J-034

LLM による Web 議論への適切なタイミングでの関連情報推薦エージェントの開発
LLM-based Agent for Recommending Information Related to Web Discussions at Appropriate Timing

櫻井 崇貴† 白松 俊† 木下 良輔†
Takayoshi Sakurai Shun Shiramatsu Ryosuke Kinoshita

1. はじめに

近年，インターネット上において情報交換や議論を行うことが可能となった．特に Web 議論は，多様な意見が集まり，広範囲にわたるトピックについての知識と理解を深める重要な手段となっている．しかし，Web 議論では，議論内容に精通している人も，全く背景知識を持っていない人も参加可能であるため，持っている情報量や議論の理解度の差によって発言しづらい人が存在することが考えられる．議論参加者が満足に発言できないことは議論全体の合意形成の妨げとなってしまうため，参加者の発言を促し，発言量の差を改善できるような仕組みが必要である．そこで LLM(大規模言語モデル)を用いて議論に関連した情報を議論参加者に推薦することは，参加者の内容理解や議論進行に役立つと考えた．しかし，LLM は通常，1 回の入力に対して 1 回の出力を行うため，議論参加者の発言に対して毎回応答をしていたら情報量が多く，議論の妨げになってしまうことが考えられる．

木下らの以前の研究[1]では，Web 上の議論において，議論参加者が発言しやすくなるような情報を自動的に提供する関連情報推薦エージェントが作成された．このエージェントの投稿は時間間隔で行われており，情報推薦のタイミングの最適化については検討されていなかった．

そのため我々は，LLM が応答すべきタイミングを推定し，そのタイミングで情報推薦を行うシステムを開発することにより，Web 議論における LLM による関連情報推薦エージェントの有用性を高めることができると考えた．

本研究の最終的な目的は，Web 議論において関連情報を推薦することにより，議論参加者の発言量の差を改善することで，合意形成を支援することである．そのためには，議論参加者が必要とする，適切なタイミングで情報推薦を行う必要がある．

本研究では，議論参加者の投稿から，いつ情報推薦を行うべきかを判断することに焦点を当て，情報推薦を行うタイミングを推定するシステムを開発した．そして，LLM による情報推薦や，情報のファクトチェックなどを行う木下ら[2]との共同研究を通じて，LLM による Web 議論への適切なタイミングでの関連情報推薦エージェントを開発し，推薦する情報の有用性についても着目して研究を行った．

2. 関連研究

先に述べた木下らの以前の研究[1]では Web 上の議論において，議論参加者が発言しやすくなるような情報を自動的に提供する関連情報推薦エージェントが作成された．このエージェントは，議論の理解度や情報量の差を埋めることによって，議論の進行と合意形成を促進することを目的としている．しかし，エージェントが情報を提供するタイミングの最適化については検討されていなかった．具体的には，情報推薦の間隔は 3 分毎と設定されており，議論の流れや，参加者の反応の基づいたものではなかった．また，ファクトチェックを行っていないことから，推薦情報が参加者を混乱させることがあった．

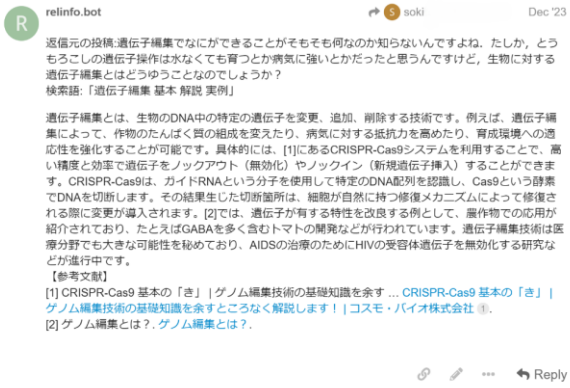

図 1　システムの動作例

実環境でロボットが複数人と音声対話する場面を想定し，ロボットの応答義務を推定した研究[3]がある．この研究では，ロボットへ向けたユーザからの発話だけでなく，ユーザ同士の会話や独り言，周辺雑音に対して応答義務を推定している．推定時には，ユーザの身体の動きや顔の動きといったユーザの状態に着目している．これは多人数での対話を対象としているところは同じだが，テキストベースではないため，本研究とは異なると言える．

