書評レビューを用いた
科学リテラシーを持つ人物の特徴分析

斉藤壮司 **掛谷英紀**

amago2981@gmail.com kake@iit.tsukuba.ac.jp

筑波大学

概要 本研究は、自然言語処理技術を用いてAmazon.jpに存在する書籍のレビュー情報を分析し、科学リテラシーを持つ人物と持たない人物の特徴を見出すことを目的とする。近年、科学リテラシーが社会一般に広く問われた事象として、STAP細胞事件があった。この事件をめぐり、関係者2名が著書を発表している。そこで、Web上に存在するこれらのレビュー情報を、科学リテラシーを持つ集団と持たない集団に定義付けしたうえで収集した。レビュー文章からは機械学習を用いることにより両集団の上位素性を、レビュー商品からは両集団の商品嗜好と書籍嗜好を得た。

キーワード: STAP細胞事件、科学リテラシー、自然言語処理、機械学習、Amazon.jp

1. はじめに

昨今、メディアからの情報に接する個人が、科学リテラシーについて問われる事態が相次でいる。東日本大震災を端に発した原発と再生可能エネルギーを巡る問題、疑似科学とされるマイナスイオン、水素水、EM菌(有用微生物群)の類を利用したビジネス、そして本研究が取り上げる題材であり、研究者の自殺にまで発展したSTAP細胞事件など、具体例は枚挙に暇がない。これらの問題で事実を見誤ると、多数の消費者が経済的被害を受けたり、非現実的な政策決定を生み出したりする恐れがある。さらには、将来の健全な科学技術の発展を阻害する要因ともなりうる。

以上の問題を乗り越えるには、個々人が自らの科学リテラシーを高める必要がある。ところが、これを阻害する3つの大きな問題がある。第一に、正確な情報がメディアを通じて個人に届かないことがある。第二に、インターネットの発達により個人による一次情報の入手が容易となったにも関わらず、事実を誤認してしまうことである。第三に、科学リテラシーの有無やその大小を判断する明確な基準が存在しないことである。この基準を定義することは、これまで行われていない。

しかし、ビッグデータ全盛の現在、上記の指標を提示しうる情報資源がWeb上に蓄積しつつある。佐藤は、Web上の書籍レビューを利用し、先見性のある人物の特徴を検討している[1]。これは、発売期間中に社会的評価が大きく変動した複数の書籍に注目し、それら書籍レビューの特徴的表現を機械学習により抽出したものである。佐藤の手法は、科学リテラシーを持つ人物の特徴分析に応用可能であると考えられる。

以上を踏まえ、本研究では科学について世論を二分する出来事に注目し、その関係者が出版した書籍のレビュー情報を自然言語処理により分析することで、科学リテラシーを持つ人物の特徴を明らかにする。

2. 手法

2.1 概要

個人の科学リテラシーを反映するデータとして、Amazon.jp[2]に掲載されている商品レビューを利用する。分析対象の書籍には、小保方晴子氏著「あの日」、須田桃子氏著「捏造の科学者」の2冊を選定する[3,4]。

小保方晴子氏は、2014年にSTAP細胞の作成で一躍有名となった元理化学研究所の研究者である。彼女がNatureに投稿した論文からは剽窃が発見され、その撤回を余儀なくされている。また、後の検証実験においてもSTAP細胞の存在は確認されていない。しかし、現在も小保方氏はSTAP細胞の存在を主張しており、「あの日」はその釈明本といえる。

須田桃子氏は、STAP 細胞事件を批判的に取材した毎日新聞の科学記者である。「捏造の科学者」は彼女が追った事件の一部始終を収めた書籍であり、第 46 回大宅壮一ノンフィクション賞を受賞している。

