書籍のレビューに基づく先見性のある人物の特徴分析

掛谷英紀　　**佐藤裕也**
kake@iit.tsukuba.ac.jp　　s1420791@u.tsukuba.ac.jp
筑波大学

概要　先見性のある人物と先見性のない人物の特徴を見出す方法として、Amazon のカスタマーレビューを使用した機械学習を提案する。本研究では、レビューの評価の平均が大きく変化した本について、後に優勢となる評価を初期にしていた人物と先見性のある人物と定義する。この定義に基づき先見性のある人物の特徴を分析したところ、洋書を含む本のレビューの割合が多いこと、本のジャンルでは文学・評論作品のレビューが多いことが分かった。また、機械学習の結果、先見性のない人物はマスメディアに流されるレビューをする傾向が強いことが分かった。

1　はじめに

現代は、将来を占うのが非常に難しい時代となっている。政治の世界では、55 年体制が崩壊し、政局は不安定となっている。その結果、マスコミが称賛した政党の失政や政策の失敗が繰り返されており、有権者はどの政党に投票するか判断しにくい状況が生じている。また、学生の就職活動においても、不安定な経済状況が続いており、どの企業に就職するかの判断がつかないこともしばしばある。このような時代において、政治家や経営者の先見性を見抜く手がかりとなる情報は極めて価値の高いものである。

政治に関して有権者への判断材料を提供する手段として、東らはツイッターを用いて国会議員を分類する手法[1,2]を提案している。これは、国会議員のツイッター上の発言を用いて議員間の類似度を自己組織化マップ上に可視化し、投票先の判断の手掛かりを提供することを意図したものである。橋本らは政治家に対する Web 上のレビュー記事を用いて政党のイデオロギー別に文書分類する手法[3]を提案している。一方、ネットへの書き込みではなく、国会議員自身の国会での発言に着目した研究もある。畑中らは国会会議録を用いて国会議員を分類する手法[4]を提案している。また、尾崎らは国会会議録に基づく政党類似度マップの作成や、国会議員の国会内での発言を要約する手法[5]を提案している。しかし、これらはあくまでも政治的主張の分類や類似度情報を提供するものであって、どの主張に先見性があるかについての判断材料は全く提供されない。

一方、企業についての判断材料を提供する研究としては、橋本らによる経営トップのメッセージを分析がある[6]。同研究では、企業の経営トップのメッセージを株価変動率ごとに分類し、各カテゴリのメッセージ群の特徴分析をしている。また、企業風土の特徴分析を行う研究として、佐藤らは自然言語処理を用いた広報文書に基づく企業風土の特徴分析を行っている[7]。しかしながら、以上の先行研究は、いずれも企業の安定性のみに注目している面があり、企業経営者の先見性を見抜くことは意図していない。

先見性に着目した研究として、黒川らは先見力検定における回答傾向の分析を行っているが、先見力のある人物の特徴分析にまで踏み込んでいない[8]。

そこで、本研究では先見性のある発言と先見性のない発言が大量に混在する Web 上の言語資源として、通信販売サイト Amazon[9]のカスタマーレビューに注目する。Amazon のカスタマーレビューでは、ユーザは商品に対する評点を 5 点満点でつけ、その理由や商品を使用した感想などを自由にコメントとして残すことができる。各ユーザの評点の平均点がその商品の評価として各商品の下に表示される。

数ある商品レビューの中には、発売当初は評価が高いものの、後にその評価が下がるものや、逆に発売当初は評価が低いものの、後に評価が上がるものがある。この原因の一つとして、発売当初は広告・宣伝などの情報に流され、多数派の意見に賛成するように評点をつけるユーザが多いことが挙げられる。ゆえに、発売当初において、多数派の意見に賛成せずにその商品の後の評価につながる評点をつけたユーザには先見性があり、多数派の意見に賛成した評点をつけたユーザには先見性がないと見なすことができる。先見性のあるユーザと先見性のないユーザのコメントの特徴を比較することで先見性のある発言の特徴が明らかになると考えられる。

