一般に、ユーザ自身が情報を閲覧するたびに適切な情報源を探すことは、ユーザに大きな負担となり現実的でない。また、ユーザ自身で適切な情報源を見つけられないことも十分に考えられる。従って、情報の客観的な根拠を効率よく探索できるシステムの実現が望まれる。

一方で、政府や企業により、膨大な各種データが統計表データとして Web 上で広く公開されている。これらの情報は、政府や企業が調査・集計したものであり、一般に信頼度の高い情報源と考えることができる。ただし、これらのデータは Excel 形式や CSV 形式による表形式で公開されており、中には見出しを階層関係にするなど複雑な形式のものも存在する。コンピュータでこのようなデータを扱うには表の認識や分析といった課題を克服する必要がある。

そこで、本稿では、ソーシャルネットワーク上の言説、特に、ある対象の時間的変化を記述した動向情報を対象に、その客観的根拠を統計表データから探索するシステムを提案する。ユーザが、ソーシャルネットワーク上で見つけた信頼性を判断したい情報をクエリとして入力すると、そのクエリの根拠となる統計表データ（あるいは関連するが矛盾する統計表データ）が探し出され、理解しやすい形に変換された上でユーザに提示される。これにより、ユーザによる情報の信頼性判断を効率よく支援することができると期待される。本稿では、試作システムの実装、および、評価実験による利点と問題点について考察する。

本稿の構成は、以下の通りである。2 節では、関連研究を紹介し、提案手法の位置づけを明確化する。3 節では、提案手法について詳細に述べる。4 節では、評価実験とその結果を示し、考察を行う。最後に 5 節で、まとめと今後の課題を整理する。

2.　関連研究

これまでにも表やその構造を認識する研究は数多く行われている。

Kieninger ら[1]は、文書画像から表を認識する手法を提案している。ビジネスレター文書の画像を入力として、表中のセルなどのユニットを認識し、それらユニットの空間的配置からレイアウトを分析することで、表の位置とその内部構造を認識している。また、Wang ら[2]は、文書画像を対象に最適化手法を用いた表認識を提案している。予め文書画像中の線や文字の領域が大雑把に分割されたものを入力とし、ページのカラムスタイルのラベリング、表がもつ一貫性評価による表候補の絞り込み、ページ全体の分割を最大化する反復更新による最適化を行っている。本稿では、文書画像ではなく、CSV 形式の表データを解析対象としている点が異なる。

また、塚本ら[3]は、HTML 形式で記述された表を対象とし、表の項目名と項目データの境界を認識し、解像度の比較的小さい携帯端末でも見やすい形式の表に変換する手法を提案している。表の行間あるいは列間で類似度を定め、類似度が低い場合には行間あるいは列間に内容的な切れ目があると認識する。本稿では CSV 形式で記述された表を対象としている点が異なるが、表項目の階層関係の認識も含めた見出し部分の解釈手法については本研究と共通している。

表を利用した応用システムに関する研究も数多く行われている。

Kerpedjiev[4]らは、マルチメディアプレゼンテーションを生成するシステム AutoBrief を提案している。アナリストや専門家が大量のデータ集合中のパターンや変化を理解することを支援し、そこで得た情報をテキストと図表によって要約し、プレゼン目的に応じて適切な形式で提示するシステムを、輸送計画のシミュレーションを例に実装している。Wang ら[5]は、HTML 形式で記述された比較的大きな表を、画面サイズの小さい携帯端末でも見やすい形式の表に変換するシステムを提案している。HTML で記述された表を、データテーブルとレイアウトテーブルに分類し、データテーブルに対しては、元々の構造情報が保存されるようにしながら、表全体を 1 カラムビューに収めることで閲覧体験の向上を目指している。松下ら[6]は、統計 DB 等から得られる時系列数値情報と、それに関連する内容の一連のテキスト情報を関連付けて視覚化し、ユーザの探索的データ分析を支援する可視化インタフェースを提案している。佐伯ら[7]は、テレビ番組中の図表画像を利用し、Web 上の動向情報の根拠を探索・提示するシステムを提案している。テレビ番組中に登場する図表画像を信頼性の高い情報源と考え、信頼性の不確かな Web 上の動向情報と関連づけることで、ユーザによる情報の信頼性判断を効率よく支援することを目指している。本研究は、政府や企業が公開する統計表データを信頼性の高い情報源と考え、信頼性の不確かな Web 上の動向情報の根拠を探索するシステムについて検討する。