両書籍には Amazon.jp において多くのレビューがなされており、本研究で利用するデータを取得する対象として、適切と判断する。

Amazon.jp では、レビューアが書籍に 5 段階の評価を与えられる。STAP 細胞研究の科学的手順に多くの不正が認定されている事実に照らすと、「あの日」を高評価（評価点 4、5）または「捏造の科学者」を低評価（評価点 1、2）としたレビューアは科学的判断に誤りがあると見なすことができる。そこで、彼らを「科学リテラシーを持たない人物」として定義する。逆に、前者を低評価または後者を高評価としたレビューアを「科学リテラシーを持つ人物」として定義する。

これらの定義付けのもと、両者が購入した「あの日」と「捏造の科学者」を含む全ての商品レビューを取得する。そして、各レビュー文章に現れる上位素性を機械学習によって求め、特徴的表現を定量化する。また、各レビューアのレビュー商品に差異が見られるか検討し、科学リテラシーの有無と商品嗜好や書籍嗜好の関係を分析する。以上により、科学リテラシーを持つ人物の特徴を見出す。

2．2　機械学習

本研究で用いる機械学習システムの概要を図 1 に示す。

まず、レビュー文章を単語単位に分割して品詞情報を付与するため、形態素解析システム MeCab[5]を利用する。本研究では、動詞、形容詞、そして名詞に限って利用する。その他の単語は科学リテラシーに関係しないものが多いと考えられ、機械学習に利用する素性から除外する。

次に、日本語教育語彙表 ver1.0[6]を利用し、ここに掲載のある素性だけに絞る。この語彙表は、言語学習への利用を想定した辞書であり、約 18,000 語を収録する。我々が日常的に利用する素性に注目することで、機械学習の結果が特定の分野に関する語句に影響されることを防ぐ。

最後に、得られた素性群を最大エントロピー法による機械学習へ入力し、それらが科学リテラシーを持つ集団と持たない集団どちらに属する確率が高いかを計算することで、上位素性を得る。最大エントロピー法を実行するツールには Maxent 2.03 を利用し、クロスバリデーションによって正解率を検証する。

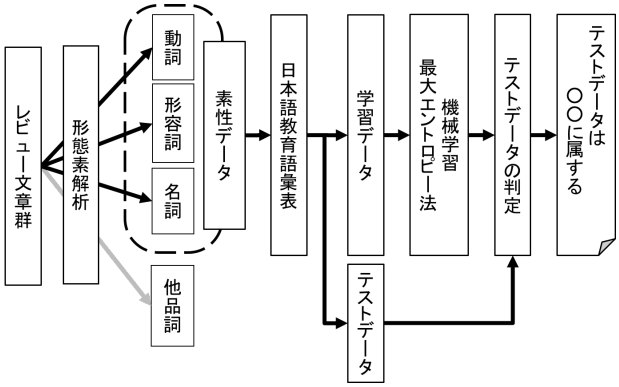

図 1　機械学習システム

3．実験と結果

3．1　データ収集

本研究で利用する商品レビューは、2016 年 8 月 8 日から 2016 年 10 月 6 日の期間に、Web スクレイピングにより Amazon.jp から取得した。その集計結果について、科学リテラシーをもつ集団は表 1、持たない集団は表 2 に示す。

取得したレビューア人数は、「あの日」の高評価が 455 件、低評価が 149 件であった。また、「捏造の科学者」の高評価が 79 件、低評価が 60 件であった。「あの日」や「捏造の科学者」を肯定しているか明確な立場を示していないと考えられるため、評価点が 3 のレビューアは利用しなかった。

取得したレビュー件数は、「あの日」の高評価が 12,442 件、低評価が 8,147 件であった。また、「捏造の科学者」の高評価が 12,349 件、低評価が 3,185 件であった。