2 データ収集

本研究で分析対象とする書籍及びその著者、発売日を表 1 に示す。これらの本は全て人手で探したものである。「福祉国家の戦い」は発売当初評価が低かったものの後に評価が上がった書籍であり、その他の書籍は発売当初の評価が高く後に評価が下がった書籍である。

まず、これらの商品の Amazon の評点をレビュー順の古いものから収集し、その移動平均を求めることにより発売当初と比較してその評価が大きく変動した転換期を求める。たとえば、『1 ドル 50 円時代を生き抜く日本経済』は 2012 年 12 月に安部政権が発足して急激に円安が進んだ時点が転換期となっている。

本研究では、転換期よりも前の時点でその商品に対してレビューをしているユーザのうち、「発売当初評価が高かった商品に 1 点または 2 点の評点をつけている、または発売当初評価が低かった商品に 4 点または 5 点の評価をつけているユーザ」を先見性があると定義し、「発売当初評価が低かった商品に 4 点または 5 点の評点をつけている、または発売当初評価が低かった商品に 1 点または 2 点の評価をつけているユーザ」を先見性がないと定義する。

この基準で先見性のあるユーザと先見性のないユーザを定義すると、初期の多数派が先見力のないユーザなので、ユーザ数に著しい差異が生じる。そこで、レビュー総数が圧倒的に多い『お金は銀行に預けるな』と『殉愛』については、先見性のあるユーザのみを抽出することとする。そして、先見性があるユーザと先見性がないユーザについて、各ユーザの Amazon での商品レビュー全てを収集する。使用するレビューを収集した書籍へのレビューに限定した場合、肯定的な意見か否定的な意見かによる差異が機械学習の結果に大きく影響すると考えられる。これを防ぐため、抽出したユーザが過去に書いた全てのレビューを対象とする。

表 1　分析対象とする書籍

書籍	筆者	発売日
1 ドル 50 円時代を生き抜く日本経済	浜矩子	2011/1
お金は銀行に預けるな	勝間和代	2007/11
小保方晴子さん守護霊インタビュー	大川隆法	2014/4
憲法	青柳幸一	2015/2
知らずに他人を傷つける人たち	香山リカ	2007/2
殉愛	百田尚樹	2014/11
2015 年放射能クライシス	武田邦彦	2011/9
福祉国家の闘い	武田龍夫	2001/2
招かれざる大臣	長妻昭	2011/2
劣化する日本人	香山リカ	2014/7
悪いのは私じゃない症候群	香山リカ	2009/8

上の条件に当てはまるカスタマーレビューを収集した結果、先見性のあるユーザを 72 人、先見性のないユーザを 71 人収集できた。先見性のあるユーザのレビュー数の合計は 16242 件、先見性のないユーザのレビュー数の合計は 11801 件であった。なお、これらのレビューは 2015 年 10 月から 12 月にかけて Amazon のホームページで公開されていたものである。

先見性のあるユーザが過去にレビューした商品のカテゴリ別の割合を図 1 に、先見性のないユーザが過去に

レビューした商品のカテゴリ別の割合を図 2 に示す。先見性のあるユーザがレビューしている商品では、本と洋書で 74%、DVD が 12%、ミュージックが 8%を占める。一方、先見性のないユーザのレビュー商品では、本が 61%（洋書 0%）、DVD が 16%、ミュージックが 5%を占める。よって、先見性のあるユーザは先見性のないユーザに比べ 13 ポイント多く書籍類をレビューしている。

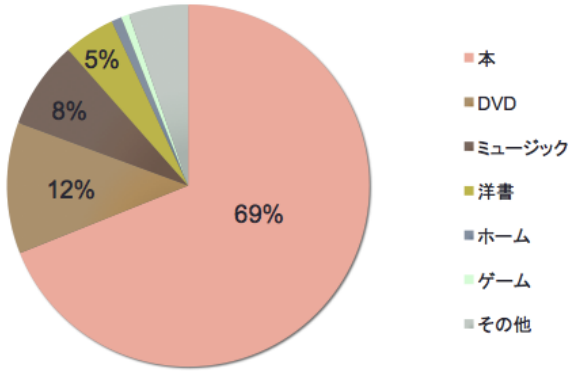

図 1　レビュー商品カテゴリ比率(先見性のあるユーザ)

