1D2-5in2

構造化データとテキストデータを組合せた消費者の声分析

Voice of Consumer Analysis by Combining Structured Data and Text Data

顔玉蘭　金英子　市川秀樹　室住淳一
Yulan Yan　Yingzi Jin　Hideki Ichikawa　Junichi Murozumi

アビームコンサルティング株式会社 BI セクター
ABeam Consulting Ltd. Business Intelligence Sector

Organizations rely on data analysts to improve their business processes and make better business decisions. Though numerous analysis and visualization tools have been built on structured enterprise data, there has been little research on how to utilize text information which is unstructured data. In this paper, to identify root causes of low transfer intention in the case of liberalization of retail electricity sales, we extracted features from free-text answers and combined them with features from selective answers of questionnaire data. As a result, we identified characteristics of customers who are more likely than others to not transfer to other electricity suppliers.

1. はじめに

近年、企業のデータ分析による様々な取組が報じられている。ただし現状、ビジネスに活躍されているほとんどのデータは構造化されたデータであり、非構造化データであるテキスト情報の活用はまだ少ない。本手法では、家庭向け【電力自由化】に関する 調査結果[マーシュ調べ]において、探索型多変量手法による選択問題の構造化データ分析と、テキスト分析による自由回答文の分析を組み合わせることで消費者の声を分析し、自宅の電力会社を変える意向が低い消費者の特徴を明らかにした。

2. 解析手法

本研究では、家庭向け【電力自由化】に関する調査結果を分析することで、家庭向け「電力自由化」が適用される時に自宅の電力会社を変える意向が低い消費者の特徴を明らかにする。調査項目には、選択型項目と自由記述型項目があり、前者は構造化されたデータで、後者は自由記述文（非構造化データ）である。本研究では、まず探索型多変量手法の「HyperCube®[1]」を用いて構造化されたデータを分析する。次に、テキスト分析により非構造化データである自由記述文を分析する。さらに、自由記述文からキーワードクラスターという新たな説明変数を作成し、構造化データとして HyperCube®に導入することで、消費者の声情報を利用して電力会社の乗り換え意向が低い消費者の特徴を一層明らかにする。

2.1 探索型多変量手法で要因分析

HyperCube®は、データの多次元空間を総当りかつ網羅的に探索することで、あらゆる現象の発生条件を特定可能にする探索型多次元分析手法である。従来の統計手法とは異なるアプローチにより、局所的な傾向も漏らさない探索を可能にする。HyperCube®は企業における多様なデータの中からビジネス課題の根本原因を発見する解析ツールとして、欧米をはじめ日本でも多数の実績を有する。

連絡先：顔玉蘭，株)アビームコンサルティング BI セクター，yulyan@abeam.com，本研究の内容はアビームコンサルティング株式会社の公式見解を示すものではありません．

[1] http://www.institut-hypercube.org/en_index.html

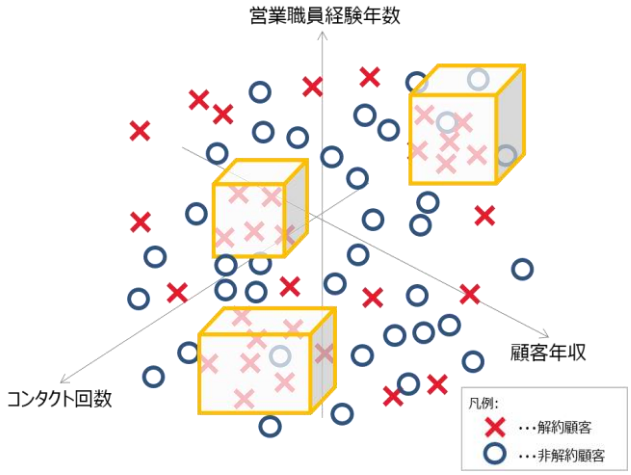

図1：HyperCube の原理

(1) HyperCube®のアルゴリズム

HyperCube®ではまず、データを分析対象項目により生成された次元中にプロットする。解約傾向の高い顧客の特徴分析を例（図1）とすると、分析対象とする顧客を、「顧客年収」、「営業職員経験年数」、「コンタクト関数」の説明変数項目を軸とする3次元空間にプロットする。データをプロットする際に、解約顧客は「×」、非解約顧客は「○」とする。そして、プロットしたすべての現象（○と×）を、多次元空間の中で総当りかつ網羅的に探索し、あらかじめ設定した閾値を上回る濃度をもつ領域を抽出する。解約顧客の特徴を分析するために、「×」の濃度を以下の式により計算する。

