6N-4

対象物と属性値の印象比較による統計表からの意外な事実の抽出

松井浩平† 西野享† 松村冬子‡ 原田実‡

青山学院大学理工学部情報テクノロジー学科†

1. はじめに

誰でも自由に機械可読形式でデータを活用することができるオープンデータが注目されており、日本政府をはじめ、各省庁や自治体などが統計表など保有する様々なデータをデータカタログの形態で公開しはじめている。行政側としては、公開されたデータを用いて一般の開発者により市民の目線で有用なWebサービスなどが開発されることに期待している。一方で、多くのデータカタログサイトに備えられている従来のキーワード検索やカテゴリ検索のみでは開発者などのユーザがデータに興味を持つことが難しく、データの活用が進まないことが懸念される。そこで本研究では、公開されたデータに含まれる事実とユーザの思い込みとの差分から意外性のある事実を提示するデータを提示し、そのデータへの興味を喚起することを目指す。

2. 本研究における意外性の定義

奥[1]の報告によると、新規性、意外性、セレンディピティ向上を目指した情報推薦の取り組みは、1)情報リスト編集に基づく方式、2)ユーザモデルに基づく方式、3)ユーザインタラクションに基づく方式の3つに分類することができ、本研究のアプローチは2)に分類することができる。同じ方式の先行研究としては、ユーザが習慣的に選択するデータを予測する習慣モデルと好みのデータを予測する嗜好モデルの2種類のモデルを導入し、その予測結果の差異に基づいて意外性を推定する手法[2]や、ブックマークのタグ群から推薦結果を多様化させることで、嗜好に合致し、かつ未知の情報を推薦する手法[3]などが提案されている。これに対して，本研究の意外性の定義としては、図1に示すようにある対象物の属性値について、対象物自体に対するユーザの印象から推測される属性値と、統計表に含まれる実際の属性値を比較した際に算出される乖離の大きさを意外性の高さとみなす。この乖離が大きいほどユーザにとってそのデータは意外性が高いと評価されると考える。例えば、図1のように東京都という対象物についての印象から推測される平均労働時間という属性の値と、実際の東京都の平均労働時間の値との差異が意外性を示す指標となる。先行研究では習慣性や未知性によって意外性を表現しているが、本研究では対象物に対する思い込みと事実との差異により意外性を表現する点が異なる。

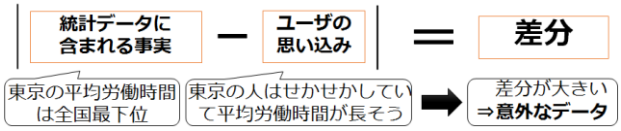

図1 本研究における意外性の概念

3. 意外性の表現とその評価

意外性の算出には、1)印象と属性の値の関係、2)対象自体に対するユーザの印象の2つを求める必要がある。1)については、その属性の値が大きい対象と小さい対象についての印象をSD法（Semantic Differential法）で獲得し、被験者全員から獲得したSD法の各尺度に対する評価値を説明変数、属性値を目的変数として重回帰分析を行いその属性の値と印象の関係を式(1)に示す重回帰式として求める。SD法の尺度の総数がnであるとき、式(1)における$a_1, a_2, \cdots, a_n$は各尺度に対応する偏回帰係数、bは切片、$x_1, x_2, \cdots, x_n$は各尺度の評価値となる。

$$y = a_1x_1 + a_2x_2 + a_3x_3 + \cdots + a_nx_n + b \quad (1)$$

例えば、対象物を都道府県、属性を平均労働時間と設定すると、まず平均労働時間が長い都道府県および短い都道府県それぞれに関する印象を獲得する。重回帰分析の目的変数には、平均労働時間が上位もしくは下位の都道府県の平均値を用いることで、平均労働時間と都道府県に対する印象の関係を重回帰式として表すことができる。

次に、2)として各都道府県自体についての印象をSD法で獲得し、平均労働時間と印象の関係を表す重回帰式に対象となる都道府県の印象を表す各尺度の評価値を代入することで、被験者の印象から予測される対象の都道府県の平均労働時間の値が算出される。図1に示したように、統計表上の実際のその都道府県の平均労働時間の値と、算出された被験者のイメージする平均労働時間の予測値の差が意外性を評価する尺度となる。つまり、この差分が大きいほど意外な事実として評価されることになる。例えば、東京にせかせかしていて多忙な印象を持つ被験者の印象を代入

Extracting Unexpected Fact from Statistical Data based on Comparison of Impression between Item and Its Attributes
Kohei Matsui†, Toru Nishino†, Fuyuko Matsumura‡ and Minoru Harada‡
†Undergraduate school of Integrated Information Technology, Aoyama Gakuin University.
‡Department of Integrated Information Technology, Faculty of Science and Engineering, Aoyama Gakuin University.