† 京都産業大学, Kyoto Sangyo University

3. 提案手法

3.1. 提案システム

本システムの目的は、信頼度の不確かな Web 上の動向情報をユーザがクエリとして入力することで、その客観的根拠となるような統計表データを効率よく探索・提示することである。これにより、クエリとして入力された情報の信頼性判断を促進することが期待される。

3.2. システム構成

本システムの構成を図 1 に示す。

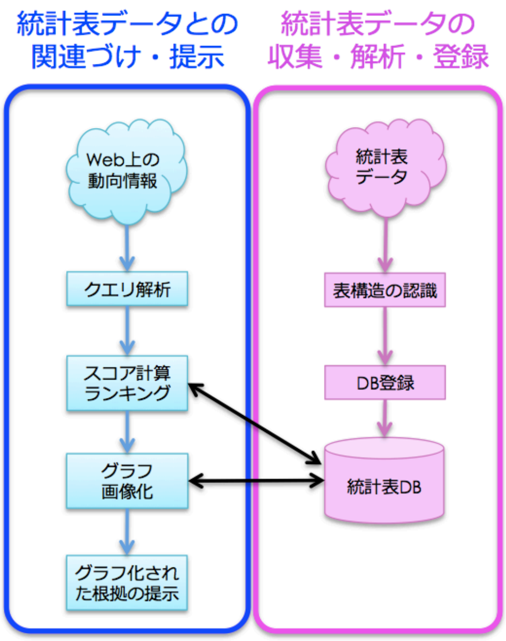

図 1　本システムの構成

本システムは大きく 2 つの部分から構成される。

統計表データの収集・解析・登録では、まず、政府や企業等により Web 上で公開されている信頼性の高い統計表データを収集する。次に、統計表データの構造を認識する。具体的には、表のタイトルや見出し、数値、注釈、単位に当たる項目とそれらの構造を認識する。これには、見出しの階層関係の解釈も含まれる。最後に、解析された統計表データを DB に登録する。

統計表データとの関連づけ・提示では、ユーザから入力された Web 上の動向情報をクエリとして解析し、統計表 DB 中の統計表データと関連づける。関連度の高い統計表から該当部分をグラフ画像化し、ユーザに提示する。

3.3. 統計表データの収集

提案システムの目的を実現するためには、信頼性の高い統計表データを収集する必要があるため、本稿では、政府が公開している政府統計[8]を収集対象とした。

政府統計では、各府省等が収集した各種統計が取りまとめられ、Excel 形式や CSV 形式等で Web 上に公開されている。公開されている政府統計のうち、本稿では日本語で記述されている CSV 形式の統計表データを収集対象とした。さらに、収集した統計表データの中にはライブラリでの読み込みに失敗するデータも存在した。本稿では、それらを除いた CSV 形式の統計表データ 49,961 件を用いることとした。

3.4. 統計表構造の認識

3.4.1. 認識対象

統計表データでは、1 つの表に以下の項目が含まれることがほとんどである。

- タイトル
- 注釈
- 上側の見出し
- 左側の見出し
- 数値

ここでは、これらの項目に対応するセル(区画)またはセル範囲を認識対象とする。

タイトル、注釈についてはそれぞれ該当するセルを認識する。見出しと数値についてはそれぞれ該当するセル範囲を認識する。なお、見出しについては、表の上側と左側に記述されていることが一般的であり、これら 2 か所の範囲で認識する。

