

# 集団評価の二部グラフモデルと相互強化型解析手法

俵本 一輝<sup>†</sup> 川本 淳平<sup>†</sup> 浅野 泰仁<sup>†</sup> 吉川 正俊<sup>†</sup>

<sup>†</sup> 京都大学大学院情報学研究科 〒 606-8501 京都府京都市左京区吉田本町

E-mail: †{tawara,j.kawamoto}@db.soc.i.kyoto-u.ac.jp, ††{asano,yoshikawa}@i.kyoto-u.ac.jp

あらまし 本稿では、集団評価の二部グラフモデルと、特異な評価者の特定及び評価対象への一般的な評価を計算するための相互強化型解析手法を提案する。本モデルでは、節点が評価者、評価対象を表し、辺が評価を表す特徴量を持つ。既存のグラフを用いたモデルでは、辺に関連付ける特徴量として実数値のみを扱うことが一般的であるが、本稿では、実数値以外のベクトル、分布といった表現形式の特徴量も扱うことができるモデルと手法を提案し、それぞれの特徴量を用いた評価実験を行う。

キーワード 二部グラフモデル, 相互強化, 外れ値検出

## Bipartite Graph Model and Mutually Reinforcing Analysis for Group Evaluation

Kazuki TAWARAMOTO<sup>†</sup>, Junpei KAWAMOTO<sup>†</sup>, Yasuhito ASANO<sup>†</sup>, and Masatoshi YOSHIKAWA<sup>†</sup>

<sup>†</sup> Graduate School of Informatics, Kyoto University Yoshidahonmachi, Sakyo-ku, Kyoto, 606-8591 Japan

E-mail: †{tawara,j.kawamoto}@db.soc.i.kyoto-u.ac.jp, ††{asano,yoshikawa}@i.kyoto-u.ac.jp

### 1. はじめに

Web 上にある膨大な情報は知識として広く活用されているが、ページの情報には知識とともにそれに対する提供者の意見や感情が含まれているものもある。例えば、製品に対するレビューやニュース記事には筆者の意見や感情が含まれていると考えられる。また、ブログやソーシャル・ネットワークキングサービスなどを通じて人々により発信される情報の中にも、意見や感情が含まれていると考えられる。ここでは、そのような意見や感情を広義の意味での評価と捉える。

ある対象への評価が妥当かどうか判断するには、評価者について十分に理解しているか、もしくは、評価された対象について十分に理解している必要がある。通常、web 上に現れる評価の評価者について十分に理解することは困難であるし、また、対象を理解するために評価を利用するのであれば、対象について十分に理解できていない場合が多いと考えられる。一方、定量化された評価であれば、多くの人は同一の対象に対して似通った評価を行うという考えから、評価の平均をとるなど、統計的に扱うことで対象への一般的な評価を知ることが可能である。このような一般的な評価であれば、対象について十分に理解していなくても、受け入れやすいと考えられる。しかしながら

ら、そのような統計的な処理は、評価者が少ない場合には、特異な評価値を与える人の影響により、歪められた評価となる可能性がある。そこで本稿では、特異な評価をする人や対象への一般的な評価を特定するための手法を提案する。

評価の定量化にはいくつかの手法が存在している。例えば、レビューサイトの評価スケールのようにユーザが点数を与えることで評価が行われる場合やコンテンツに対して付与されるタグについては、ベクトルを用いて定量化することが可能である。一方で、ニュースサイトやブログのようにテキスト中に自然言語を用いて意見や感情といった評価が行われる場合もある。そのような際には、tf・idf による特徴ベクトルを評価ベクトルとしたり、positive と感じたか negative と感じたかの極性値を評価値とすることが可能である。また、「驚き」や「悲しみ」といった感情のように、positive とも negative とも判断され得る場合には、基本となる感情が混合することで新しい感情が生じるという考えに基づき、分類した基本感情と各基本感情の強度でベクトル表現することも行われている。

一人の評価者からの評価を個別に扱うのではなく、複数の評価者からの評価を統一的に扱う（集団評価）には、モデル化する必要がある。そこで、人・対象をそれぞれ節点として、評価を辺とする二部グラフでのモデル化を行う。集団評価を扱う際

には「ある人が評価を行ったすべての対象」や「ある対象へ評価を行うすべての人」が扱いやすいと便利であるが、二部グラフでモデル化することで、それらを隣接関係を用いて扱うことができる。特異な評価をする人や対象への一般的な評価の特定は、この二部グラフ上で解析することで行う。