$$濃度＝×の数/(○の数＋×の数)$$

図1で例えば、解約率80％の濃い領域を閾値とした場合、×を6つ○を3つ含む右上キューブの領域は濃度が 66.7%で廃却され、左側真ん中と下のキューブはそれぞれ濃度 100%と88.9%で採択される。その結果、分析目的への寄与率や適応させた場合の確率を含め、分析項目と値を組み合わせた形となっ

て、ルール形式にて分析の解が複数導出される。例えば、「顧客の年齢が 20~25 かつ 営業職員経験年数が 2 年未満 かつ コンタクト回数が 3 回未満の顧客は、解約に至る確率が 88.9%」のようなルールが出力される。

HyperCube®は、2 次元/3 次元で人間が目視でデータの集まりを見つけるのと同じことを多次元(説明変数が多数)の場合にも適用することができる。そして、全ての「濃い領域」を網羅的に提示してくれる。多次元で対象データが集まっている領域(hyper-cube)を各変数の条件の範囲(ルール)として抽出する。

(2) HyperCube®の特徴

HyperCube®は従来の統計手法と比べいくかの特徴がある。まず、従来の統計解析手法では、仮説を設定し、仮説立証に有効と想定されるデータ項目のみ分析し、有効な示唆が得られるまで繰り返す。HyperCube®では全データ項目の組合せを総当りで探索するため、試行錯誤の繰り返しが不要である。つぎに、従来の統計解析手法では、データを 1 つの塊として分析するため、全体傾向の把握が可能であるが、HyperCube®は、データを個々にあらゆる条件別に集計し、通常と比較して優位差のある条件の組み合わせを発見するため、全体傾向に加えて局所傾向も細くすることが可能である。さらに、従来の統計解析手法では、欠損値・はずれ値がある場合、補填や置換等のデータ修正が必要となり、修正量が多い場合はそのデータの利用をあきらめていましたが、HyperCube®では欠損値・はずれ値は結果に影響を与えないため、データ修正をすることなくデータ分析が可能である。

本分析では、これらの HyperCube®の特徴を利用し、消費者調査における回答結果から、どのような属性の回答者が、どのような質問に、どのような回答をすると、特定の傾向をつながるかの特徴を分析する。選択型質問として、例えば「*SC2. あなたは「電力の自由化」について、どの程度ご存知ですか。(1つ選択)*」の質問に対し、「内容をよく知っている」「内容をなんとなく知っている」「名前だけ知っている」「聞いたことがあるような気がする」「知らない・聞いたことがない」の回答から答えを選択する。調査結果に対し、本分析では質問項目を説明変数とし、個々の回答者に対し質問項目に該当する回答結果を値とする。HyperCube®はこれらの分析データに対し、仮説なしでスタートし、データの全次元空間を総当りかつ網羅的にルール探索を行うことで、従来の統計手法では見えなかった局所的な傾向も漏れなく把握する。

2.2 消費者の声のテキスト分析

HyperCube®では構造化されたデータのみが分析可能である。電力自由化調査には自由記述型項目もある。例えば、「*Q6. 家庭向けの「電力自由化」になることであなたが新しい電力会社に期待することはどのようなことですか。*」の質問に対し、自由記述文によりコメントすることを求めている。本研究では、これらの回答文に対し以下のテキスト分析を行った。

