1E3-2in

Web アクセス履歴に基づくユーザの価値観の類推

Inferring consumer values by analyzing web access logs

木虎 直樹 Naoki Kitora
久保 征人 Masato Kubo

シナジーマーケティング株式会社
Synergy Marketing, Inc.

In order to increase corporate competitive power, it is important to understand consumer decision-making behavior regarding purchases by analyzing data produced by consumers. We assume that consumer action is based on personal values; therefore, we also assume that actions on the internet are based on personal values. We propose a method that can infer these values based on the analysis of web pages viewed by the consumer.

1. はじめに

インターネットおよび, PC, スマートフォン, タブレットなど情報端末の進化により消費者が生み出すデータの量は加速度的に増えている. この消費者が生み出す大量のデータを分析し, マーケティングに活かすことが求められているが, 現状はデモグラフィック情報や購買履歴などの行動履歴データを分析し, 表層的なマッチングを行うにとどまっている. このようなマーケティング手法は引き続き有効である場面も多いが, 一方で顧客をより深く理解し, よりよい需給マッチング, 幸せな消費活動を追求することが望まれている.

我々は消費者の購買行動における意思決定を左右する重要な因子の一つに価値観があると考え, この価値観成分により説明される社会的類型を中心に据えて, 消費者の行動を説明するモデルの研究をしている [馬場 2012] [馬場 2013] [谷田 2013].

人はその嗜好や価値観に基づいて行動するという仮説にたつと, 個人の Web 上での行動も嗜好・価値観に基づいていると考えられる. 本研究では消費者の行動履歴の一つである Web アクセス履歴を元に個人の閲覧したページを解析し, その価値観を類推することを試みる.

2. 全体像

2.1 Societas

我々は 2012 年から消費者の行動に関する価値観調査を始め, デモグラフィック情報, 価値観, 消費行動に関する 1,000 程度の質問項目からなる定量調査 (Web 調査) を行い, 約 1.1 万人のサンプルを得た. この調査結果を元に定義した社会的類型を Societas (ソシエタス) と名付け, 12 の Societas を定義している (表 1). 各 Societas には識別子があり, これを Societas 番号と呼んでいる.

表 1: Societas

Societas 番号	説明
#1-1	受け身な隠者タイプ
#1-2	受け身な清閑タイプ
#2-1	家族大好き悠々タイプ
#2-2	家庭的な真面目タイプ
#3-1	こだわりインドア派タイプ
#3-2	アウトロータイプ
#4-1	自分中心的なアクティブタイプ
#4-2	好奇心旺盛なバランス人間タイプ
#5-1	家族思いの多忙ワーカータイプ
#5-2	社交的な堅実ホームメーカータイプ
#6-1	繊細な個人主義タイプ
#6-2	好奇心旺盛な人生謳歌タイプ

Societas は価値観成分をクラスタリングして得たものであり, 得られた Societas 番号と価値観成分を教師データとして学習し, ベイジアンネットワーク [本村 2006] による確率モデル, 価値観モデルを構築している.

2.2 提案モデル

推論をするにあたり, ユーザの興味のあるカテゴリと持っている価値観の間には関連があるという仮説をたて, その仮説を以下のように分解した.

- ユーザが持っている価値観と閲覧する Web ページには関連がある
- Web ページには対応するカテゴリがある

図 1 に提案モデルの概要を示す.

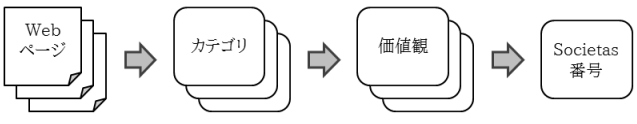

図 1: 提案モデルの概要

3. Web ページのカテゴリ推論

最初にポータルサイトなどの情報を元に 46 のカテゴリを定義した. 例えば以下のようなものである.

- 政治・経済
- 住宅・不動産
- 車・バイク

連絡先:木虎直樹, シナジーマーケティング株式会社 システム開発部 研究開発グループ, 電話番号: 06-4797-2900, メールアドレス: kitora.naoki@synergy101.jp

- 料理・レシピ
- …

Web ページのカテゴライズをするにあたっては以下の仮説をおいた.

- Web ページは 0 以上 3 以下のカテゴリに属す
- 推論により 4 以上のカテゴリに属すとされたページはどのカテゴリにも属さないものとする

このような仮説をおいたのは, 多くのカテゴリに属すと判定されたページは種々雑多な情報からなるポータルのような性質を持つものであり, このようなページを閲覧することに価値観はあまり影響せず, また興味のあるカテゴリだけでなく興味のないカテゴリも多く含まれると考えたためである.