表1　科学リテラシーを持つ集団

書籍名	あの日	捏造の科学者
評価点	1, 2	4, 5
レビューア	149 人	79 人
商品レビュー	8,147 件	12,349 件

表 2　科学リテラシーを持たない集団

書籍名	あの日	捏造の科学者
評価点	4, 5	1, 2
レビューア	455 人	60 人
商品レビュー	12,442 件	3,185 件

3．2　商品嗜好

科学リテラシーを持つ集団が過去にレビューした商品について、カテゴリ別の割合を図 2 に示す。同様に持たない集団について図 3 に示す。

科学リテラシーを持つ集団は、上位から本が 70%、DVD が 4%、ミュージックが 4%、食品・飲料・お酒が 3%を占める。科学リテラシーを持たない集団は、上位から本が 44%、ミュージックが 13%、DVD が 9%、家電&カメラが 4%を占める。

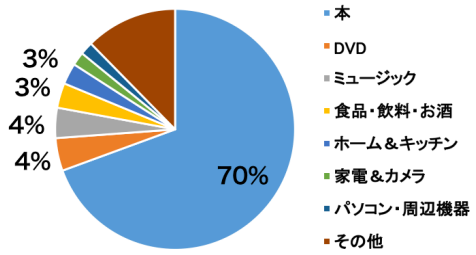

図 2　科学リテラシーを持つ集団の商品嗜好

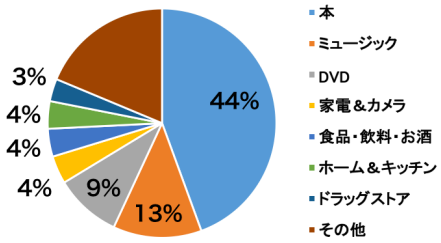

図 3　科学リテラシーを持たない集団の商品嗜好

3．3　書籍嗜好

3.2 節の結果について、さらに本、洋書、Kindle 本の 3 分類に絞り、ジャンル別の内訳を集計した。科学リテラシーを持つ集団を図 4 に、持たない集団を図 5 に示す。

科学リテラシーを持つ集団は、上位から文学・論評が 18%、コミック・ラノベ・BL が 14%、社会・政治が 10%、人文・思想が 10%を占める。科学リテラシーを持たない集団は、上位から文学・論評が 24%、社会・政治が 14%、人文・思想が 13%、ビジネス・経済が 5%を占める。

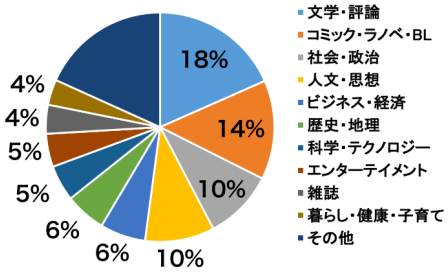

図 4　科学リテラシーを持つ集団の書籍嗜好

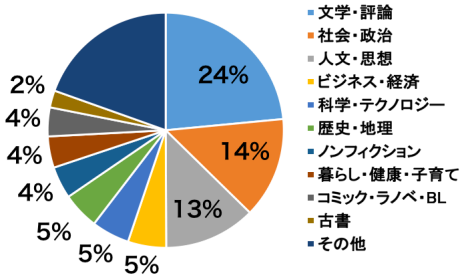

図 5　科学リテラシーを持たない集団の書籍嗜好

3．4　機械学習

機械学習を用いて両集団の上位素性を求めるにあたり、以下の 3 条件を満たすデータセットを作成した。

第一に、短いレビュー同士はまとめ、一つのデータセットあたりの文字数を 8,000 字以上とした。佐藤によれば[1]、各データセットの文字数に大きな偏りがあると、文字数の多いデータセットに機械学習の結果も偏る。また、判定のヒントとなる素性を多数含むため、文字数の多い方が正解率は上昇する。

第二に、利用するレビューは同一レビューアあたり新着順に最大 100 件までとする。レビューアの約 9 割が 100 件以下のレビュー数であるが、中には 1,000 件以上のレビューアも存在する。この条件により、機械学習の結果が特定のレビューアに偏るのを防ぐ。