一般的な統計表において、各認識対象に該当するセルおよびセル範囲を示した例を図 2 に示す。

各項目の認識にあたっては、以下に記述する順序で処理する。

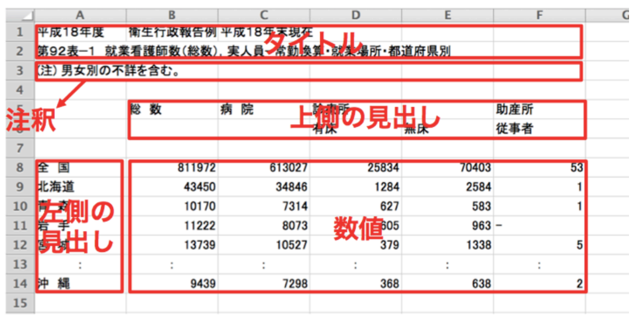

図 2　統計表の認識対象項目に該当するセル
およびセル範囲の例

3.4.2. 空の行および列の無視

表の中には、Excel 等の表計算ソフトで閲覧した際の見栄えを整えるために、1 行全て空のセル、また 1 列全て空のセルとなっている行や列が存在する。このような行や列はデータとしては特別な意味を持たないため、解釈をする際には空の行や列はないものとして処理を続ける。

3.4.3. 数値のセル範囲の認識

表における数値部分は、通常、最も右下の部分に配置され、特定しやすいと考えられるため、まず、数値のセル範囲を認識する。ただし、表の最も左下の部分には注釈が記述される場合があるため、該当箇所が注釈かどうかの判定を行った後、数値のセル範囲を認識する。以下、処理手順を説明する（図 3）。

1. 入力として与えられた表中の最も右下のセルを特定し、現在の注目セルとする。

2. 注目セルの左端のセルにおいて、3 文字目以内に「注」の文字が出現する場合、そのセルは注釈であると解釈し、上にある次の非空白行の右端へ移動する。
3. この段階での注目セルを、数値のセル範囲の**右下セル**であると決定する。次に、注目セル内の文字列が、数字または記号類(「・」「-」「...」「X」など)で記述されている場合、そのセルは数値であると判断し、数値ではないと判断されるまで、注目セルを左へ 1 セルずつ移動しながらこの判断を繰り返す。その後、上向きにも、1 セルずつ注目セルを移動しながら、そのセルが数値かどうかの判断をそうでないと判断されるまで繰り返す。この時点の直前で得られた注目セルを、数値のセル範囲の**左上セル**であると決定する。

	A	B	C	D	E	F	G
1	平成18年度	衛生行政報告例	平成18年末現在				
2	第92表—1 就業看護師数(総数), 実人員—常勤換算・就業場所・都道府県別						
3	(注) 男女別の不詳を含む。						
4							
5		総 数	病 院	診療所		助産所	
6				有床	無床	従事者	
7							
8	全 国	811972	613027	25834	70403	53	
9	北海道	43450	34846	1284	2584	1	
10	青 森	10170	7314	627	583	1	
11	岩 手	11222	8073	605	963	–	
12	宮 城	13739	10527	379	1338	5	
13	:		:	:	:		
14	沖 縄	9439	[illegible]	[illegible]	[illegible]	2	
15							
16	注：宮城県の石巻医療圏、気仙沼医療圏及び福島県を除いた数値である。						
17							

図 3 数値のセル範囲の認識手順

3.4.4. 見出し(左側・上側)の認識

一般的な表では、数値のセル範囲の左側に見出しが置かれる。また、左側の見出しでは、見出し文字列を異なる列のセルに記述することで、見出し内容の階層関係を表す場合がある(図 4)。よって、左側見出しは以下のように特定する。まず、数値のセル範囲の左隣の列について、数値のセル範囲の左側に該当するセル群を、現在の注目セル群とする。注目セル群に非空白文字が含まれるセルが 1 つ以上ある場合、その注目セル群は見出しであると判断し、見出しでないと判断されるまで、注目セル群を左へ 1 セルずつ移動しながらこの判断を繰り返す。

なお、上側の見出しについても、数値のセル範囲の一つ上の行について、数値のセル範囲の上側に該当するセル群を、現在の注目セル群として処理を開始し、判断と移動を同様に繰り返すことで、特定することができる。