カテゴリの推論は分類器を用いて行った. 使用した分類器はナイーブベイズ [Manning 2008] である. 正解データはカテゴリ毎に対応する Web ページを 50-60 ページ, 合計 2,486 ページ用意した. ページ内の単語をそのまま使うとノイズが多く, また高次元になるため以下の手順で単語を選別した.

1. HTML を形態素解析
2. tf-idf [Manning 2008] で単語を選別
3. カテゴリ毎に相関の高い単語を選別

学習に使用する正解データとは別に, 任意のアクセスログからランダムに 100 件抽出し, 人の目でカテゴライズしたものを利用してオープンテストを行い単語の次元数を調整した (表 2).

表 2: カテゴリ分類モデルの評価

単語次元数	precision	recall	F-measure
約 2,300	0.45	0.78	0.57
約 1,600	0.62	0.74	0.67
約 300	0.84	0.42	0.56
約 70	0.89	0.24	0.38

単語次元数が少ないほうが, カテゴリに特徴的な単語に絞られるため precision が高く, recall は低くなる傾向にある. 単語次元数が 1,600 程度となるようにしきい値を設定したときの F-measure [Manning 2008] が一番良く, 数値としては高いとは言えないものの, precision 0.62, recall 0.74 とランダム正解率よりも高いため, カテゴリの分類は可能であると考える.

4. カテゴリから社会的類型の推論

4.1 アンケート

カテゴリと価値観を結びつける情報を得るため, 価値観および Societas を推論するための設問と, 興味のあるカテゴリに関する設問からなる定量調査 (Web 調査) を行った. 使用したカテゴリは Web ページのカテゴリ推論で定義した 46 カテゴリである. 男女比を 1:1 とする以外の制約は設けずに 1,000 程度のサンプルを得た. そのサンプルから価値観成分の計算を行い, Societas 定義時に構築した価値観モデルによって Societas 番号を得た.

4.2 社会的類型推論モデル

アンケートで得られた価値観成分と Societas 番号にカテゴリ情報を結合して教師データとし, いくつかのパターンでカテゴリから Societas 番号を推論するベイジアンネットワークのモデルを構築した. 以下に評価したパターンのうち主要なものを挙げる. 全価値観成分は 62 個であるが, 今回取得した定量調査結果から直接計算できる価値観成分は 22 個であるため, それらだけで構成されるモデルも作成した.

1. 全価値観成分と Societas 番号から構成された既存のベイジアンネットワークにカテゴリを組み込んだモデル
2. 定量調査結果を元に, カテゴリ, 22 の価値観成分, Societas 番号から構築したモデル
3. 定量調査結果を元に, カテゴリ, 62 の価値観成分, Societas 番号から構築したモデル

1 は全ての価値観成分が, 全てのカテゴリそれぞれの親になり得るものとした (図 2). 2, 3 については Societas 番号と全ての価値観成分が、全てのカテゴリそれぞれの親になり得るものとし, 価値観成分間に親子関係はあるもの (2-1, 3-1) と, ないもの (2-2, 3-2) の 4 パターンでモデルを作成した. すべてのモデルにおいてカテゴリ間に親子関係はないものとした. 図 2 に 1 のモデルを, 図 3 に 2-1, 3-1 のモデルを図示する. 2-2, 3-2 は図 3 で価値観成分間に親子関係がないものであるため図は省略する. なお, 図中の価値観成分は 22 及び 62 の価値観成分の上位のものである.