第三に、クロスバリデーションの分割単位をまたがないよう、同一レビューアのレビューをデータセットに配置する。理由は、学習データとテストデータに同一レビューアのレビューが存在すると、そのレビューアの特徴を機械学習する恐れがあるからである。クロスバリデーションの分割は 20 分割とした。

以上を踏まえ、動詞、名詞、形容詞を用いた機械学習を行った結果、63%の正解率を得た。この判定において重要な手掛かりとなった特徴的な 30 個の素性を上位から表 3 に示す。

表 3　上位素性

科学リテラシーを持つ集団	科学リテラシーを持たない集団
指摘	マスコミ
上がる	十分
性格	テレビ
安っぽい	元気
勧める	意図
解決	応援
タイトル	報道
笑い	終える
辛い	実用
一生	役
用語	語
変	含める
好き	始める
数	正義
驚く	潰す
食べる	サラリーマン
予想	歌
末	組織
出版	集団
逆	一部
唯一	非難
通り	取材
彼	平和
触れる	一人
質	映画
濃い	騒動
概念	番組
イギリス	頭
丁寧	構造
文庫	姿勢

4. 考察

3.2 節で示した商品嗜好の違いから、科学リテラシーを持つ人物は、持たない人物と比較して読書習慣があると考えられる。本カテゴリの割合に注目すると、科学リテラシーを持つ集団は 70%を占めるのに対し、持たない集団は 44%を占める。この差が科学的リテラシーの有無に影響していると考えられる。

3.3 節で示した書籍嗜好の違いから、科学リテラシーを持つ人物は持たない人物と比較して漫画などのフィクション作品を好むと考えられる。コミック・ラノベ・BL カテゴリの占める割合に注目すると、科学リテラシーを持つ集団は 14%を占め上位 2 位であるのに対し、持たない集団は 4%で 9 位である。科学リテラシーを持つ集団は物語背景の理解を必要とする漫画やライトノベルを多数読むことにより、その読解記述力を涵養していると考えられる。

3.4 節で求めた両集団の上位素性を比較すると、科学リテラシーを持たない集団はマスメディアに関する素性が数多くみられる。これらの登場場面の例を挙げると、「マスコミ」の一方的な「報道」、「テレビ」と DVD を衛星アンテナに接続、綿密な「取材」とデータ、米倉涼子の「テレビ」「番組」、などがある。マスメディアへの関心の高さや各種媒体からの情報量が、科学リテラシーの大小に影響している可能性がある。

以上の特徴は、佐藤により見出された先見力のある人物の特徴と重なる部分が多くあり、先見性と科学リテラシーの関連性が示唆される。

5. おわりに

本研究では、科学リテラシーを持つ人物の特徴を見出すにあたり、STAP 細胞事件の関係者 2 名が出版する書籍のレビュー情報を Amazon.jp から収集し、自然言語処理技術により分析することを試みた。その結果、レビュー商品やレビュー書籍の嗜好に違いが見られた。また、機械学習により両集団のレビュー文章における特徴的な素性を得た。クロスバリデーションによる正解率は 63%と低調であったが、科学リテラシーを持たない人物はマスメディアへの関心が高いことがわかった。

今回は、STAP 細胞事件に関する Web 上の言論のみに注目した。科学リテラシーについて、一般性のある特徴を見出すためには、この事件と同様の事例を他に多数収集し、併せて分析する必要がある。

参考文献

[1] 掛谷英紀、佐藤裕也 : 書評レビューに基づく先見性のある人物の特徴分析, 言語処理学会第 22 回年次大会, 2016

[2] Amazon.jp, https://www.amazon.co.jp

[3] 小保方晴子, あの日, 講談社, 2016

[4] 須田桃子, 捏造の科学者 STAP 細胞事件, 文藝春秋, 2015

[5] MeCab, http://taku910.github.io/mecab/

[6] 砂川有里子, 日本語教育語彙表 ver1.0, http://jhlee.sakura.ne.jp/JEV.html