ウェブメディア探索を通じた信念形成の認知モデル化の検討

Towards Cognitive Modeling of User Belief Formation through Web Media Exploration

福地庸介 [1*]

[1] 東京都立大学
[1] Tokyo Metropolitan University

Abstract: SNS 等のウェブメディアが普及するのにしたがって、フェイクニュースやプロパガンダを通じた誤信念の形成が深刻な問題となっている。本研究では、ウェブメディア探索の主体としてのユーザの認知に着目し、ウェブメディアとのインタラクションを通じて特定の信念を形成するまでの認知過程を計算モデル化することを検討する。具体的には、ユーザの信念形成と探索行動の関係を自由エネルギー原理に基づいて定式化することで、ウェブメディア探索における確証バイアスを計算論的に再現することを試みる。さらに、仮想 SNS におけるユーザ実験の結果とシミュレーションの結果を比較することで、認知モデルの検証を行う。検証の結果、提案モデルで表現される信念の初期値と信念更新の学習率が、ユーザの多様な探索行動を再現する可能性が示唆された。

1 はじめに

本研究の最終的な目標は、SNS 環境におけるユーザの探索行動や信念変化を再現し、実際の言語コンテンツをもとに誤信念形成のプロセスをシミュレーション可能な認知モデルを構築することである。ウェブには膨大な情報が存在し、効果的に活用できれば多種多様な知見を得ることが可能である。特に、SNS の普及によって、ユーザが異なる立場や意見にアクセスできる環境はかなり整ってきたと言える。しかし、現実には、ユーザがウェブの偏った情報をもとに誤信念を形成してしまう、という負の側面に注目が集まっている。ユーザの行動やその背後にある認知プロセスの計算モデルが実現すれば、ウェブメディアとユーザの相互作用をシミュレーションすることができるようになり、メディア設計がユーザの信念形成や行動に与える影響の予測や、ユーザを誤信念に傾倒させにくくする健全なウェブメディアの設計に貢献することが期待できる。

ユーザのウェブ探索を歪める要因の 1 つとして、確証バイアスがある。確証バイアスとは、自分の信念や先入観に合致する情報を優先的に探索し、信念に反する情報を避けたり無視したりする傾向を指す [1]。Tanaka et al. は、ファクトチェックサイトを題材とした実験で、参加者の約半数が自身の信念に反する情報を積極的に取得した一方で、残りのユーザは信念に反する情報を回避し、結果として誤った信念を保持する傾向があることを報告している [2]。つまり、確証バイアスは、ユーザが自身の信念を強化する情報を過剰に取得し、逆に信念に反する情報を取得する機会を逃す原因となりえる。そこで本稿も、ウェブ探索におけるユーザの確証バイアスと信念形成の関係に着目する。

これまで、確証バイアスをベイズ推論の枠組みに基づいてモデル化するさまざまな試みが行われている。例えば、Pilgrim et al. は、限定された認知資源の下での近似的なベイズ推論として確証バイアスを説明できることを、シミュレーションによって示している [3]。また、Chattoraj et al. は、視覚的探索において、既存の信念をもとにした能動的推論が確証バイアスを引き起こすことを指摘している [4]。内海らは、信念が確証的な視覚的注意を引き起こすことで錯視画像の認識を行う過程を、ベイズ推論としてモデル化している [5]。

以上のように、従来研究では、確証バイアスの認知的な基盤のモデル化において、ベイズ推論による定式化の有効性が示されている。しかし、SNS のような言語情報を中心としたウェブメディア探索への適用は行われていない。

そこで本稿では、ベイズ推論、特に自由エネルギー原理 [6] をベースに、SNS 環境における確証的な探索行動のモデル化を試みる。自由エネルギー原理は、エージェントが観測と信念のズレを最小化することで、不確実性を減らしながら行動を選択するメカニズムを説明する理論である。提案する認知モデルでは、ユーザが自己の信念に適合した情報を得ようとする確証的行動と、自己の信念を広げるために新しい情報を得ようとする行動の間で、ユーザがどのような行動を選択す