1. **キーワードの抽出**:回答文のテキストに対し、まず形態素解析を行いシステム、名詞、動詞、形容詞に限定し上位頻出語を抽出した。例えば、「料金」、「安定」、「自由」など上位キーワードが抽出される。
2. **キーワードクラスターの作成:**自由記述文は回答者により表現のゆれが生じる。例えば、料金を表す意味にも、「料金」「価格」「電気代」など様々な語が使用されている。今回は手動でシソーラス辞書を作成し、類似意味を表す語を同じクラスターとしてまとめた。各クラスターで一番頻出する語を用いて『安定』『料金』『対応』などのようにクラスターを命名した。
3. **キーワードクラスターの共起ネットワーク**:各キーワードクラスターに属する語が他のクラスターに属する語と同じ回答文に出現した場合、2つのクラスターの共起回数を1としてカウントすることで、全回答文におけるキーワードクラスターの共起ネットワークを作成した。キーワードクラスターを共起ネットワークで視覚化することで、回答者のコメントの傾向及び中心課題を把握することができた。自由記述項目には、家庭向け電力自由化について「期待すること」と「不安に思うこと」をそれぞれ質問してあり、本分析ではその回答結果を「期待」と「不安」に分類する。例えば、質問「*Q7. 家庭向けの「電力自由化」になることであなたが不安に思うことはどのようなことですか。*」の回答文は「不安」と分類する。回答文を「期待」「不安」に分類してそれぞれ分析することで、消費者の期待や不安に至った傾向を比較することができる。

2.3 テキストの構造化で要因を深掘り

本研究では、自由記述型項目の回答内容から抽出したキーワードクラスターを、構造化データの説明変数として分析に使用する。1回答者を1レコードとし各キーワードクラスターを説明変数として分析データを作成する時、当該レコードによる説明変数(クラスターC_i)の値は、次のようにスコアリングした。

$f_1(kw)$ = キーワード(kw)が期待コメントに出現頻度
$f_2(kw)$ = キーワード(kw)が不安コメントに出現頻度

$$Score(C_i) = \frac{\sum_{\mathrm{kw}\in C_i} f_1(kw) - \sum_{\mathrm{kw}\in C_i} f_2(kw)}{\sum_{\mathrm{kw}\in C_i} f_1(kw) + \sum_{\mathrm{kw}\in C_i} f_2(kw)}$$

まず、当該キーワードクラスターC_i(説明変数)に属する語 *kw* が当該回答(レコード)での出現頻度を、期待コメントと不安コメント別にそれぞれ数える。$f_1(kw)$を期待コメントでの出現回数、$f_2(kw)$を不安コメントでの出現回数にし、その出現頻度の差を計算する。そして出現頻度の差を、*kw* が期待・不安の両方で出現する回数により割り算することで、値の範囲を[-1,1]間に正規化する。この値を、説明変数 C_i の当該レコードにおける値とする。例えば、キーワードクラスター『安定』を説明変数C_Xとした場合、回答者 *Y* の回答コメントの中には C_X クラスターに属する語が、期待コメントには 9 回出現し、不安コメントには 1 回出現したとする。この場合、回答者 *Y* における説明変数 C_X の値は、(9-1)/(9+1)=0.8と、正の1に近い数字になり、回答者 *Y* は自由電力の『安定』した供給について期待していると解釈できる。

自由記述型項目の回答内容から抽出したキーワードクラスターを説明変数として HyperCube®のルール探索に使用することで、選択型項目のみを使用した結果より説明力があるルールを探索することが可能になる。

3. データ概要

本研究では、家庭向け【電力自由化】に関するインターネット調査結果を分析データとする。詳しくは、2016 年 4 月から始まる予定の家庭向け【電力自由化】について、全国 30～69 歳の電力が自由化することを認識している男女 477 名を対象に、株式会社マーシュが実施したインターネット調査結果[マーシュ調べ]である。調査内容には、対象者の性別、年齢、年収、住まいなど対象者の属性に関する選択項目や、「電力自由化」について

知っている内容、情報収集方法、興味度、重視点などに関する選択項目、ほかに「電力自由化」に期待すること、不安に思っていることなどをコメントできる自由記述項目もある。回答形式には、単一回答(SA)、複数回答(MA)、自由回答(FA)がある。