グラフ上での解析には、相互強化関係を利用する。相互強化関係とは「ある性質  $a$  を持つものは、性質  $b$  を持つものから支持されている」、「ある性質  $b$  を持つものは、性質  $a$  を持つものを支持している」というように二つの性質が互いの支持によって強化し合う関係のことを言う。この関係を利用した解析を相互強化型解析と名づける。

以下、2. 節では関連研究についてまとめ、3. 節で提案手法である二部グラフでのモデル化と相互強化型解析手法について説明する。4. 節では提案手法を用いた実験を行い、その結果の考察を行う。最後に5. 節で本稿のまとめと今後の課題について述べる。

## 2. 関連研究

### 2.1 相互強化型解析

相互強化関係を利用する解析を、我々は相互強化型解析と呼んでいるが、そのような解析手法は従来から研究されている。HITS [1] はその代表であり、リンク構造を用いて Web 上に存在するページから良質な情報を多く含んだページを発見するリンク解析の手法である。良質な情報を多く含んだページである Authority と、良質な情報を含んだページへの多くのリンクを持つページである Hub には、「良い Authority は多くの良い Hub からリンクされている」、「良い Hub は多くの良い Authority へリンクしている」という相互強化関係が存在し、この関係を用いて両者の特定を行っている。

[1] は、一般的なグラフ解析同様に実数値のみを扱っているが、我々はそれ以外の形式の特徴量も扱うことができる手法を提案している。

### 2.2 外れ値検出

外れ値検出 (outlier detection, anomaly detection) は、データ中から傾向の異なるものを特定するという問題を扱い、近年では、データマイニングの分野でも注目されている [2]。外れ値の中には、全データ中で見た際に外れ値となるもの (global outlier) や、ある文脈で見た際に外れ値となるが、別の文脈で見た際には外れ値とならないもの (contextual outlier) などが存在する。

我々は、グラフ上で解析を行うことで、外れ値検出を行っているが、同様に、グラフ上で解析を行うことで外れ値検出を行う研究が存在する。例えば、Sun ら [3] は、二部グラフで表されるデータについて、グラフ上での隣接関係を用いて外れ値検出を行っている。ある節点到隣接するという文脈において、その中で特異な節点の特定を行っている。一方で、文脈を決定する事前知識を用いることが一般的であるが、事前知識を利用できない場合も存在する。Wang ら [4] は、そのような際にも有効な、事前知識を用いた文脈の特定を行わない手法を提案している。データを節点とし、データ間の類似度を遷移確率とする

グラフ上で、ランダムウォークを行うことで、文脈とその文脈における外れ値の特定を行っている。

個々の手法は、想定する文脈が異なるという点で区別することができる。我々の手法では [3] と同様にグラフ上での隣接関係という文脈を想定した外れ値検出を行うものである。しかしながら [3] では、隣接関係に基づく、節点同士の類似度を利用するのに対して、我々は辺に関連づけた定量化した評価を利用するという点で異なっている。

## 3. 提案手法

特異な評価者や評価対象への一般的な評価を特定は、集団評価をモデル化したグラフ上での解析により行う。3.1 節では、モデル化して作成する二部グラフについて説明し、3.2 節では、二部グラフ上で行う、相互強化型解析について説明する。

### 3.1 集団評価グラフ

評価者、評価対象はそれぞれ節点として表現され、節点集合  $V_S$  が評価者の集合、節点集合  $V_O$  が評価対象の集合となる。評価者  $p$  が評価対象  $q$  に評価を行っている場合、 $p$  の  $q$  に対する評価を定量化した  $D_{pq}(p \in V_S, q \in V_O)$  が関連付けられた枝  $(p, q)$  が存在し、枝集合  $E$  に含まれる。集団評価グラフは、グラフ  $G(V_S, V_O, E)$  として作成することができる。

ここで、 $V_S$  内、 $V_O$  内に枝が存在しなければ、 $G(V_S, V_O, E)$  は二部グラフとなる。特に、評価対象が人のような評価を行うことができるものではなく、物や出来事のようなものだけである場合には二部グラフとなる。逆に、評価を行うことができるものが評価対象にもなっている場合には、 $p \in V_S \cap V_O$  となるが、 $p' \in V_S, p'' \in V_O$  として区別することで、二部グラフとして扱うことができる。

### 3.2 相互強化型解析手法

特異な評価者と評価対象への一般的な評価を特定には、評価者、評価対象に与えられる、相互強化関係に基づいて定義される特徴量を用いる。評価者の特異な程度と評価対象への一般的な評価には、次のような相互強化関係がある。

- 特異な評価者ほど、多くの評価対象について一般的な評価と極端に異なる評価を行っている
- 評価対象への一般的な評価は、特異でない多くの評価者の評価に近いものとなる

