

# 人の価値観を表すカテゴリを対象にした自動分類

石田栄美\*, An-Shou Chen\*\*, Douglas W. Oard\*\*, Kenneth R. Fleischmann\*\*

駿河台大学文化情報学部\*      University of Maryland\*\*  
emi@surugadai.ac.jp\*

抄録：内容分析の自動化を試みるために、公聴会で発言された 28 の証言に対して、人の価値観を表すカテゴリを付与した新しいコーパスを作成した。用いたカテゴリは Schwartz Values Inventory である。コーパスには複数のカテゴリが付与されているデータも含まれているため、分類器のための学習用データの表現方法、閾値を用いた分類結果のカテゴリ数の推定方法を実験し、内容分析が自動的に実行される可能性があるか、検討した。

## 1. はじめに

内容分析は、発言や文書から人がどのような価値観をもっているか、世の中の価値観がどのように変化しているかなどを調べるための重要な研究手法の一つである<sup>1)</sup>。内容分析は、質的な作業を伴うため、そのほとんどが人手で行われているが、大変な労力や時間を必要とし、未だに大規模なテキストに対して適用している例は少ない。内容分析を自動的に行うことは、同じ概念を用いて繰り返しコーディングする場合や大量のテキストに対してコーディングを行いたい場合などには有効である。しかしながら、正確性に欠けることや新しい概念(カテゴリ)を用いたコーディングをすることはできないなどの問題がある。

人による分析もコンピュータを用いた分析も一長一短があるが、本研究の目的は、人とコンピュータの利点をうまく組み合わせたシステムを構築し、大規模なテキストに対して、質の高い内容分析を行えるようにすることである。内容分析が自動的に実行できるようになれば、ニュースやブログ、スピーチなどさまざまな文書に適用することもでき、また、より広い範囲を対象に価値観の変化を長期的に観察することも可能になる。本稿では、まず、公聴会における 28 の証言を含む陳述書に人の価値観 (human values) を表しているカテゴリをコーディングし、テストコレクションを作成した。つぎにそれを用いて、自動分類実験を行った。テストコレクションには複数のカテゴリが付与されているデータがあったため、自動分類実験では学習用データの学習手法といくつかのカテゴリを最終的な分類結果とするかを定める方法について、提案されている手法を検討した。

## 2. テストコレクション

### 2.1 コーパス

価値観とポリシーの相互作用に着目しているため、コーパスには、2006年2月7日に米国上院商務・科学・運輸委員会によっておこなわれた“ネットの中立性 (net neutrality)”に関する公聴会<sup>2)</sup>と

2008年4月17日連邦通信委員会が開催したブロードバンドネットワークに関する公聴会<sup>3)</sup>における証言を用いた。各委員会のウェブサイトには、これらの証言を文字起こししたものが提供されており、合計で28の証言を得た。公聴会における証言からは、ステークホルダーの価値観を反映した公序良俗に関するポリシーを分析することができる。

### 2.2 カテゴリとコーディング

カテゴリには社会科学の研究で広く用いられている Schwartz Values Inventory (以下、SVI とする) を用いた。SVI は、文化、言語、地理、宗教、人種などの分野で広く適用されており、その他の分野にも適用することが可能である<sup>4,5,6)</sup>。たとえば、心理学分野における行動と価値観との関係<sup>4,5)</sup>や価値観と所属政党の関係の調査<sup>5,6)</sup>などにも適用されている。SVI は階層構造になっており、4つの第一階層、10の第二階層、56の第三階層で構成されている。第三階層のカテゴリには、“Social Power”, “Successful”, “Equality”, “Politeness”, “Social Order”などがある。

28の証言に対し、2番目の著者がコーディングを行った。価値観を表している表現には、複数の文で一つの価値観を表している場合や文の一部分だけに価値観が表れている場合があるが、本研究では文単位でコーディングを行った。コーディングには ATLAS.ti を利用した。その後、ATLAS.ti から結果を xml 形式で出力し、さらに、分類器に入力するためのフォーマットに整形した。