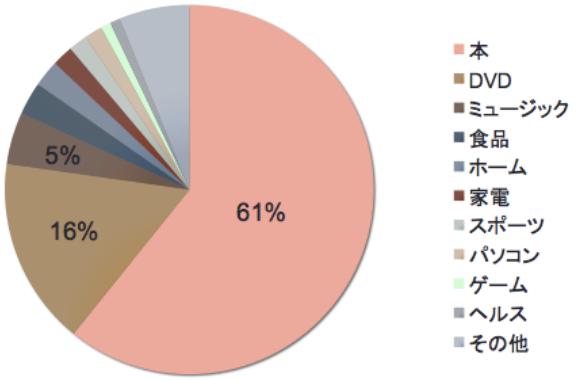

図 2　レビュー商品カテゴリ比率(先見性のないユーザ)

次に書籍の中で、どのようなジャンルの本のレビューをしているかを調べる。先見性のあるユーザが過去にレビューした本のカテゴリ別の割合を図 3 に示し、先見性のないユーザが過去にレビューした本のカテゴリ別の割合を図 4 に示す。先見性のあるユーザがレビューしている本では、文学・評論が 31%、人文・思想が 10%、ビジネス・経済が 9%を占める。一方、先見性のないユーザがレビューしている本では、文学・評論が 22%、人文・思想が 18%、ビジネス・経済が 13%を占める。先見性のあるユーザは先見性のない対象者に比べ 9 ポイント多く文学・評論にレビューをしていて、先見性のないユーザは先見性のある対象者に比べ 8 ポイント多く人文・思想にレビューをしていることが分かる。また、先見性のないユーザがレビューする本には、古賀茂明氏の著書が多いという特徴も得られている。

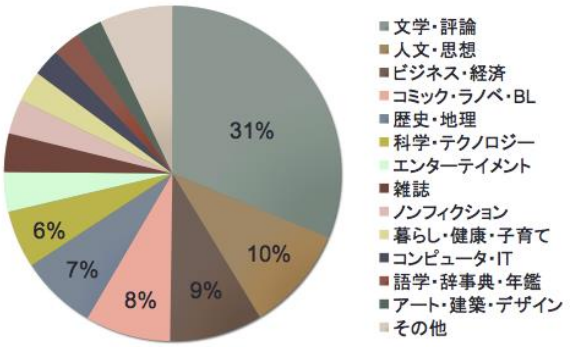

図 3　レビュー本のカテゴリ比率(先見性のあるユーザ)

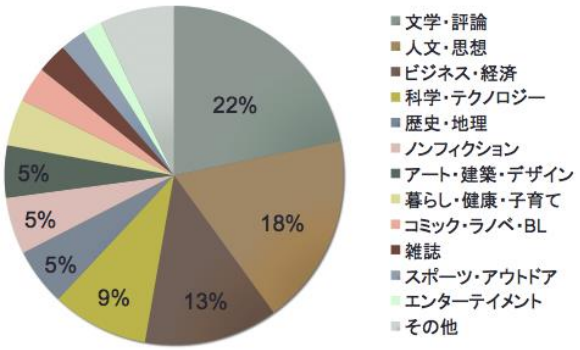

図 4　レビュー本のカテゴリ比率(先見性のないユーザ)

3　機械学習

上で収集したレビューについて形態素解析ツール ChaSen[10]を使って、それぞれの単語すなわち素性ごとに分割し、最大エントロピー法[11]を用いた機械学習にかける。素性は名詞、形容詞、動詞を用いて学習する。

レビュー件数はユーザごとに異なる。各カテゴリ内のデータの文字数がある程度均一でないと文字数の多い対象者のレビューに機械学習の結果が強く影響される。そこで、同一ユーザのレビューは 130 件までの使用とする。また、判定するデータセットの一件あたりの文章量