この相互強化関係に基づき、評価者  $p$  には、特異な程度を表す実数値の特徴量  $x_p$  を、評価対象  $q$  には、一般的な評価を表す、定量化された評価と同じ形式の特徴量  $Y_q$  を与える。

まず、評価者  $p$  の特異な程度は、集団評価グラフ上で隣接するすべての評価対象に対して  $D_{pq}$  と  $Y_q$  との非類似度から決まると考えられる。よって、特徴量  $x_p$  を次のように定義する。

$$\mathbf{x}^{<p>} = \{x_q^{<p>}\}, x_q^{<p>} = distance(D_{pq}, Y_q) \quad (1)$$

$$x_p = f(\mathbf{x}^{<p>}) \quad (2)$$

ここで、 $x_q^{<p>}$  は、距離関数  $distance$  を用いて求まる、 $D_{pq}$  と  $Y_q$  の非類似度であり、評価対象  $q$  への評価者  $p$  の評価の特異な程度を表している。 $D_{pq}, Y_q$  の形式に応じて、距離関数

*distance* を選択することで特徴量に依らず適用することができる．評価者  $p$  自身の特異な程度  $x_p$  は、各評価対象での特異な程度を統合して求めることになり、関数  $f$  により、それを実現する． $f$  は、例えば、 $x^{<p>}$  のすべての成分の平均値をとることが考えられる．ここで、各評価対象での特異な程度を統合するにあたり、尺度を揃える必要がある．そこで、式 1 では、 $x_q^{<p>}$  を次のように正規化する．

$$\mu_q = \sum_{p:(p,q) \in E} \frac{x_q^{<p>}}{N_q}, \sigma_q^2 = \sum_{p:(p,q) \in E} \frac{(x_q^{<p>} - \mu_q)^2}{N_q}$$

$$norm(x_q^{<p>}) = \frac{x_q^{<p>} - \mu_q}{\sigma_q}$$

$N_q$  は  $q$  に隣接する節点の数を表す．このとき、 $norm(x_q^{<p>})$  は、特異でない評価ほど大きな負の値をとり、特異な評価ほど大きな正の値をとることになる．

また、評価対象  $q$  が持つ一般的な評価は、集団評価グラフ上で  $q$  に隣接するすべての評価者  $p$  に対する  $D_{pq}$  と  $x_p$  から決まり、極端な評価を行う評価者の評価ほど反映されないと考えられる．よって、特徴量  $Y_q$  を次のように定義する．

$$w_{pq} = \frac{S_a(x_p)}{\sum_{p:(p,q) \in E} S_a(x_p)}, S_a(x) = \frac{1}{1 + \exp(ax)} \quad (3)$$

$$Y_q = \sum_{p:(p,q) \in E} w_{pq} D_{pq} \quad (4)$$

ここで、 $w_{pq}$  は、評価者  $p$  の特異な程度を基に算出される、 $q$  に対する重みである．シグモイド関数を用いることで、特異な評価者ほど 0 に近い重みを、特異でない評価者ほど 1 に近い重みが与えられる．評価対象  $q$  への一般的な評価  $Y_q$  は、 $w_{pq}$  を用いた  $D_{pq}$  の重み付き平均をとることで求める．

各特徴量を求めるには、初期化した後、上述の定義を用いて収束するまで更新を繰り返す．更新を繰り返した後、最終的に収束した値が得られる特徴量となる．

ここで、収束することを前提としたアルゴリズムとなっているが、定性的に収束することは保証されていない．よって、実際には一定回数更新を繰り返し、 $x$  の差分が十分小さくなったところで処理を終了することになる．

## 4. 実験

提案手法を用いて、特異な評価者の特定と評価対象への一般的な評価の特定を行い、その有効性の評価を行う．4.1 節では、今回用いたデータセットについて説明し、4.2 節では、実験の結果を示し、その考察を行う．

### 4.1 データセット

実験には、Yahoo!映画に投稿されたユーザレビューを用いた．Yahoo!映画では、各映画に対して、ユーザが自由にレビュー文を記述している．また、レビュー文以外にも、映画に対する 5 段階での採点や、物語、配役、演出、映像、音楽の 5 項目について 5 段階の評価を行うことができる．集団評価グラフを作成するにあたり、評価を定量化する必要がある．そこで、5 段階