コーパスには 2,294 の文が含まれており、そのうち 2,003 文に対し 3,160 の SVI カテゴリが付与されていた。文には、カテゴリが付与されていないものもあれば、複数カテゴリが付与されているものもあった。一つの文に対し最大で 7 カテゴリ付与されており、付与数の中央値は 1、平均は 1.58 であった。実際に付与されたカテゴリは、第一階層で 4 カテゴリ、第二階層で 10 カテゴリ、第三階層で 46 カテゴリであった。図 1 にコーパス全体で、20 回以上付与されているコードを示した。

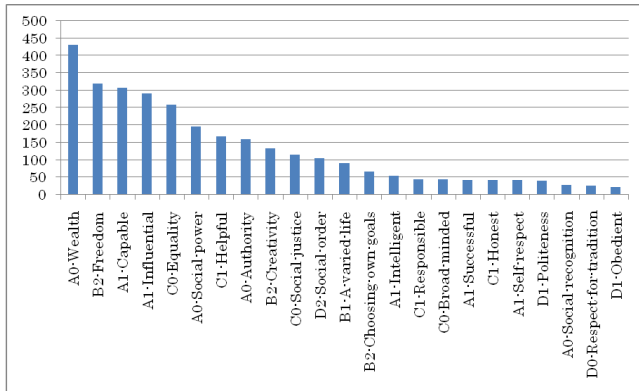


図1 20回以上出現するSVIカテゴリの分布

### 2.3 コーダー間の一致率

コーパスの中から4つの証言を選択し、第二コーダー(4番目の著者)がカテゴリを付与した。第一コーダーの結果を正解カテゴリとみなし、第二コーダーとの一致率をみたところ、第二コーダーの再現率は0.420、精度は0.359、F値は0.387であった。また、コーダー間の一致率を示す尺度であるCohen's Kappa値を求めた。コーパスの中で両コーダーが2回以上付与した17カテゴリを対象に、カテゴリごとのKappa値を求め平均したところ0.383と適正(fair agreement)であったが、決して一致率が高いとは言えない。対象とするものにもよるが、コーディングは人が行ったとしても同じ結果を得るのはむずかしい作業であるといえる。

自動分類実験においては、複数の人によるコーディングの結果を利用することが望ましいが、現段階では一人分のコーディング結果しか入手できないため、本実験では一人分のデータのみを使った。

## 3. 複数ラベルへの自動分類

テストコレクションには複数のカテゴリが付与されているものもある。本研究で用いた分類器は、カテゴリが付与されているデータを用いて学習することが必要であるが、データに複数のカテゴリが付与されている場合、学習用データをどのように表現すれば最も効果的であるかを検討しなければならない。また、分類器が出力した分類結果の候補の中から、最終的にいくつのカテゴリを分類結果として選択するかという問題もある。

以下では、この学習手法と分類手法の2つの問題に対し、提案されている手法の中でどの手法が適切であるかを、検討した。

### 3.1 学習用データの表現

テストコレクションの中で、複数ラベルが付与されている文を表1に示した。TsoumakasとKatakisからは、複数ラベルが付与されているデータの学習方法として5つの手法を提案している。

表1 複数カテゴリが付与されている文の例

カテゴリ	文
A1-Influential, B1-A-varied-life	As a result high speed access to the Internet is revolutionizing the way we work learn seek medical advice gather our news engage in public discourse interface with government socialize and almost every aspect of the way we live.
A0-Wealth, A1-Capable, B2-Freedom	To compete in the world we need a simple inexpensive and open network not a costly complex and balkanized one.
C1-Helpful, D2-Healthy	A Veterans Administration study showed you could cut hospital stays in half for many patients - and yet monitor and watch over them for longer periods of time.