*連絡先：東京都立大学システムデザイン学部情報科学科
〒 191-0065 東京都日野市旭が丘 6 丁目 6
E-mail: fukuchi@tmu.ac.jp

るかという過程を、自由エネルギー原理において即時的な満足を表現する「実利的価値」と、新しい知識を得ることに関する「認識的価値」の関係に対応づけることで定式化する。また、信念の初期分布や学習率といったパラメータを変化させることで、ユーザの多様なウェブ探索行動を再現できることを期待している。さらに、認知モデルを言語埋め込み空間に実装することで、言語情報に基づくウェブ探索行動をシミュレーションすることを目指す。

本稿ではまず、シミュレーションの題材として用いた仮想 SNS と、これを用いたウェブ探索のユーザ実験について説明する。この中で、仮想 SNS とインタラクションしたユーザの一部が、確証的選択によって既存の信念を強化する傾向を示したことを報告する。次に、自由エネルギー原理をベースにした認知モデルの定式化と実装を示す。最後に、シミュレーションの結果とユーザ実験の結果を比較し、提案モデルの妥当性を議論する。

2 仮想 SNS 探索実験

2.1 目的

本実験では、著者が構築した仮想 SNS 環境におけるユーザの探索行動と信念形成のプロセスを観察し、確証バイアスとの関係を分析する。具体的には、(i) 本仮想 SNS においてユーザの確証バイアスは確認されるか、(ii) ユーザの行動が、初期意見を強化する、反転させる、または中立化させるなど、意見形成のパターンにどのような違いをもたらすのか、を検証する。

2.2 仮想 SNS 環境の設計

実験のために、ニュースポータルサイトを模倣した仮想 SNS を用意した。この仮想 SNS は、以下の 3 つの要素からなる。(a) **ニュース記事：** 実験開始時に、ユーザに対して 1 つのニュース記事を提示する（図 1）。(b) **ハッシュタグ：** 各ハッシュタグは、ニュースに対して投稿された意見の要約を表現する。また、ユーザが選択できるリンクとして機能する（図 2）。(c) **投稿:** ハッシュタグに紐づけられた、ニュース記事に関する賛成、または反対意見が投稿の形で表示される（図 3）。ニュース記事は Yahoo!ニュース[1]から、記事に寄せられたユーザコメントが多かったものを選定した[2]。この記事は、安倍晋三元総理の夫人である昭恵氏とトランプ新大統領の面会の是非に関する議論を題材としたものである。当該記事に実際に投稿されていたユーザコメントは、面会の是非以外にも多岐に渡る内容を含んでおり、実験内容の統制を支障すると懸念されたため、実験におけるコメントは、面会の是非に内容を絞らせるプロンプトをもとに GPT-4o[3]を用いて生成しなおしたものを用いた。ハッシュタグの生成にも GPT-4o を用いた。

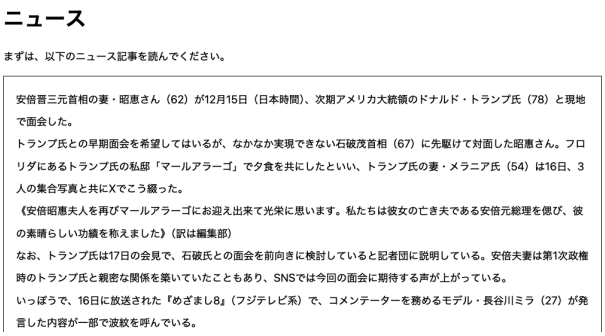

図 1: ニュース記事

図 2: ハッシュタグの選択

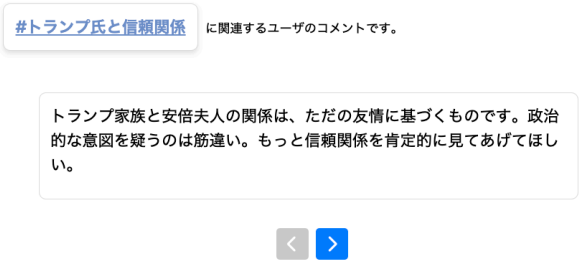

図 3: 投稿の観測

2.3 実験手順

実験の手順は以下の通りである。(1) ユーザは、ニュース記事を 1 つ閲覧し、その内容に対する賛否を 1-100 のスコアで評価する（初期意見の記録）。(2) 投稿に関する 6 つのハッシュタグが提示される。ハッシュタグは、賛成の意見を代表するものと反対の意見を代表するものから、毎回 3 つずつ選ばれる。ユーザはこれらの中から 1 つを選択し、対応する投稿を 8 件閲覧する。(3) 投稿を閲覧した後、次のラウンドに進み、再びハッ