本研究では、自宅の電力会社を変える意向が低い回答者の特徴を分析する。目的変数は以下のように定義しる。調査質問の「*Q4. あなたは、家庭向けの「電力自由化」が適用されたら、自宅の電力会社を変えたいと思いますか。(1つ選択)*」の質問項目に対し、「変えたい、やや変えたい、どちらともいえない、あまり変えたくない、変えたくない」の5つの回答内容から、「あまり変えたくない」「変えたくない」のいずれを選択した回答者を、自宅の電力会社を変える意向が低い回答者(以下「低意向者」)とする。「低意向者」は全回答者の 10.5%(50 人)を占めてあり、本研究では、これら「低意向者」の回答の特徴を分析することを目的とする。説明変数は各調査項目とし、各回答者の回答結果を説明変数の値とする。なお、複数回答形式の項目については、回答の数ほど説明変数を横展開で作り、2 値の説明変数として分析に導入する。自由回答文については、2.2~2.3 節で提案したキーワードクラスターを作成して説明変数として用いる。従って、本研究の説明変数に使用する構造化データは、選択型項目と回答文テキストから作成したキーワードクラスターを用い、非構造化データは回答文テキストそのままを用いる。

表 1: 分析対象データ

対象	「電力自由化に関する調査」	
対象者	447 人	
目的変数	乗り換え意向が低い回答者(「低意向者」): × その他の回答者:○	
説明変数	構造化項目	選択型項目:単一回答、複数回答から作成した説明変数 84 項目
		自由記述型:回答文から抽出したキーワードクラスター説明変数 13 項目
	非構造化項目	自由記述型:回答文テキスト

4. 分析結果

本分析では、まず HyperCube®を使用し構造化データ分析を実施した。選択型項目の構造化データを分析データとし、乗り換え意向につなぐ要因を探索する。次に、非構造化データである自由記述の回答文に対しテキスト分析を行う。自由記述文からキーワードを抽出し、キーワードクラスターを作成するとともに、キーワードクラスターの共起ネットワークを視覚化することで、期待・不安コメントにおける傾向の特徴を示す。さらに、テキスト分析から作成したキーワードクラスターを説明変数と定義し、他の構造化データとともに HyperCube®を使用したルール探索を行う。テキスト情報から作成した説明変数を追加することで、より表現力の高いルールが得られた。各ステップにおける詳細結果を以下で説明する。

(1) 構造化データ分析の結果

家庭向け【電力自由化】に関する調査項目から作成した説明変数(84 項目)の構造化データに対し、HyperCube®探索(2.1 節)を実行することで、乗り換え意向が低い回答(以下、「低意向者」とする)の特徴を抽出した。 全説明変数の多次元空間上 447 件の回答結果から、「低意向者」の密度が通常の2倍を上回る HyperCube ルールを出力した。その結果、最大3変数の組合せによるルールが約 6,673 個抽出され、例えば、以下のようなルールが抽出された。Rule1 では、*選択項目「F2. あなたの年齢」の回答結果が「60～69 歳」で、かつ、選択項目「Q5. あなたがご自宅の電力会社を選ぶ際に最も重視すること」の回答結果が「ライフスタイルにあった料金プラン」或は「安定した電力供給」*の回答者が 31 人いて、彼らの総合評価は平均より 3.1 倍「低意向者」になりやすいことを示す。

表 2:HyperCube ルールの例

Rule 1: 本ルールの条件は以下の通り	
「F2. あなたの年齢」	「60～69 歳」
「Q5. あなたがご自宅の電力会社を選ぶ際に最も重視すること」	「ライフスタイルにあった料金プラン」或は「安定した電力供給」
上記条件に当てはまる回答者は 31 人がいる、平均より 3.1 倍「低意向者」になりやすい	
Rule 2: 本ルールの条件は以下の通り	
「Q1. あなたは、「電力自由化」について情報収集をしたもの。」	「インターネットのニュース(yahoo!ニュースなど)」
「Q5. あなたがご自宅の電力会社を選ぶ際に重視することをお知らせください。」	「契約期間の縛りがないこと」を特に重視しない
上記条件に当てはまる回答者は 88 人がいる。平均より 2.1 倍「低意向者」になりやすい	