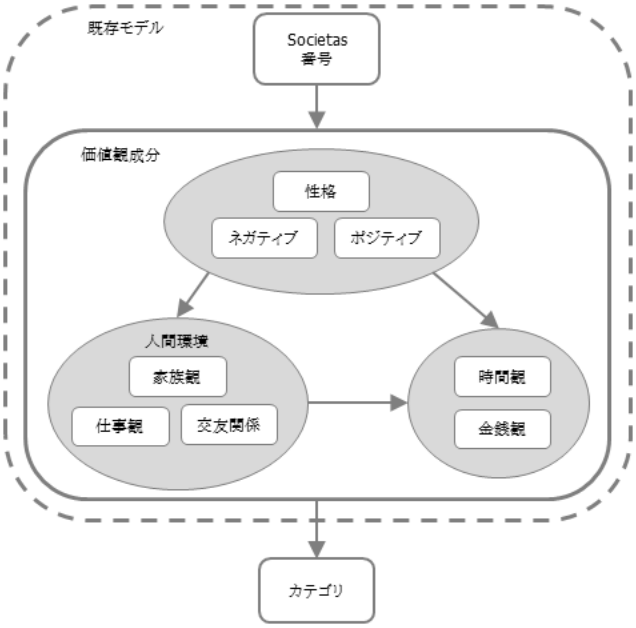

図 2: 社会的類型推論モデル 1

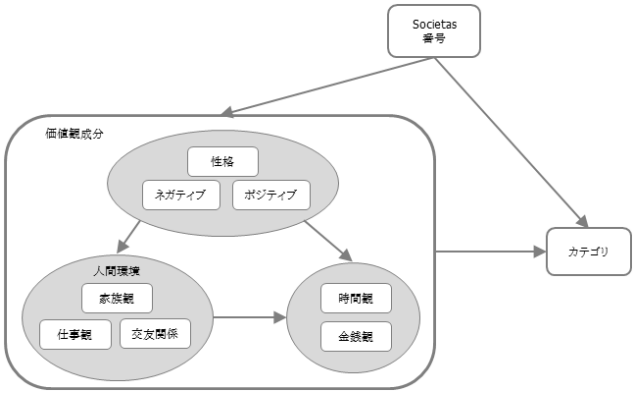

図 3: 社会的類型推論モデル 2-1, 3-1

4.3 評価

構築したモデルを使いクローズドテストによる評価を行った. ここで, 1 位正解率とは推論結果の 1 位が正解データの 1 位である確率であり, 2 位以上正解率とは推論結果の 1 位が正解データの 2 位以内に含まれる確率をいう.

表 3: 社会的類型推論モデルの評価 (クローズド)

パターン	1 位正解率	2 位以上正解率	3 位以上正解率
1	10.58%	21.55%	30.68%
2-1	52.91%	69.90%	77.38%
2-2	30.19%	44.37%	54.37%
3-1	23.50%	37.28%	47.57%
3-2	16.02%	27.77%	37.18%

次に 2-1 のモデルを使い, 正解データを 10 分割することによる交差検定でオープンテストを行った結果を表 4 に示す.

表 4: 社会的類型推論モデルの評価 (オープン)

1 位正解率	2 位以上正解率	3 位以上正解率
18.74%	33.50%	45.53%

この結果から以下のことが考えられる.

- クローズドテストとオープンテストとで, 特に 1 位正解率の乖離が大きく, クローズドテストにおいてオーバーフィッティングを起こしている可能性がある.
- 価値観成分間に親子関係をもたせている 2-1, 3-1 の正解率のほうが高いのは, 正解データを生成したモデルにおいて価値観成分間に親子関係をもたせており, モデルのネットワーク構造が親子関係を持たせない場合に比べて近づくためだと考えられる.
- ランダム正解率は 1 位で 8.3%, 2 位以上で 16.7%, 3 位以上で 25%であるため, オープンテストにおいても 2 倍程度は良い結果を得られている. カテゴリによる推論は有効だと考える.

5. まとめ

精度に関する課題は多くあるものの, Web ページからカテゴリ, カテゴリから価値観成分を推論すること, すなわち Web アクセス履歴から価値観および価値観から構成される Societas を推論することは十分に可能であることが確認できた.

カテゴリの分類精度はランダムよりも良いというだけでまだまだ向上の必要がある. 社会的類型推論モデルについてもより精度を高めるためサンプルを増やして構築したいと考えている. また, 実証実験の実施とその結果検証を行い本提案手法の有効性を確認したい.

さらにその先では Web アクセス履歴以外の行動履歴も価値観推論の証拠データとできるようにすることで, 複数の行動履歴データから証拠データを生成できるようにする. 証拠となるデータを増やすことで, 推論可能対象を増やすこと, および推論精度を高めることを目指す.

謝辞

本研究を進めるにあたり数々の助言を頂いた谷田泰郎氏, Mathieu Bertin 氏に心より感謝いたします.

参考文献

[馬場 2012] 馬場彩子, 木虎直樹, 谷田泰郎, 後迫彰, 井上哲浩, 加藤卓: 社会知を還元するクラウド型データベース「iNSIGHTBOX」の構築, 情報処理学会関西支部大会, 2012.

[馬場 2013] 馬場彩子, 谷田泰郎, Mathieu Bertin: 社会知としての消費者価値観構造モデルと類型「Societas」の構築, 人工知能学会全国大会(第 27 回)JSAI2013, 2013.

[谷田 2013] 谷田泰郎, 河本裕輔, 馬場彩子: マイクロブログにおける潜在的価値観の推定, 人工知能学会全国大会(第 27 回)JSAI2013, 2013.

[本村 2006] 本村陽一: ベイジアンネットワーク技術, 東京電機大学出版局, 2006.

[Manning 2008] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze: Introduction to Information Retrieval, Cambridge University Press, 2008.