[1] https://news.yahoo.co.jp/

[2] “「首傾げてる国民は多い」人気モデル トランプ氏と面会した安倍昭恵さんへの“苦言”に賛否…SNS には“荒らしコメント”続出の事態に”（https://news.yahoo.co.jp/articles/020ba4acb21022773eaffc9b34b19cc8f3cc65b3）2024/12/25 アクセス

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

シュタグを選択する。このプロセスを 10 ラウンド繰り返す。(4) 最終ラウンド終了後、ニュース記事に対する最終的な賛否スコアを再び記録する（最終意見の記録）。実験を統制するため、実験参加者間で各ラウンド同じハッシュタグを用いた。ただし、ハッシュタグの表示順をランダム化することで、選択肢が特定の順序に偏らないように設計した。ラウンド間で、ハッシュタグや投稿の内容には重複がないようにした。実験は Yahoo! クラウドソーシングを通じて実施し、100 人の参加を得た（男性 85 人、女性 14 人、無回答 1 人；19-65 歳（M=47.7, SD=10.0））。

2.3.1 評価項目

以下の 2 つの項目を評価に用いた。**確証的選択回数：** ユーザの確証バイアスの程度を定量化するため、初期意見と一致する側のハッシュタグを選択した回数を測定した。本研究では、10 回中 8 回以上の確証的選択を「確証的探索」、2 回以下を「反証的探索」と定義した。**意見変化タイプ：** 初期スコアと、最終スコアとの差分を用いて、ユーザの意見形成のパターンを分類した。分類は以下のように行った：(a) 初期意見を強化した (b) 維持した (c) 中立に近づけた (d) 賛否を反転させた (e) その他。ここで、(b) 維持した は、最終スコアが初期スコアとの差の大きさが 5 ポイント以内の場合とした。

2.3.2 結果

図 4 に、確証的選択回数の分布を示す。結果、28 名が確証的探索を、3 名が反証的探索を、69 名が中立的な探索を行っており、本仮想 SNS において一部のユーザが実際に確証的探索を行うことがわかった。

図 5 に、初期意見から最終意見への変化を分類した結果を示す。数が多い順に、50 名が初期意見を維持、22 名が中立に近づけ、17 名が強化、10 名が賛否を反転させた。残りの 1 名は、初期意見が中立（50 ポイント）であったところを最終意見で反対に転じた。

図 6 は、意見変化タイプごとの確証的選択回数を示している。主な傾向として、初期意見を強化したユーザは、確証的探索をする傾向が強かった。一方、意見を中立に近づけたユーザや反対の意見に転じたユーザは、賛成のハッシュタグと反対のハッシュタグをバランスよく選択する傾向が見られた。