上記のようなルールが約6千個以上抽出され、「低意向者」につながる局所的な傾向を網羅的に抽出することができた。

(2) テキスト(非構造化データ)分析の結果

自由記述型項目の回答文のテキストに対しは、2.3 節のテキスト分析を実施した。まず、KH Coder [樋口 2014]を使用して、形態素解析を実施し名詞、動詞、形容詞に限定しキーワード抽出した。「期待」コメントと「不安」コメントの上位 15 語は以下のようである。

表 3:自由回答文のキーワード(上位 15 個)

Rank	「期待」コメント	「不安」コメント
1	安い	安定
2	料金	供給
3	安定	安定供給
4	価格	電力
5	供給	停電
6	安定供給	電力供給
7	電力	不安定
8	電気料金	料金
9	電力供給	会社
10	電気代	不安
11	電力会社	電力会社
12	サービス	企業
13	プラン	高い
14	安全	対応
15	原発	倒産

次に、回答コメントの表記ゆれを解消するため、類似意味を表す語を同じクラスターとして手動でまとめることで、シソーラス辞書を作成した。表4のような 13 個のキーワードクラスターを作成した。

表 4:キーワードクラスターの例

No.	クラスター	クラスターに属する単語
1	電力会社	電力会社, 会社, 企業, 購入先, …
2	電力	電力, 電気, 発電
3	安定	安定, 停電, 不安定, 不安, 信頼, …
4	供給	供給, 電力供給, 提供、…

5	選べる	選べる, 自由, 選択, 自分, …
6	料金	料金, 価格, 電気料金, 電気代, …
7	契約	契約, プラン, 料金プラン, …
8	競争	競争, 価格競争, …
9	買える	買える, 買う, 購入, 販売, 売電, …
10	対応	対応, サービス, …
11	使用	使用, 利用, 使う
12	災害	災害
13	原発	原発

最後に、キーワードクラスターの共起ネットワークを作成し、共起が強い 60 個の共起関係をネットワーク上で視覚化した(図 2)。図 2 に示したように、消費者家庭向け電力自由化に対し不安と思うコメントには、『対応』と『解決』という 2 つの話題を中心とし、『業者』の『リスク』や『詐欺』、『電力』の『安定』『供給』に関するコメントに多く述べられていることが分かる。

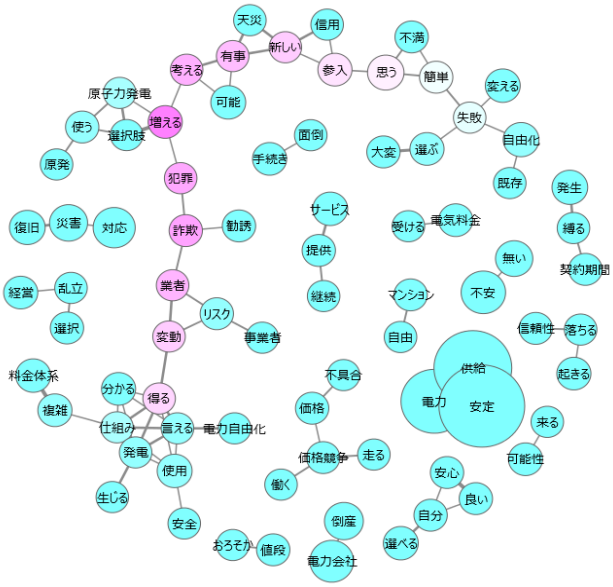

図 2: ネットワーク分析図(消費者の不安なコメント)

(3) 構造化データとテキスト情報を組合せた分析の結果

テキスト分析で作成した 13 個のキーワードクラスターを説明変数とし、2.3 節の数式により回答者ごとに各説明変数のスコアを計算した。これらの説明変数を(1)の分析で使用した構造化データに加え、再度 HyperCube®ルール探索を実行した。同様のパラメータで「低意向者」回答の密度が通常の 2 倍を上回る HyperCube ルールを出力した結果、最大 3 変数の組合せによるルールが約 5,544 個抽出された。例えば Rule3 では、*選択項目「F2. あなたの年齢」の回答結果が「60～69 歳」で、かつ、選択項目「Q5. あなたがご自宅の電力会社を選ぶ際に最も重視すること」の回答結果が「ライフスタイルにあった料金プラン」或は「安定した電力供給」で、かつ、テキスト情報から作成した説明変数『安定』キーワードクラスターの値が「-1～1」*の回答者が 27 人いて、彼らの総合評価は平均より 3.5 倍「低意向者」につながりやすいことを示す。