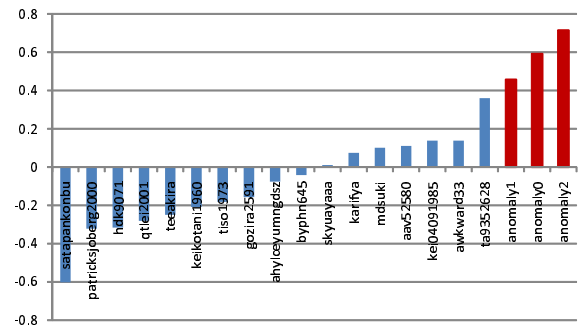


図 1 評価者の特異な程度 (評価値)

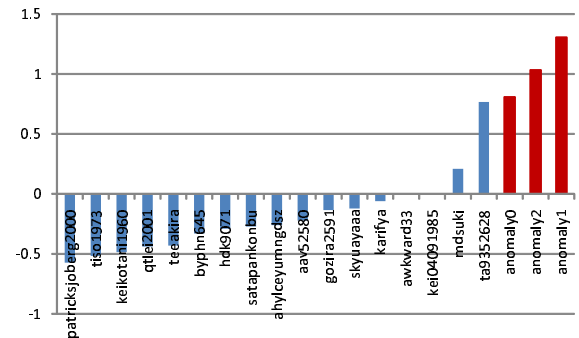


図 2 評価者の特異な程度 (評価ベクトル)

の採点を利用した評価値と 5 項目への評価を利用した評価ベクトルの 2 種類を作成した．提案手法が定量化した評価の表現形式に依らず適用可能であるかを評価するために、それぞれの定量化した評価を用いて実験を行う．

また、特異な評価者が問題となるのは、評価対象へ評価を行う者が少ない場合である．上映が始まった初期の段階の映画やマイナーな映画に対してはレビュー数が少なく、特異な評価者が問題となり得る．このことから、レビュー数が少ない映画タイトルを含むデータセットを作成した．具体的には、レビュー数 17、映画タイトル数 80 からなるデータセットを作成した．

一方で、実際のユーザレビュー中には、特異な評価者が存在しない可能性が存在する．そこで、ランダムに映画タイトルを作成し、ランダムな評価を行う特異なレビュー 3 人をデータセットに追加した．実験では、この (17+3) レビューから成るデータセットを用いて実験を行った．

### 4.2 結果と考察

実験で適用した提案手法の設定について説明する．定量化した評価はベクトル表現であるので、距離関数 *distance* にはユークリッド距離を用いた．また、各評価対象での特異な程度を統合する関数  $f$  は、すべての成分の平均値をとるとした．そして、シグモイド関数のパラメータ  $a$  は 1 とした． $x$  の差分が十分小さくなるまで処理を繰り返すことになるが、今回は更新回数を 20 とした．

まず、評価者の特異な程度を表すスコアを図 1、図 2 に示す．図 1 は評価値を、図 2 は評価ベクトルをそれぞれ定量化した評価として用いた場合の結果である．縦軸が特異な程度を表すス

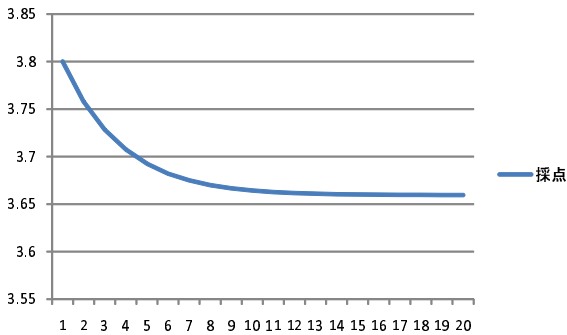


図 3 評価対象の一般的な評価 (評価値)

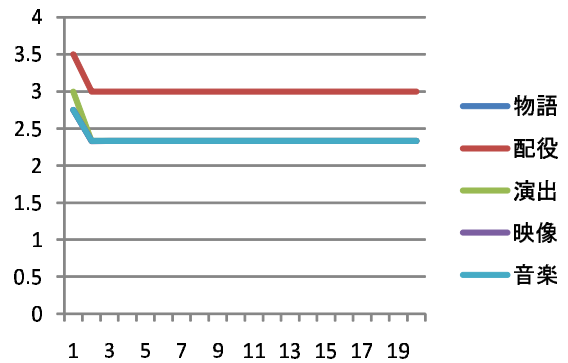


図 6 各評価対象での特異な程度を利用

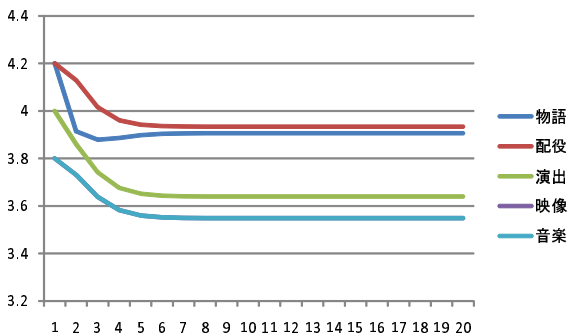


図 4 評価対象への一般的な評価 (評価ベクトル)