以上の結果は、本仮想 SNS において一部のユーザの確証的探索を引き起こし、それによって初期意見を強化したことを示唆している。

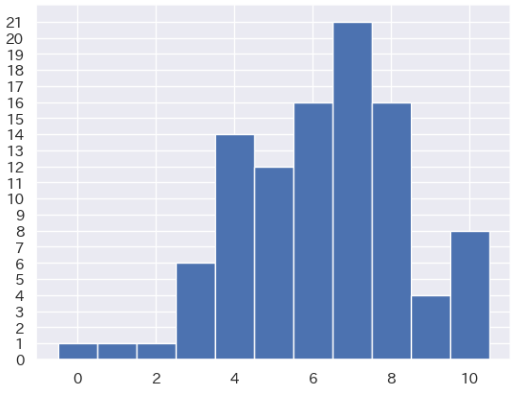

図 4: 確証的選択回数

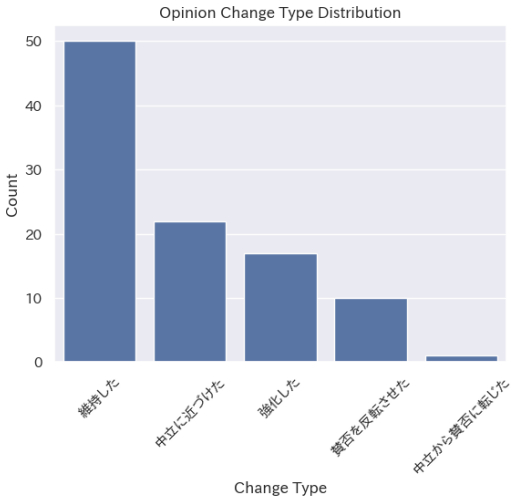

図 5: 意見変化タイプの分布

3 SNS 探索認知モデル

3.1 概要

本稿で提案する認知モデルは、仮想 SNS 実験で観察されたユーザの探索行動や信念形成プロセスを再現することを目指す。提案モデルのベースとなる自由エネルギー原理は、エージェントが観測データと内部モデル（信念）のズレを最小化することで、不確実性を減少させつつ行動を選択する仕組みを説明する枠組みである。自由エネルギー原理では、人は (i) 外界の観測、(ii) 観測情報をもとにした信念の形成、(iii) 信念をもとにした外界の状態の予測、(iv) 予測を確かめる行動の選択、(i') 行動によって変化した外界の観測…というループを、自由エネルギーと呼ばれるコスト関数を最小化することで辿るとされる。

モデルでは、ユーザの探索行動を自由エネルギー原理における 2 つの価値のトレードオフとして定式化する。(1) 実利的価値：現在の信念に適合する情報に対する価値。(2) 認識的価値：新しい情報を取得すること

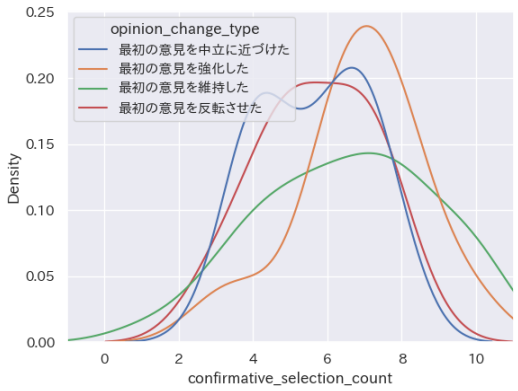

図 6: 意見変化タイプごとの確証的選択回数

で信念を更新し、不確実性を減少させることに対する価値。

3.2 定式化

提案する認知モデルは、ユーザの探索行動と信念形成を、言語埋め込み空間上で自由エネルギー原理に基づいて記述する。

3.2.1 信念の表現

ユーザの信念は、言語埋め込み空間内のベクトル $x \in \mathbb{R}^d$ を中心とする確率分布 $q(x)$ として表現される。ここで、x は賛成意見や反対意見といった命題を表現する埋め込みベクトル、$q(x)$ はエージェントが x をどの程度「正しい」と考えるかを表す確率分布、o はユーザが観測するデータ（例: ハッシュタグ、投稿）であり、同じ言語埋め込み空間内のベクトルとして表現される。本稿では $q(x)$ を多次元正規分布としてモデル化する。

信念の更新は、観測データ o に基づく自由エネルギー最小化として定式化される：

$$F(q) = -\mathbb{E}_{q(x)}[\ln p(o \mid x)],$$

実装では、確率的勾配降下法を用いて $F(q)$ を最小化することで、観測データ o に基づく信念 $q(x)$ を更新する。ここで、α を学習率とする。

3.2.2 価値関数と行動選択

本モデルでは、ユーザの探索行動を、自由エネルギー原理に基づく以下の 2 つの価値のトレードオフとして定式化する：**実利的価値（Pragmatic Value）**：現在の信念 $q(x)$ に基づいて観測データ o をどれだけうまく説明できるかを表す。**認識的価値（Epistemic Value）**：新しい観測データが信念 $q(x)$ に与える変化、すなわち不確実性の削減量を表す。