選択肢項目のみを使用した構造化分析結果(1)に、テキストからの抽出したキーワードクラスターを説明変数として構造化分析に加えることで、乗り換え意向が低い回答につなぐ要因を更に深堀することができた。例えば、Rule3(表 5)は Rule1(表 2)の 1つの条件である「*Q5. あなたがご自宅の電力会社を選ぶ際に最も重視することを知らせてください。*」の回答結果が「ライフスタイルにあった料金プラン」或は「安定した電力供給」という要因を、Rule3 では、回答文のテキスト情報から作成した説明変数(ここでは、『安定』キーワードクラスタ)を利用することで、自由電力の安定した供給に対し不安も期待もある消費者([-1,1])に対し、その不安を解消する施策を施すことにより、消費者の乗り換え意向を向上させる可能性を示唆する。ルールの条件に更にテキスト情報からの条件を加えることで、ルールがカバーするサイズは小さくなっているが、寄与率がより高い局所的なルールを抽出することができた。

表 5:HyperCube ルールの例

Rule3: 本ルールの条件は以下の通り	
「F2. あなたの年齢」	「60～69 歳」
「Q5. あなたがご自宅の電力会社を選ぶ際に最も重視すること」	「ライフスタイルにあった料金プラン」或は「安定した電力供給」
『安定』キーワードクラスター	[-1,1](不安或は期待している)
上記条件に当てはまる回答者は 27 人がいる、平均より 3.5 倍「低意向者」になりやすい	
Rule4: 本ルールの条件は以下の通り	
「Q1. あなたは、「電力自由化」について情報収集をしたもの。」	「インターネットのニュース(yahoo!ニュースなど)」
「Q5. あなたがご自宅の電力会社を選ぶ際に重視することをお知らせください。」	「契約期間の縛りがないこと」を特に重視しない
『供給』キーワードクラスター	[-1,1](不安或は期待している)
上記条件に当てはまる回答者は 39 人がいる、平均より 2.7 倍「低意向者」になりやすい	

5. 結論と今後の展望

本研究では、構造化データと非構造化データを組合せて両方を分析に使用する手法を提案した。構造化データ分析については HyperCube®ツールを使用し、従来の統計分析ツールではカバーできない局所的な要因を漏れなく探索することができた。構造化データ分析についてはテキスト分析を実施し、自由記述回答文の傾向を明らかにした。さらに、非構造化データ(テキスト情報)から構造化分析のための説明変数を作成して構造化データ分析に加えることで、家庭向けの「電力自由化」が適用されたら、自宅の電力会社を変える意向が低い消費者の特徴を明らかにした。

今回はテキスト情報からより厳密な説明変数を作成するために、キーワードクラスターを手作業で作成した。今後より大量のデータ分析においては、自動化作業が必要になり、高精度なクラスタリングモデル構築や[橋本 2008]テキスト評判分析モデル[鍛治 2009]などを構築する必要がある。

参考文献

[樋口 2014] 樋口耕一 2014『社会調査のための計量テキスト分析 ―内容分析の継承と発展を目指して―』ナカニシヤ出版

[鍛治 2009] 鍛治伸裕, テキストから評判分析と機械学習, 人工知能学会, 第 73 回 人工知能基本問題研究会, 2009.

[橋本 2008] 橋本泰一,村上浩司,乾孝司,内海和夫,石川正道, 文書クラスタリングによるトピック抽出および課題発見, 社会技術研究論文集, 2008

[マーシュ調べ] 株式会社マーシュが実施した家庭向け電力自由化に関するインターネット(自主)調査:https://www.marsh-research.co.jp/examine/2709denryoku.html