3.4.5. タイトル・注釈の認識

上側の見出しのセル範囲よりも上にあるセルのうち、非空白文字が含まれる全てのセルを、タイトルまたは注釈であると判断する。ここで、条件に該当するセル内の文字列について、3 文字目以内に「注」の文字が出現する場合、そのセルは注釈であるとし、そうでないセルはタイトルであると判断する。

3.4.6. 見出し内容の階層関係の認識

3.4.4 節で述べたように、統計表の中には、見出しを異なる列のセルに記載するなどして、見出し内容の階層関係を表現する例が少なくない。ここでは、見出し内容の階層関係を認識する手順を説明する。

統計表における見出しの階層関係は、多くの場合、以下のいずれかのパターンで表現される。

- 異なる列または行のセルに見出し内容を記述し、段違いにして表現する
- 同じ列のセルについて、見出し内容の開始位置に異なる文字数の空白文字を埋めることで表現する
- 数値が何も記述されていない行を上位の見出し行として表現する

本稿では、上記の各パターンのどれに該当するかを判断し、そこで表現されている階層関係を、最上位の要素から順に子の要素をスラッシュでつなぎ、順番に記述する形式で表現することで、左側および上側の見出しをそれぞれ 1 列、1 行のみとした単純な形式の表に変換する。階層関係で記述されている見出しの例と、変換後の見出しの例を、それぞれ、図 4 と図 5 に示す。

		総数	歯科診療台	
			施設数	台数
2000年				
総数	総数	9026	1758	10867
	国	294	221	2719
	厚生労働省	22	21	43
	その他	272	200	2676
	公的医療機関	1362	489	1833
	都道府県	303	153	574
	市町村	757	238	865
2005年				
総数	総数	9362	1925	11533
	国	298	243	3015
:	:	:	:	:

図 4 階層関係で記述されている見出しの例

	総数/	歯科診療台/ 施設数/	歯科診療台/ 台数/
2000年/総数/総数/	9026	1758	10867
2000年/総数/国/	294	221	2719
2000年/総数/国/厚生労働省/	22	21	43
2000年/総数/国/その他/	272	200	2676
2000年/総数/公的医療機関/その他/	1362	489	1833
2000年/総数/公的医療機関/都道府県/	303	153	574
2000年/総数/公的医療機関/市町村/	757	238	865
2005年/総数/総数/	9362	1925	11533
2005年/総数/国/	298	243	3015
:	:	:	:

図 5 階層関係で記述されている見出しの単純な見出しへの変換例

3.4.7. タイトル、注釈、見出しからのキーワード抽出

後述するクエリとの関連づけにキーワードが必要となるため、タイトル、注釈、見出しの各テキストから、名詞、複合名詞、数値表現を抽出する。なお、名詞と複合名詞の抽出には形態素解析ツール JUMAN[9]を利用し、数値表現の抽出と正規化には normalizeNumexp[10]を利用した。なお、形容詞や動詞はノイズとなる可能性があるため抽出対象から除外した。

3.5. クエリ解析

本システムでは、ユーザは、客観的根拠を確認したい Web 上の文章をクエリとして入力する。入力されたクエリからは、3.4.7 節と同様にキーワードが抽出され、関連度の高い統計表データが探索される。さらに、最も関連度の高

い列あるいは行に絞り込みを行い、該当部分を動向情報の変化を表す図表画像に変換し、結果を提示する。

3.5.1. クエリからのキーワード抽出

ユーザから入力されたクエリからは、3.4.7 節と同一の方法で、名詞・複合名詞・数値表現がキーワードとして抽出される。

3.5.2. 統計表データのランキング

ベクトル空間モデルによるコサイン類似度を用いて統計表データをランキングし、最もコサイン類似度が高かった 1 件のみを後に提示する統計表データとして取り扱うものとする。クエリベクトル V_q と統計表データに対応する文書ベクトル V_d のコサイン類似度は以下の式で表される。

