能動的推論としての確証的SNS探索の認知モデル化の試み
A Trial of Cognitive Modeling of Confirmatory SNS Exploration as Active Inference

福地 庸介
Yosuke Fukuchi
東京都立大学
Tokyo Metropolitan University
fukuchi@tmu.ac.jp

概要

本稿では、SNS 環境における情報探索行動を通じたユーザの信念形成過程、特に探索における確証バイアスを、能動的推論としてモデル化する。提案するモデルは、観測される投稿による信念更新量（認識的価値）と、既存の信念に適合する情報を観測することで得られる満足（実利的価値）とのバランスにより、探索行動の確証性を定量的に説明する。仮想 SNS を用いたユーザスタディでは、参加者の一部が自身の初期信念に一致する情報を選好し、結果として初期信念を維持・強化する傾向が確認された。さらに、モデルを仮想 SNS に適用したシミュレーションにより、信念の初期値の偏りと学習率が、確証バイアスを含めたユーザの行動を再現できる可能性が示唆された。

キーワード：SNS, ユーザモデル, 能動的推論, 自由エネルギー原理, 確証バイアス

1. 序論

SNS 等のインタラクティブなウェブメディアが普及するにつれて、メディアを通じたフェイクニュースやプロパガンダの伝搬が深刻な問題となっている。真贋入り混じったさまざまな情報が氾濫する中、陰謀論に代表される誤った信念をユーザがどのように形成・強化してしまうかを解明することは、中立なウェブメディアの設計や、健全なウェブメディア利用の促進のために重要な問題である。ウェブメディア探索におけるユーザの認知過程を再現する認知モデルを構築できれば、メディアのデザインやユーザのパーソナリティが探索行動や信念形成に与える影響をシミュレーションできるようになり、健全なウェブメディア構築への貢献が期待できる。

本研究では、メディア探索における確証バイアスに着目する。確証バイアスは、自らの信念に適合する情報ばかりを選択し、信念に反する情報を無視しようとする傾向のことである [4]。Tanaka et al. は、ファクトチェックサイトにおいて、参加者の約半数が自己の信念に反する情報を回避し、結果的に誤った信念が保持されたことを報告している [6]。このように確証バイアスは、ユーザの情報探索を歪めることで、誤信念の形成・強化を促進する可能性がある。

従来、確証バイアスをベイズ推論の枠組でモデル化する研究がなされている。Pilgrim et al. は、限定合理的なベイズ推論として確証バイアスをモデル化し、エージェントシミュレーションによってその妥当性を示した [5]。Chattoraj et al. は、既存の信念をもとにした能動的推論として確証バイアスをモデル化し、視覚探索タスクにおける人の振る舞いと整合することを示した [1]。内海らは、錯視画像の認識過程を、信念をもとにした確証的な視覚的注意の結果として定式化したベイジアンモデルを提案した [7]。このように、確証バイアスのモデル化におけるベイズ推論の有効性が示されてきたが、テキスト情報が中心となるウェブメディア探索への応用可能性は、明らかでない。

そこで本稿では、SNS 環境における確証的な探索行動を、ベイズ推論、特に自由エネルギー原理 [2] における能動的推論としてモデル化する。自由エネルギー原理では、新たな情報を獲得して信念を更新する**探索**と、既存の信念にもとづいて目的の達成のための行動を選択する**活用**のトレードオフを、探索によって新たに得られる情報量の期待値である**認識的価値**と、活用によって得られる満足を定量化する**実利的価値**のバランスが表現するとされる。本稿で提案するモデルは、ウェブメディアとのインタラクションで得られる観測をよく説明する信念の獲得を目的関数とするエージェントのモデルである。新たな情報の獲得によって得られる情報量（認識的価値）と、信念を確かめることで得られる満足（実利的価値）のバランスを取る過程としてウェブメディア探索が表現される。両者のバランスが崩れ、実利的価値の影響が認識的価値よりも大きくなる時、探索行動は確証的になる。

モデルの検証にあたり、まずニュースプラットフォームを模した仮想 SNS におけるユーザスタディを

図 1　ハッシュタグの選択

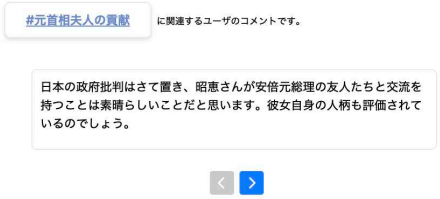

図 2　投稿の観測

行った。ここでユーザは、あるニュースの議論に関する賛否を回答した後、(i) ハッシュタグを選択し、(ii) 選択したハッシュタグに紐づいた投稿を観測する、という (i), (ii) のプロセスを 10 回繰り返し (図 1, 2)、最後に改めて賛否を回答した。結果、ユーザはそれぞれ多様な情報探索を行っており、その一部が確証的なハッシュタグ選択によって初期信念を強化した可能性が示唆された。次に、提案する認知モデルをこの仮想 SNS に適用して挙動をシミュレーションした。結果、信念の初期位置と信念更新の際の学習率が実利的価値と認識的価値のバランスに影響することで、ユーザスタディで見られたユーザの振る舞いを再現することが示唆された。本稿の内容は、[3] の内容に追加の検証を加えたものである。