すると、その値は平均労働時間が多い都道府県に近づく。実際には東京は他府県と比較して平均労働時間が短いため、この事実は意外とみなされる。

4. 評価実験

4.1 実験内容

統計表に含まれる事実とユーザの思い込みの差分から、意外性のあるデータを提示できるか検証する実験を行った。後述する実験システムを使用して、被験者に SD 法によるアンケートに解答してもらった。被験者は学生 20 名（男性 17 名、女性 3 名、21 歳～24 歳）とした。SD 法に用いる尺度（形容詞対）は、井上らの報告[4]における「社会」の分野の実験において使用頻度が 8 以上の項目である計 12 項目を選択した。

実験に用いる統計データは、総務省統計センターにより e-Stat で公開されている平成 23 年社会生活基本調査 調査票 A に基づく生活時間に関する結果の生活時間（地域）編 1-1 表とし、被験者が学生ということを考慮して、学生がイメージしやすい属性値として男女それぞれが「食事」「学業」「趣味・娯楽」に必要とする生活時間を選択した。目的変数とする大きい属性値の代表値としては、属性値の大きさが上位 5 位以内に入る都道府県の属性値の平均値を、小さい属性値の代表値としては、属性値の大きさが下位 5 位以内に入る都道府県の属性値の平均値を用いた。また、印象を獲得する対象の都道府県については、対象都道府県を北海道、東京都、神奈川県、大阪府、沖縄県に限定して実験を行った。

4.2 実験システムの概要

SD 法を用いた実験システムを Web アプリケーションとして実装した。被験者はシステムを用いて、対象となる各都道府県に対する印象、各属性の値が大きい都道府県および値が小さい都道府県に対する印象を 5 段階で回答する。また、被験者から取得した各尺度についての評価値をもとに、R を用いて属性の値と印象の関係を表す重回帰式を導出する。導出された式に基づき、各被験者が都道府県の印象として回答した各尺度の評価値を代入し予測値を算出することで、実際の属性の値との差分を提示する.

5. 実験結果および考察

実験結果から差分が提示できた事例を示す。属性値「男性が生活時間の中で“学業”に費やす時間」に関しては、沖縄県において、実際の値が 53 分/日（全国 1 位）であるのに対し、ある 2 名の被験者の印象によるその予測値は、34.7 分/日や 33.8 分/日となり、これらは全国 46 位の北海道および和歌山県の値、35 分/日とほぼ一致しており、大きな差が見られた。東京においては、実際の値が 36 分/日（全国 44 位）であるのに対し、予測値は、54.8 分/日や 53.6 分/日などが見られ、これらは全国 1 位の沖縄県の値とほぼ一致しており、大きな差が見られた。

また、都道府県別に算出されたデータを見ると、沖縄県および東京都に関しては多くの属性値において被験者の印象と実際の値の差分が大きかった。一方で、神奈川県・大阪府に関してはほぼ全ての属性値において、被験者の印象と実際の値が一致していた。このことから、印象と事実との乖離が顕著に見られる都道府県とそうでない都道府県があることがわかった。

実験を通して、被験者間で属性値に対する印象に極端なばらつきが存在する場合、予測精度が低くなり意外性の提示が難しくなることがわかった。これより属性値に対しての印象が被験者間で一貫性が高いことが求められるということが、提案手法における問題点であると言える。現時点では差分の抽出まで行ったが、今後被験者に差分の大きいデータを提示し意外かどうかを判定してもらい、意外性の判断基準についても検証を行う。

6. おわりに

本稿では、意外性のある事実を提示することでデータへの興味を喚起することを目的とし、統計データにおける対象に対する思い込みと事実との差異により意外性を表現する推薦手法を提案した。評価実験からは、重回帰式により印象から予測された属性値と実際の属性値の差分を確認することができた。しかし、現時点では意外性の評価については検証できていないため、今後も継続して被験者実験を行い意外性の予測が可能かどうかを検証していく。また、属性値の印象にばらつきが出た場合の分析方法についても検討していく必要がある。

参考文献

[1]奥健太: セレンディピティ指向情報推薦の研究動向、知能と情報（日本知能情報ファジィ学会誌）-特集: Web インテリジェンスとインタラクションⅡ-、Vol25、No.1、pp.2-10、2013.
[2]村上知子、森紘一郎、折原良平: 推薦の意外性向上のための手法とその評価、人工知能学会論文誌 24(5)、428-436、2009.
[3]藤原誠、中川博之、田原康之、大須賀昭彦: タグクラスタ多様化による未知性を考慮した推薦手法の提案、情報処理学会研究報告.ICS、[知能と複雑系]2012-ICS-167(4)、1-6、2012.
[4]井上正明、小林利宣: 日本における SD による研究分野とその形容詞対尺度構成の概観、教育心理学研究、33(3): pp.253-260、1985.