は多いほうがカテゴリ分類する上でヒントとなる素性の数が増えるため正解率は上昇すると考えられる。そこで実験では一つのデータセットをレビュー10 件、20 件、30 件でまとめた場合のそれぞれで機械学習を行う。また、先見性のある対象者と先見性のない対象者がレビューする商品のカテゴリに大きく差があることから、商品カテゴリの差によって機械学習と判定が行われてしまうことを防ぐため、商品カテゴリを本に限定して機械学習を行うこととする。

以上の条件を満たすレビューを各カテゴリ 2690 件ずつ使用し機械学習を行い、クロスバリデーションの分割数を変えて実験を行った。クロスバリデーションでの正解率をまとめたものが図 5 である。また、最大エントロピー法の判定で得られる確率（確信度）とクロスバリデーションの正解率の関係を図 6 に示す。

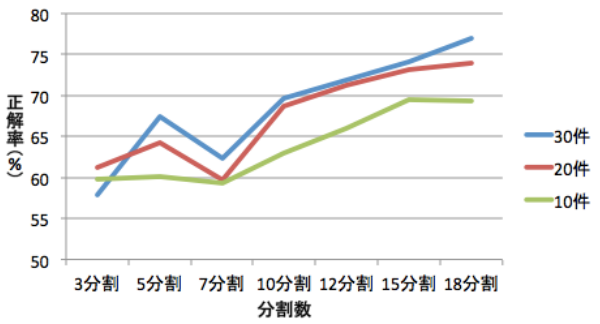

図 5　各データセットの分割数別の正解率

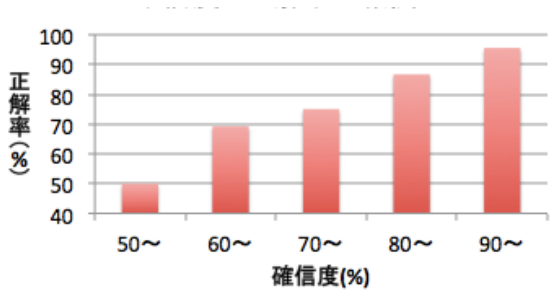

図 7　確信度と正解率の相関

機械学習で得られた素性の特徴としては、先見性のあるレビューには、「作者」の「自己」満足、分かり「にくい」、「新しい」切り口といった本の感想や、「最初」の一冊におすすめ、「十分」、不「十分」といった他のユーザへの推薦に言及したものが多く見られる（括弧付きがカテゴリの上位素性）。一方、先見性のないレビューには「テレビ」化した本、「メディア」や「テレビ」に出ている著者といった本を手にした経緯に言及したものが多く見られる。このことから、マスコミの情報に流される人は先見性に欠ける傾向があると考えられる。

4　まとめ

本研究では先見性のある人物と先見性のない人物の特徴を見出す方法として、Amazon のカスタマーレビューを使用した機械学習を提案した。その結果、先見性のある人物とない人物に関する特徴を得ることについて一定の成果がえられた。今回は評価が大きく変化した本を人手で 11 冊見つけてレビューを収集したが、評価が大きく変化した本を自動的に判定・収集することができれば標本データが増えて正解率を高くできる可能性がある。先見性のある人物と先見性のない人物の特徴が明らかになることで、先見性のある人物への投票や先見性のある経営トップのいる会社への就職などの支援ができると考えられる。

参考文献

[1]東ほか，言語処理学会第 17 回年次大会, 2011
[2]東，掛谷，言語処理学会第 18 回年次大会, 2012
[3]橋本，掛谷，言語処理学会第 16 回年次大会, 2010
[4]畑中ほか，言語処理学会第 15 回年次大会, 2009
[5]尾崎，掛谷，言語処理学会第 20 回年次大会, 2014
[6]橋本，掛谷，第 5 回メディア情報検証学会, 2009
[7]佐藤、掛谷，言語処理学会第 21 回年次大会,2015
[8]黒川，掛谷，第 5 回メディア情報検証学会, 2009
[9]http://www.amazon.co.jp
[10]http://chasen.naist.jp/hiki/ChaSen/
[11]http://www2.nict.go.jp/univ-com/multi_trans/member/mutiyama/index-ja.html