2. 提案モデル

提案するモデルは、SNS 探索で観測される投稿をよく説明する信念の獲得を目的として行動するエージェントのモデルである。図 1, 2 に示した仮想 SNS において、ハッシュタグの選択を行動 a、選択されたハッシュタグに紐づいた投稿群の観測を o とする。また、信念 $q(x)$ を、命題 x が真であるとするエージェントの主観確率を表現する確率分布とする。a, o, x はいずれも、埋め込み空間におけるベクトルとして表現することにする。また、q を埋め込み空間における多変量正規分布として、その重心を g_q、共分散行列を V とする。エージェントは o が与えられた際、$q(o)$ の対数尤度を目的関数とする勾配法によって、$q(o)$ が増加する方向に q を更新する。

$$\mathbb{E}_{p(o|a)}[\ln q(o)]. \tag{1}$$

この時の学習率を α とする。

エージェントは認識的価値 V_{epist} と実利的価値 V_{prag} の和に基づき、a を確率的に選択する。

$$P(a) \propto \exp(V_{\text{epist}}(a) + V_{\text{prag}}(a)). \tag{2}$$

[3] では価値を最大化する行動を選択し続ける貪欲法を採用したが、本稿ではユーザの選択の確率的なばらつきを表現するため、上式を用いた。

V_{epist} は、a により得られる o_a によって、信念がどれだけ更新されるかを示す。

$$V_{\text{epist}}(a) = \mathbb{E}_s[\mathrm{KL}(q_{\text{new}}(s) \| q(s))]. \tag{3}$$

q_{new} は、o_a を受けて式 1 によって更新された後の信念である。o が既存の信念と異なるほど信念が大きく更新されるので、この項は反証行動を増やし確証行動を減らす作用がある。注目すべき点として、信念の更新量は α の影響を受け、α が小さいと V_{epist} の影響が V_{prag} に対して相対的に小さくなる。その結果、行動は確証的になる。

$V_{\text{prag}}(a)$ は q における o_a の対数尤度であり、信念更新の目的関数（式 1）そのものである。

$$V_{\text{prag}}(a) = \ln q(o_a). \tag{4}$$

この項は a によって得られると見込まれる o が既存の信念に適合するほど大きくなるので、エージェントの確証行動を促進する作用がある。

自由エネルギー原理では、エージェントは a によって期待される o_a を予測するとされる。しかし、本稿では、「ハッシュタグの内容がそれに紐づくコメントをよく代表している」、つまり、釣り見出しなどがなく、ハッシュタグの表現が投稿を適切に表現しているとユーザが想定していると仮定し、式 3 の計算では a を o_a にそのまま置き換えた。

3. 実験

3.1 ユーザスタディ

3.1.1 目的と方法

ユーザスタディの目的は 2 点である。(i) モデルシミュレーションの題材とする仮想 SNS においてユーザが実際に確証的探索を行うかを調べる。(ii) モデルを検証するためのデータを収集する。

実験は Yahoo!クラウドソーシングを通じて行い、100 名の参加を得た (男性 85 名、女性 14 名、無回答 1 名; 19-65 歳 ($M = 47.7, SD = 10.0$))。ユーザはまずニュース記事を読み、記事で取り上げられている議論（元総理夫人である安倍昭恵氏が、石破茂総理に先んじて、新たに大統領に就任したトランプ氏と私的に面会したこと）の賛否を 0 から 100 の 101 段階のスケールで回答した。意見の入力はブラウザのレンジス