行動 a（ここではリンクの選択）に対する総合価値 $V(a)$ は、以下のように定義される：

$$V(a) = V_{\text{prag}}(a) + V_{\text{epist}}(a),$$

実利的価値 $V_{\text{prag}}(a)$ は、信念 $q(x)$ に基づいて観測データ o_a（リンク a をクリックして得られるデータ）を説明する尤度に対応し、次式で表される：

$$V_{\text{prag}}(a) = \mathbb{E}_{q(x)}[\ln p(o_a \mid x)].$$

実装では、ハッシュタグ自体の埋め込みベクトルを o_a とした。

認識的価値 $V_{\text{epist}}(a)$ は、観測データ o_a によって信念分布 $q(x)$ がどれだけ変化するかを定量化する。これは、現在の信念 $q(x)$ と観測後の更新信念 $q_{\text{new}}(x)$ の間の KL ダイバージェンスを用いて次式で定義される：

$$V_{\text{epist}}(a) = \text{KL}(q_{\text{new}}(x) \| q(x)),$$

ここで、$q_{\text{new}}(x)$ は観測 o_a を用いて更新される新たな信念分布である。$q_{\text{new}}(x)$ への更新量は、学習率 α の影響を受ける。

エージェントは、リンク a に対する総合価値 $V(a)$ が最大となるリンクを選択する：

$$a^* = \arg\max_a V(a).$$

3.3 言語埋め込みベクトル

本モデルで利用する言語埋め込みベクトルを準備するために、まず国立情報学研究所大規模言語モデル研究開発センターが公開する大規模言語モデル（LLM）である llm-jp-3-13b-instruct[4] を用いた。具体的には、LLM の最終層の埋め込みベクトルを、各トークンに最大値プーリングを適用することで 1 つのベクトルにまとめ、さらに 2 次元ベクトルに線形変換することで得た。この線形変換の学習のために、クラウドソーシングによって投稿の対の距離を調査し、3600 対に関して類似度のラベルを得た。そして、埋め込み空間内の 2 ベクトルの距離とクラウドワーカーが回答した距離（1-類似度）の相関を最大化するよう、線形変換を学習した。線形変換後のベクトル間の距離は、クラウドワーカーから取得した距離と高い相関を示した（$r = 0.80$）。

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

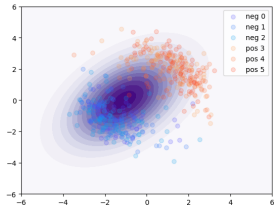

図 7: ケース 3 の初期信念

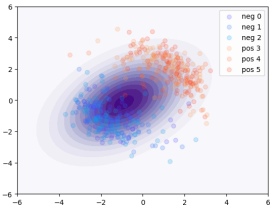

図 8: ケース 3 の最終的な信念

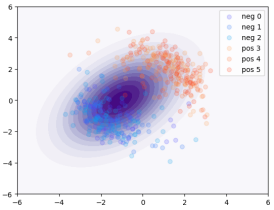

図 9: ケース 4 の初期信念

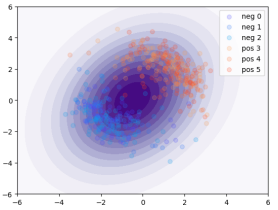

図 10: ケース 4 の最終的な信念

4　モデルによるシミュレーション

4.1　目的

本シミュレーションでは、特に初期信念の偏りと学習率に着目し、提案モデルの挙動を確認する。また、ユーザ実験の結果とシミュレーションの結果を比較することで、提案モデルの妥当性を検証する。