$$\cos(V_q, V_d) = \frac{V_q \cdot V_d}{|V_q||V_d|}$$

3.5.3. 列および行、範囲の絞り込み

関連度の高い表について、クエリの客観的根拠としてふさわしい列、行、適宜、その範囲を特定し、画像化する必要がある。今回はこの部分が未実装のため割愛する。

4. 評価実験と考察

4.1. 統計表の認識

4.1.1. 実験の目的と方法

本実験では、3.4 節で述べた CSV 形式の統計表データに対する表構造の認識手法の妥当性を明らかにすることを目的とする。

手動で収集した 252 件の CSV 形式の統計表データおよび、自動で収集した 49,961 件の CSV 形式の統計表データから無作為に選んだ 70 件の CSV 形式の統計表データに対し、3 章で述べた認識手法を用いて、表中の各項目を抽出し、正しく抽出できた割合（抽出率）により、評価・考察する。

なお、実験を行うにあたり、表の各項目について正解データの定義を表 1 に示す。

表 1 各抽出項目の正解データの定義

項目名	*正解データの定義*
タイトル	正しい見出し(上側)よりも上の非空白文字からなるセル群のうち、注釈以外のセル群
注釈	正しい見出し(上側)よりも上の非空白文字からなるセル群のうち、注釈と考えられるセル群、および、表の下に記述されている注釈と考えられるセル群
見出し(上側)	正しい数値のセル範囲の上に記述されている見出しを含むセル群
見出し(左側)	正しい数値のセル範囲の左に記述されている見出しを含むセル群
数値	数字または記号セルのみを含む最大ブロックを構成するセル群

4.1.2. 実験結果

提案手法による表項目の抽出結果を表 2 および表 3 に示す。

表 2 本手法による表の各項目の抽出率(252 件)

項目名	*抽出率*
タイトル	44% (113/252)
注釈	51% (129/252)
見出し(上側)	61% (155/252)
見出し(左側)	65% (165/252)
数値	91% (230/252)

表 3 本手法による表の各項目の抽出率(70 件)

項目名	*抽出率*
タイトル	49% (34/70)
注釈	59% (41/70)
見出し(上側)	53% (37/70)
見出し(左側)	87% (61/70)
数値	87% (61/70)

4.1.3. 考察

実験結果から、数値の抽出率は 90%を超え概ね良好であることが分かる。ただし、それ以外の項目についてはあまり芳しい結果とはいえず、手法を改良する必要がある。

数値の抽出における誤りの原因としては、以下が挙げられる。

- 1 つの CSV ファイル内に複数の表が存在した
- 数値を表現する際に、3 桁毎の区切り記号としてカンマ「,」を使用せず、半角スペースが使用されていた

また、見出し(上側)や注釈、タイトルの抽出における誤りの原因としては、以下が挙げられる。

- 見出し(上側)が記述されているセル群とタイトルや注釈が記述されているセル群との間に空白行がなく、タイトルや注釈のセル群まで見だし(上側)であると判定されていた
- 注釈の記述において「注」の文字が出現せず、「*」や「※」印などの記号で表現されていた

これらの要因に対処するためには、抽出手法のさらなる改良が必要である。また、ルールベースでの認識だけでなく、機械学習による認識手法を検討したいと考えている。

4.2. クエリと統計表との関連づけ

4.2.1. 実験の目的と方法

本実験では、3.5 節で述べたクエリの解析と統計表データとの関連づけ手法の妥当性を明らかにするのが目的である。本手法により、ユーザに提示すべきデータが含まれている適切な統計表データが選択されたかどうかを評価する。

実験用のクエリは 10 種類用意した。例を以下に示す。

- 京都市の出生率は 2001 年から減少しています。
- 大阪市の離婚件数は 2005 年から増加しています。
- 三重県の歯科関連における薬剤師の数は毎年減少しています。
- 乳児の死因のうちウイルス肝炎が占める割合は 2000 年から減少してきている。
- 東京都における医療保険の加入者数は 2000 年から減少してきている。

各クエリを用いて各統計表データをランキングし、上位10件についてMAP(Mean Average Precision)を求めた。なお、検索結果が正解かどうかの判定基準については、表4に示す3種類を用いることとした。