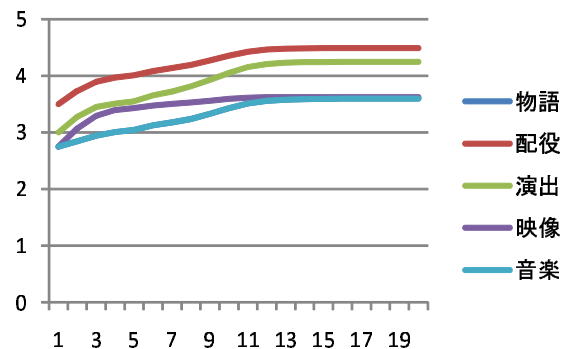


図 5 評価者の特異な程度を利用 (提案手法)

コアの値を示している．図中で赤く示されているものが、我々が加えた特異な評価者である．定量化した評価として、評価値、評価ベクトルのどちらを用いた場合でも、我々が加えた特異な評価者は、いずれも大きなスコアが与えられている．また、加えたレビュー以外で大きなスコアが与えられたレビューは、多くの人が高い評価をする中で低い評価をするというケースが多かった．

次に評価対象への一般的な評価の一例を図 3、図 4 に示す．図 3 は評価値を、図 4 は評価ベクトルをそれぞれ定量化した評価として用いた場合の結果である．縦軸が評価値の値、または評価ベクトルの各成分の値を表し、横軸は提案手法の繰り返し数を表す．この例では、高い評価をした人が特異とされ、その影響が更新を繰り返すことで抑えられていったと考えられる．

最後に、提案手法では、各評価対象での特異な程度を統合し

た評価者の特異な程度を用いた重み付き平均をとることで、一般的な評価を算出しているが、一方で、各評価対象での特異な程度に基づく重みを用いた重み付き平均をとることで一般的な評価を算出することも考えられる．評価ベクトルを用いた場合の両者の一例を図 5、図 6 に示す．提案手法である、評価者の特異な程度に基づく重み付けを行った場合である図 5 は、低い評価をした人の影響が抑えられ、一般的な評価が高くなっていることが分かる．一方で、各評価対象での特異な程度に基づく重み付けを行った場合である図 6 は、高い評価をした人の影響が抑えられ、一般的な評価が低くなっていることがわかる．これは、低い評価をした人の方が多数派であるために、高い評価をした人が特異とされてしまうためである．提案手法が用いる評価者の特異な程度は、評価対象を横断して得られる特異な程度であり、ある評価対象へのすべての評価の中で特異な評価であっても、その評価者が特異でなければ、一般的な評価に大きな影響を与えることができる．

## 5. おわりに

特異な評価者と評価対象への一般的な評価を特定するための手法として、集団評価の二部グラフモデルとグラフ上での相互強化型解析手法を提案した．評価者の特異な程度は、個々の評価の妥当性を判断する際に活用でき、評価対象への一般的な評価は、評価対象を理解する際に活用できる．また、評価を定量化する際には、いくつかの表現形式が考えられるが、提案手法はそれらの表現形式に依らず統一的に適用可能な手法となっている．

今回、すでに定量化された評価を用いて実験を行ったが、テキストの内容との関連について議論するには、テキストからの評価の定量化が必要であると考えられる．今後は、そのようなテキストから評価の抽出、定量化を行い、提案手法を適用することを検討したい．

謝辞 本研究の一部は、NICT 委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」及び日本学術振興会特別研究員奨励費 (21・3998) の助成を受けたものである．ここに記して謝意を表します．

## 文 献

- [1] J. M. Kleinberg , “Authoritative sources in a hyperlinked environment , ” Journal of the ACM(JACM) , 46(5) , pp. 604-632 , 1999.
- [2] V. Chandola, A. Banerjee, and V. Kumar, “ Anomaly detection: A survey, ” ACM Computing Surveys, vol. 41, no. 3, 2009.
- [3] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos, “Neighborhood formation and anomaly detection in bipartite graphs, ”In Proceedings of the Fifth IEEE International Conference on Data Mining, pp. 418-425, 2005.
- [4] X. Wang and I. Davidson, “Discovering contexts and contextual outliers using random walks in graphs, ”In Proceedings of the Ninth IEEE International Conference on Data Mining, pp. 1034-1039, 2009.