#### 3.1.1 学習手法1: 複製手法

学習手法1は、複数カテゴリが付与されている場合でも、一つのカテゴリが付与されていた場合と同様に、各データを複製し、各々のカテゴリの学習用データとするという方法である。同じデータを複数のカテゴリに学習させてしまう可能性はあるが、分類性能が高い分類器を用いる場合は、効果的であるといえる。この手法は、すべてのコーディング結果を用いることができるため、Tsoumakasらはデータ量に制限がある場合には、有効であると述べている。表1で用いた例に対し、学習手法1を適用した結果を表2に示した。

表2 学習手法1を適用した学習用データの例

カテゴリ	文
A1-Influential	As a result high speed access to the...
B1-A-varied-life	As a result high speed access to the...
A0-Wealth	To compete in the world we need a...
A1-Capable	To compete in the world we need a...
B2-Freedom	To compete in the world we need a...
C1-Helpful	A Veterans Administration study...
D2-Healthy	A Veterans Administration study...

#### 3.1.2 学習手法2: カテゴリの出現回数による選択

学習手法1は、偏りがある学習用データの場合には、出現回数が多いカテゴリに対して過学習してしまう恐れがある。Tsoumakasらは、複数ラベルが付与されている場合、その中から適切な一つのカテゴリを選択し、そのカテゴリのみの学習用データとするという方法も提案している。本実験では、複数のカテゴリが付与されている場合、付与されているカテゴリの中でもっとも出現回数が少ないカテゴリを選択した。この手法を用いることで、データの偏りの問題が解消されると予想できる。表1の例に

学習手法 2 を適用した結果を表 3 に示した。この手法の場合は、学習用データが小さくなってしまいう問題がある。

表 3 学習手法 2 を適用した学習用データの例

カテゴリ	文
B1-A-varied-life	As a result high speed access to the...
A1-Capable	To compete in the world we need a...
C1-Helpful	A Veterans Administration study ...

### 3.1.3 その他の手法

その他に、複数カテゴリの組み合わせを一つのカテゴリとみなす方法、一つのカテゴリだけが付与されているデータだけを学習用データとして用いる方法、カテゴリごとにデータを正と負のサンプルに分けて学習する方法が提案されている。最初の 2 手法については、プレ実験で学習手法 1, 2 によりも性能が低かったため、本実験は行わなかった。3 番目の手法は、他の分類器を用いて、今後、実験する予定である。

### 3.2 分類手法

分類器として、性能が良いと報告されている  $k$  NN<sup>8,9)</sup>( $k=1,3,5,7,10,13,15,18,20$ )を用いた。本実験では、プレ実験の結果から、データに対し、ポーターのアルゴリズムを用いて語幹処理を行い、4 回以下しか出現しない語は削除した。 $k$  NN は Weka<sup>10)</sup> で提供されているものを用いた。Weka は、サンプルの近さをはかる尺度として 1) 同じ重み(vote)、2) 逆距離加重法( $w=1/\text{distance}$ , iw)、3) 類似性加重法( $w=1-\text{distance}$ , sw)の 3 つの重みづけ手法を提供している。また、Weka が提供している  $k$  NN の分類器は、評価用データに各々のカテゴリに対する確率分布を付与して、分類結果を出力することができる。本節では、この確率分布を用いて最終的なカテゴリを決定する方法を実験した。

#### 3.2.1 分類手法 1: 正解と同数のカテゴリ数を選択

実験環境においては、評価用データにいくつのカテゴリが付与されているか、あらかじめ分かっている。分類手法 1 では、この正解カテゴリとして付与されているカテゴリと同じ数のカテゴリを選択した。分類結果のカテゴリは以下のように選択した。あるデータ  $s$  に  $i$  個のカテゴリが付与されている場合、 $k$  NN の確率分布をもとに上位  $i$  番目までのカテゴリを、 $s$  に対する最終的な分類結果とした。 $i$  番目のカテゴリと同じ確率分布を持っているカテゴリがあった場合は、それらも選択した。

#### 3.2.2 分類手法 2: 閾値を用いる方法

実世界では、各データにいくつのカテゴリを付与することが適当であるかはわからない。そこで、閾値を用いて、選択するカテゴリ数を推定する方法を実験した。 $k$  NN の確率分布の値が閾値を以上であれば、最終的な分類結果のカテゴリとするという方