表4　検索結果の正解判定基準

基準名	*判定基準*
基準A	クエリで入力した内容の真偽が確認できる
基準B	クエリで入力した内容の真偽は確認できないが、クエリに深く関連した統計表データである
基準C	クエリで入力した内容の真偽は確認できないが、クエリにある程度関連した統計表データである

4.2.2.　実験結果

各判定基準で得られたMAP値を表5に示す。

表5　評価実験から得られた評価基準別のMAP値

基準名	*MAP 値*
基準A	0.156
基準B	0.453
基準C	0.718

4.2.3.　考察

実験の結果、判定基準が厳しくなるにつれてMAP値は大きく低下した。原因の1つとして、タイトルと注釈の記述内容に、その表に関する説明が十分記載されていなかったことが挙げられる。見出し部分にしか登場しないキーワードも多く存在すると考えられるため、今後は見出し部分も用いた手法を検討したいと考えている。

また、タイトルや注釈にその表に関する説明が記載されている場合でも、言い換えや表記ゆれ等により、クエリのキーワードと十分一致させられなかった例も多く存在すると考えられる。今後、表記の正規化を強化するなど改良が必要である。

また、クエリによっては、関連しそうな表が全く提示されない例も多く見られた。これは、今回用いた統計表データの数がそれほど多くなく、また、表認識の精度も十分ではないため、有用なデータ件数が不足しているためと考えられる。今後は、統計表データのさらなる収集と、認識精度の向上を目指す。特に、今回はCSV形式の統計表データのみを対象としたが、Excel形式のファイルに対しても表認識を適用することで、有用なデータを増強させられると期待できる。

5.　まとめと今後の課題

本稿では、ユーザがソーシャルネットワーク上で見つけた言説のうち、特に、動向情報をクエリとして入力すると、その客観的根拠となる統計表データ（あるいは関連するが矛盾する統計表データ）を探索し、ユーザに提示するシステムを提案し、試作システムを実装した。

評価実験の結果から、CSV形式で記述された統計表データの認識手法やクエリと統計表との関連づけ手法において、一部良好な結果が得られたものの、精度向上のために、より一層の改良を進める必要があることを確認した。

今後は、評価実験で得られた課題を克服するべく、統計表の認識手法やクエリとの関連づけ手法を改良していきたい。また、客観的根拠としてふさわしい統計表内の列、行、または、その範囲を同定し画像化する手法についても検討を行い、実装と評価実験により検証を進める予定である。

文　　献

[1] T.G. Kieninger and B. Strieder. T-Recs Table Recognition and Validation Approach. AAAI Fall Symposium on Using Layout for the Generation, Understanding and Retrieval of Documents, 1999.

[2] Yalin Wang, Ihsin T. Phillips, Robert M. Haralick, Table structure understanding and its performance evaluation, Pattern Recognition, Vol.37, 7, pp. 1479-1497, 2004.

[3] 塚本, 増田, 中川. "HTMLの表形式データの変換と携帯端末表示への応用", 情報処理学会研究報告 2002, 2002.

[4] S. Kerpedjiev, G. Carenini, S.F. Roth, and J.D. Moore. AutoBrief: a multimedia presentation system for assisting data analysis. Comput. Stand. Interfaces 18, 6-7, pp. 583-593, 1997.

[5] C.Wang, X. Xie,W.Wang,W.-Y. Ma. Improving web browsing on small devices based on table classification. Advances in Multimedia Information Processing – PCM, pp.88–95, 2004.

[6] 松下: "Elucignage : 探索的データ分析のための動向情報可視化インタフェース", 動向情報の要約と可視化に関するワークショップ第2回成果進捗報告会予稿集. pp.17-18, 2007.

[7] 佐伯, 宮森. "テレビ番組に基づくWebコンテンツの信頼性判断支援システムの提案", DEIM Forum 2012 B3-5, 2012.

[8] 日本政府, "e-Stat 政府統計", https://www.e-stat.go.jp/

[9] 京都大学 黒橋・河原研究室, "日本語形態素解析システム JUMAN", http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

[10] 成澤, "normalizeNumexp 数量表現・時間表現の規格化を行うツール", http://www.cl.ecei.tohoku.ac.jp/~katsuma/software/normalizeNumexp/