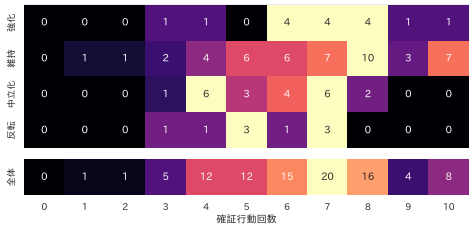

図 3　確証行動回数の分布

ライダーによって行われた。次に、ハッシュタグの選択（図 1）と投稿の観測（図 2）を 10 回繰り返した。ハッシュタグの選択では、賛成と反対の内容のものがそれぞれ 3 つずつ表示されるようにした。最後に、議論の賛否を改めて回答させた。初期の意見が 50 で、実験の前後で賛否が確証的であるか反証的であるか定義できない 6 名は除外し、初期の意見が肯定側だったユーザ 62 名と否定側だったユーザ 32 名のデータについて分析した。ハッシュタグとコメントは、GPT-4o[*1] によって生成した。

3.1.2　結果

タスクの前後でユーザの意見は多様に変化することがわかった。6 ポイント以上初期意見の側に変化したユーザを**強化**、意見の変化量が 5 ポイント以内のユーザを**維持**、6 ポイント以上初期意見と逆側に変化したユーザのうち中央値である 50 を超えていないものを**中立化**、超えたものを**反転**に分類したところ、強化、維持、中立化、反転に該当するユーザはそれぞれ 16、47、22、9 名いた。

図 3 に、ユーザが初期意見を支持するハッシュタグを選択した回数（確証行動回数）の分布を示す。確証行動回数の平均値は 6.38（SD=2.02）だった。確証行動回数が 2 回以下で初期信念に対して反証的選択を多く行ったユーザは 2 名のみであったが、反対に 8 回以上確証行動を選択したユーザは 28 名、9 回以上で 12 名いた。以上から、反証的な行動より確証的な行動を好むユーザの傾向が明らかになった。

図 3 には、意見変化の分類ごとの確証行動回数の分布も示している。意見が中立化/反転したユーザは、確証行動と反証行動をバランスよく行う傾向にあった。一方で、確証行動が多いユーザは、意見を維持/強化したグループに集中していた。維持のグループで確証行動が 9 回以上だった 10 名のうち 9 名は初期意見のスコアが 0、99、または 100 であり、それ以上信念を強化できない天井効果が示唆される。

以上の結果から、仮想 SNS においてユーザの一部が確証的な探索を行い、結果として初期の信念が維持/強化されたことが示唆された。

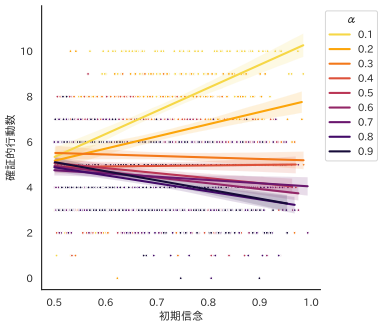

図 4　意見の初期値、α と確証行動回数の関係

3.2　シミュレーション

3.2.1　目的と方法

確証的 SNS 探索モデルを仮想 SNS に適用し、ユーザスタディで観察された振る舞いをモデルが再現することができるかを、乱数で決定した α, g によるシミュレーションで検証した。$\alpha = 0.1, 0.2, ..., 0.9$ について、それぞれ 100 回シミュレーションした。a, o, x の埋め込みベクトルとして、Transformer ベースの大規模言語モデル[*2]の最終層のベクトルに平均値プーリングを適用し、線形変換で 2 次元に変換したものを使った。線形変換には、クラウドワーカーが回答した投稿間の距離と、線形変換後のベクトル間の距離の相関を最大化するよう学習したものを用いた。意見スコアに対応するパラメータとして、賛否それぞれの投稿の埋め込みベクトルの重心から g_q までの距離の比を用いた。

3.2.2　予想

ユーザの行動、特に確証バイアスは、意見スコアの初期値と α によって再現できると予想した。具体的には、2. 章で述べたように α が小さいと V_{epist} の影響が小さくなることで確証行動が増え、その傾向は初期スコアの賛否への偏りが強いほど顕著だと予想した。

3.2.3　シミュレーション結果の傾向

図 4 に、シミュレーションでモデルが示した確証行動の回数と、意見スコアの初期値と α の関係を示す。賛否の対称性から、初期スコアが否定側の結果を反転させることで、賛否を同等に扱っている。予想通り、α が小さい時に確証行動が増加し、その傾向は初期スコアが偏っている時に強まった。結果は、提案モデルがユーザの確証バイアスを再現できることを示唆している。逆に、α を大きくすることで確証行動回数が減少することも予想通りであった。ただし、α を大きくする効果は $\alpha \geq 0.6$ 近辺で頭打ちとなった。この理由は、α が大きい場合、反証的な観測によって信念がそ

[*1] https://openai.com/index/hello-gpt-4o/

[*2] https://huggingface.co/llm-jp/llm-jp-3-13b-instruct

れまでと逆側の意見に大きく更新されるのが繰り返されることで、信念が振動するためである。以上から、確証バイアスを含めたユーザの多様な振る舞いをモデルが大局的に再現していると考えられる。