法である。この場合、適切な確率分布の値を設定することが必要である。

まず、テストコレクションを、学習用データ、閾値用データ、評価用データと 3 分割した。それぞれの割合は、80%、10%、10%である。次に以下の手順で閾値を設定した。

- 1) 学習用データを用いて分類器を学習する
- 2) 分類器に、閾値用データを入力し、各カテゴリの確率分布を得る
- 3) 閾値用のデータの正解と分類結果を比較し、閾値用データにおいて、もっとも F 値が高くなる閾値を選択する
- 4) 評価用データに対して 3) で得られた閾値を用いて、カテゴリを選択する

### 4. 結果

#### 4.1 評価尺度

評価尺度には、マクロ平均の精度、再現率、再現率と精度を組み合わせた F 値を用いた。また、以下で報告する F 値は、10 点交差検定を平均したものである。閾値を用いたカテゴリの選択の場合には、閾値用データと学習用データの間での交差検定は行っていない。

#### 4.2 学習手法の比較

表 4,5,6 にそれぞれ第三階層、第二階層、第一階層レベルでの F 値を示した。実験は、7 通りの  $k$  で実験したが、 $k$  が小さいうちは F 値もそれほど高くなかったため、表には  $k=13$  以上の結果を示した。また、この結果は分類手法 1 を用いた結果である。この結果から学習手法 1 のほうが、わずかではあるが、すべての階層で学習手法 2 よりもよい結果を示しているといえる。

表 4 第三階層レベルにおける F 値

	学習手法 1			学習手法 2		
	vote	iw	sw	vote	iw	Sw
13NN	0.3083	0.3197	0.3117	0.2951	0.2984	0.2988
15NN	0.3100	0.3184	0.3122	0.2984	0.2998	0.3010
18NN	0.3122	0.3206	0.3148	0.2997	0.3016	<b>0.3035</b>
20NN	0.3160	<b>0.3238</b>	0.3191	0.2991	0.2996	0.3008

表 5 第二階層レベルにおける F 値

	学習手法 1			学習手法 2		
	vote	iw	sw	vote	iw	sw
13NN	0.4734	0.4748	0.4757	0.4666	0.4633	0.4649
15NN	0.4704	0.4707	0.4719	<b>0.4672</b>	0.4639	0.4644
18NN	0.4716	0.4733	0.4734	0.4668	0.4637	0.4637
20NN	0.4728	<b>0.4758</b>	0.4755	0.4634	0.4607	0.4611

表 6 第一階層レベルにおける F 値

	学習手法 1			学習手法 2		
	vote	iw	sw	vote	iw	sw
13NN	0.6461	<b>0.6471</b>	0.6453	0.6414	0.6395	0.6398
15NN	0.6435	0.6455	0.6430	0.6418	0.6401	0.6399
18NN	0.6400	0.6422	0.6393	<b>0.6432</b>	0.6407	0.6410
20NN	0.6398	0.6418	0.6391	0.6400	0.6385	0.6380

#### 4.3 分類手法の比較：閾値を用いた結果

表 7,8,9 に第三階層、第二階層、第一階層レベルでの分類手法 1 と 2 の結果をそれぞれ示す。これは学習手法 1 を用いた場合の結果である。閾値は、確率分布の値 0.07 から 0.16 まで 0.01 刻みで設定した。もっとも高い F 値を示した閾値の結果を表に示している。第二階層、第三階層では、ほとんどのケースで 0.10、または 0.12 の時が最もよい結果であった。第一階層では、0.15 もしくは 0.16 であった。これらの結果から、最もよい結果を示したケースを比較すると、第一階層レベルでは 0.06 の開きがあるが、第二階層、第三階層レベルでは 0.02 の差におさまっている。第一階層レベルでは、閾値の設定に問題があったことが考えられる。これらの結果から、閾値を用いて正解カテゴリの数を推定したとしても、あらかじめ正解カテゴリ数を知っている場合と比べて遜色ない性能が得られることがわかった。