4.2　シミュレーションの設定

シミュレーションでは、エージェントの信念を潜在空間内の正規分布としてモデル化し、観測データに基づいて信念を更新しつつ行動を選択させた。主な設定は以下の通りである。**初期信念の重心**: 初期信念の重心は、ネガティブな意見の重心 ($g_{\mathtt{neg}}$) とポジティブな意見の重心 ($g_{\mathtt{pos}}$) の間で調整される。調整には `w=0` でニュートラルな信念、`w=1` で $g_{\mathtt{neg}}$ と一致するような重み `w` を用いた。**学習率**: 観測データに基づく信念更新の速度を制御するパラメータである。シミュレーションでは複数の値 ($\alpha = 0.1$ および $\alpha = 0.3$) を設定した。

以下の条件でシミュレーションを実施した：**ケース 1**: 初期信念がニュートラル (`w`=0)、学習率 $\alpha = 0.1$。**ケース 2**: 初期信念がネガティブに強く偏り (`w`=0.5)、学習率 $\alpha = 0.1$。**ケース 3**: 初期信念がネガティブにやや偏り (`w`=0.25)、学習率 $\alpha = 0.1$。**ケース 4**: 初期信念がネガティブにやや偏り (`w`=0.25)、学習率 $\alpha = 0.3$。

4.3　シミュレーション結果と考察

表 1: シミュレーション条件と確証的行動の回数

ケース	w	学習率 α	確証的行動の回数
1	0	0.1	4
2	0.5	0.1	9
3	0.25	0.1	8
4	0.25	0.3	6

表 1 に、各ケースにおける確証的行動の回数を示す。ケース 1 では初期信念がニュートラルなため、確証的行動は少なく探索的行動が増加した。確証的行動回数は 4 回であった。ケース 2 では初期信念が強く偏っているため、信念に基づく実利的価値が優先され、確証的行動が支配的であった。確証的行動回数が 9 回であった。ケース 3 では、初期の偏りがやや弱いが低学習率のため探索的行動は限定的となり、確証的行動回数は 8 回であった。また、確証的行動に伴って、信念が強化された（図 7, 8）。ケース 4 では学習率が高いため信念更新が積極的に行われ、不確実性の削減が優先され探索的行動が増加した。確証的行動回数が 6 回であった。また、4 回の探索的行動により、信念分布が中立化した（図 9, 10）。

シミュレーション結果から、初期信念が偏るほど確証的行動が増加することが確認された。これは初期信念が実利的価値を強化し、既存の信念に一致する情報を選択する行動につながるためである。また、学習率が高い場合、観測データの影響が大きくなり信念が動的に更新されるため、確証的行動が減少し探索的行動が増加することがわかった。以上の結果は、ユーザ実験におけるユーザの多様な振る舞いと整合的で、提案モデルがユーザ行動を再現する可能性を示唆している。

5　おわりに

本稿では、実際の言語コンテンツをもとにウェブ探索における信念形成のプロセスをシミュレーションすることを目指し、ウェブ探索の認知モデルを提案した。仮想 SNS におけるユーザ実験の結果とシミュレーションの結果を比較から、提案モデルがユーザの確証的探索行動を再現する可能性が示され、特に、提案モデルで表現される信念の初期値と信念更新の学習率が、ユーザの多様な探索行動を再現する可能性が示唆された。

参考文献

[1] Raymond Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of Gen-*

eral Psychology, 2:175–220, 06 1998.

[2] Yuko Tanaka, Miwa Inuzuka, Hiromi Arai, Yoichi Takahashi, Minao Kukita, and Kentaro Inui. 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, New York, NY, USA, 2023. Association for Computing Machinery.

[3] Charlie Pilgrim, Adam Sanborn, Eugene Malthouse, and Thomas T. Hills. Confirmation bias emerges from an approximation to bayesian reasoning. *Cognition*, 245:105693, 2024.

[4] Ankani Chattoraj, Sabyasachi Shivkumar, Yong Soo Ra, and Ralf M. Häfner. A confirmation bias due to approximate active inference. In *Annual Meeting of the Cognitive Science Society*, 2021.

[5] 内海 佑麻, 福地 庸介, 木本 充彦, and 今井 倫太. 曖昧性解消における視覚的注意へのトップダウン介入. **人工知能学会全国大会論文集**, JSAI2021:1H2GS1a04–1H2GS1a04, 2021.

[6] Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138, 2010.