3.2.4 ユーザスタディとの比較

モデルがユーザ個人の SNS 探索をどの程度再現できるかを更に検証するため、シミュレーションによって推定されるモデルのパラメータとユーザスタディで得られた結果の比較を行った。具体的には、各ユーザについて、10 回の行動選択のうち 6 回以上一致したシミュレーション結果を抽出し、各サンプルの初期/最終意見スコアの中央値を推定値とした。そして、この推定に基づき、シミュレーション結果からユーザの意見スコアを復元できるか調べることで、モデルの妥当性を検証した。結果として用いたサンプル数は 2,998,105 であり、2,487,064 の行動系列パターンが得られた。

表 1 に、シミュレーションによって推定された意見スコアとユーザの回答の相関係数を示す。結果、行動系列のみをもとに、ユーザの初期および最終意見スコアをそれぞれ相関係数 0.349、0.498 で推定できた。さらに、初期スコアも所与とした場合（g と初期スコアの差 < 0.1 を制約条件に加えた場合）には、最終スコアとの相関係数が 0.767 に向上した。また、本研究で着目しているのが SNS 探索がユーザに与える影響であることから、意見スコアの変化量（最終スコアと初期スコアの差）を推定できるかも調べた。結果、意見スコアの変化量を相関係数 0.350 で推定できることがわかった。これらの結果は、提案モデルがユーザ個人の探索行動および信念変化を一定程度再現できていることを示す。

一方、本シミュレーションの限界も明らかになったと言える。初期スコアを所与とした時の相関の大幅な向上は、行動系列のみからユーザの各パラメータを推定することの限界を示している。また、探索行動の影響が直接的であると考えられる最終スコアと比べて、初期スコアを行動系列のみから逆算することは、より困難であった。さらに、シミュレーションにおけるパラメータの探索空間が広く、ユーザのパラメータを特定するだけのサンプル数を得られていない可能性がある。行動系列との一致度の基準を 1 引き上げると、利用可能なサンプル数は概して 1/6 になる。そのため、今回は 6 回以上という比較的緩い基準を採用した。以上の限界を克服するため、今後はユーザの学習率とパーソナリティ特性との関係を調査することで、シミュレーションの探索空間を絞り込み、個人レベルの SNS 探索を効率的に再現することを可能にしたい。

表 1　ユーザの回答とシミュレーション結果の相関

所与条件	予測対象	相関係数
行動系列	初期スコア	0.349
行動系列	最終スコア	0.498
行動系列, 初期スコア	最終スコア	0.767
行動系列, 初期スコア	信念変化量	0.350

4. 結論

本稿では、SNS 探索を通じたユーザの信念形成を能動的推論としてモデル化した。ユーザスタディとシミュレーションによる検証の結果、提案モデルがユーザの振る舞いを一定程度再現できることがわかった。今後は、ユーザの学習率とパーソナリティ特性との関係を調査することで、SNS 探索がユーザに与える影響をより精緻に予測し、健全な SNS プラットフォーム設計に向けた議論に繋げることを目指す。

謝辞 本研究は、公益財団法人 電気通信普及財団 2024 年度研究調査助成の支援を受けた。

References

[1] Ankani Chattoraj et al. “A confirmation bias due to approximate active inference”. In: *Annual Meeting of the Cognitive Science Society*. 2021.

[2] Karl Friston. “The free-energy principle: a unified brain theory?” In: *Nature reviews neuroscience* 11.2 (2010), pp. 127–138.

[3] Yosuke Fukuchi. “A Bayesian Model of Confirmatory Exploration in Text-based Web Media”. In: *Annual Meeting of the Cognitive Science Society*. 2025 (in press).

[4] Raymond Nickerson. “Confirmation Bias: A Ubiquitous Phenomenon in Many Guises”. In: *Review of General Psychology* 2 (June 1998), pp. 175–220.

[5] Charlie Pilgrim et al. “Confirmation bias emerges from an approximation to Bayesian reasoning”. In: *Cognition* 245 (2024), p. 105693.

[6] Yuko Tanaka et al. “Who Does Not Benefit from Fact-checking Websites? A Psychological Characteristic Predicts the Selective Avoidance of Clicking Uncongenial Facts”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. 2023.

[7] 内海 佑麻 et al. “曖昧性解消における視覚的注意へのトップダウン介入”. In: **人工知能学会全国大会論文集** JSAI2021 (2021), 1H2GS1a04–1H2GS1a04.