表 7 第三階層レベルにおける F 値

	分類手法 1			分類手法 2		
	vote	iw	sw	vote	iw	sw
13NN	0.3083	0.3197	0.3117	0.2846	0.2914	0.2842
15NN	0.3100	0.3184	0.3122	0.2862	0.2930	0.2869
18NN	0.3122	0.3206	0.3148	0.2880	0.2966	0.2883
20NN	0.3160	<b>0.3238</b>	0.3191	0.2938	<b>0.2998</b>	0.2945

表 8 第二階層レベルにおける F 値

	分類手法 1			分類手法 2		
	vote	iw	sw	vote	iw	Sw
13NN	0.4734	0.4748	0.4757	0.4442	0.4491	0.4438
15NN	0.4704	0.4707	0.4719	0.4504	0.4545	0.4491
18NN	0.4716	0.4733	0.4734	0.4519	0.4566	0.4523
20NN	0.4728	<b>0.4758</b>	0.4755	0.4519	<b>0.4607</b>	0.4543

表 9 第三階層レベルにおける F 値

	分類手法 1			分類手法 2		
	vote	iw	sw	vote	iw	Sw
13NN	0.6461	<b>0.6471</b>	0.6453	0.5811	<b>0.5838</b>	0.5813
15NN	0.6435	0.6455	0.6430	0.5784	0.5804	0.5786
18NN	0.6400	0.6422	0.6393	0.5760	0.5783	0.5762
20NN	0.6398	0.6418	0.6391	0.5736	0.5768	0.5746

#### 5. おわりに

人の価値観を表すカテゴリを伴った新しいテストコレクションを作成し、自動分類実験を行い、内容分析を自動的に行うことが可能かどうかを検討した。実験の結果、第三階層レベルでもっとも高い F 値は 0.3228 であり、閾値を用いて正解カテゴリ数を推定しても 0.2998 の F 値を得ることができた。4 証言を用いた場合のコーダー間の F 値は 0.387 であった。これらの結果から、現在のところ、人によるコーディングには及ばない性能ではあるが、自動的に内容分析を行える可能性はあるといえる。実験で検討した手法は学習手法と分類手法のみであったが、今後、カテゴリの出現位置や文脈を考慮した手法を取り入れ、性能を向上を試みる予定である。

#### 謝辞

本研究は科研費(19700232)の助成を受けたものである。

#### 引用文献

- Morris, R. (1994). Computerized content analysis in management research: A demonstration of advantages and limitations. *Journal of Management*, 20(4), 903-931.
- U.S. Senate. 2006, Feb 7. Senate Committee on Commerce, Science and Transportation Hearing on Network Neutrality.
- Fed. Comm. Commission. 2008, Apr 17. Broadband Network Management Practices Public Hearing. Palo Alto.
- Schwartz, S. (1992). Universals in the Content and Structure of Values. In M. Zanna, ed. *Advances in Experimental Social Psychology*, Academic Press, 25, 1-66.
- Schwartz, S.H. (1996). Value priorities and behavior: Applying a theory of integrated value systems. In C. Seligman, J.M. Olson, & M.P. Zanna (Eds.), *The psychology of values: The Ontario Symposium*, Vol. 8 (pp.1-24). Hillsdale, NJ: Erlbaum.
- Caprara, G. V., Schwartz, S. H., Cabaña, C., Vaccane, M., & Barbaranelli, C. (2005). Personality and politics: Values, traits, and political choice. *Political Psychology*, 27(1), 1-28
- G. Tsoumakas, I. Katakis, (2007). "Multi Label Classification: An Overview", *International Journal of Data Warehousing and Mining*, David Taniar (Ed.), Idea Group Publishing, 3(3), pp. 1-13
- A. Wiczkowska and P. Synak(2006) "Quality Assessment of k-NN Multi-Label Classification for Music Data" *ISMIS 2006*, 389-398
- Zhang, M. and Zhou, Z. (2005), "A k-Nearest Neighbor Based Algorithm for Multi-label Classification." *IEEE International Conference on Granular Computing*, Vol. 2, pp. 718-721
- WEKA, [<http://www.cs.waikato.ac.nz/ml/weka/>]