3. システムの概要

我々はディスカッションプラットフォームである Discourse 上で動作する LLM が適切なタイミングで情報推薦を行うシステムを開発した．このシステムは以下の手順に従って関連情報を推薦する．

1. 議論データの取得
2. 情報推薦のタイミングの推定
3. 検索語の生成と関連情報のスクレイピング
4. ファクトチェック
5. Discourse へ関連情報の投稿

† 名古屋工業大学　Nagoya Institute of Technology

本研究で焦点を当てた情報推薦のタイミングの推定方法の詳細については 3 章にて詳しく説明する．

図 1 は開発したシステムが情報推薦を行っている例である．どの投稿に対する返信なのか，情報の検索に用いた検索語，LLM によって生成された意見，参考文献が記載されている．

4. 提案手法

この章では我々の提案手法の詳細について述べる．

4.1 情報提示要求の抽出

4.1.1 情報提示要求

我々は，LLM の 1 つである OpenAI の GPT-4 Turbo (gpt-4-1106-preview) を用い，情報推薦のタイミングを推定することにした．まずは，議論参加者の投稿から情報提示要求が含まれるか否かを判断させる．ここで情報提示要求とは，発言から読み取れる発言者が必要としている情報の要求のことをいう．発言に情報提示要求が含まれるときはその要求に対して，「情報の必要性」と「検索可能性」を 0〜100 の値で推定させる．このようにして得られた要求や推定値などを「情報提示要求リスト」に加える．このリストは JSON 形式で，各要素は表 1 に示すパラメータを持つ．

表 1　情報提示要求のパラメータ

パラメータ	説明
exractionReason	情報提示要求を抽出した理由
request	情報提示要求
name	発言者名
id	投稿 id
infoNecessityReason	情報の必要性の推定の理由
infoNecessityScore	情報の必要性
searchabilityReason	検索可能性の推定の理由
searchabilityScore	検索可能性
recommendAction	情報推薦するか否か

4.1.2 情報の必要性と検索可能性

各情報提示要求は 2 つの尺度，「情報の必要性」と「検索可能性」によって測られる．以下に GPT-4 Turbo に与えた具体的な評価基準を示す．

- **情報の必要性**：「意見の分裂」，「対話の停滞」，「感情の高まり」，「質問の有無」などの要素を組み合わせて 0〜100 で推定して下さい．
  - **0**：人間参加者同士での対話や情報交換がスムーズに行われていて，情報の提示が不要．
  - **100**：明確な質問がある，対話が停滞している，または参加者間での意見が分裂しているなど，情報の提示が必要．

- **検索可能性**：話題や問題の内容，キーワード，背景情報などを分析して 0〜100 で推定して下さい．
  - **0**：話題や問題が主観的，抽象的，または特定の内部情報に基づくもので，Web やデータベースでの検索が難しい
  - **100**：話題や問題が具体的，客観的で一般的な知識や情報に関連しているため，Web やデータベースでの検索が可能

例えば，ある議論参加者に何かわからないことがあり，疑問文を含む発言をした場合，その発言は情報提示要求を含むとみなされ，要求が抽出される．しかし，「あなたはどう思いますか？」といったような，要求がある他の参加者への個人的な質問である場合は LLM を用いたシステムが介入しても支援をすることが難しい．このような状況で情報推薦を行ってしまうと議論の妨げになってしまう可能性がある．そのため，上で示した 2 つの尺度を推定することで適切なタイミングを推定できるのではないかと考えた．

4.1.3 情報推薦のタイミングの決定

表 1 の「recommendAction」の値は True または False で表され，情報推薦するか否かを表す．これは上で述べた「情報の必要性」と「検索可能性」の値を基に GPT-4 Turbo が判断するようになっている．初めはそれぞれの値に閾値を設け，共に上回った情報提示要求を含む投稿に対して情報推薦を行うようにしていた．しかし，動作確認時に，人間が適切な閾値を決定することの難しさが判明したため，LLM に判断を任せた．

4.2 プロンプトの設計

GPT-4 Turbo などの LLM を用いる際はプロンプトの設計が重要である．ここではプロンプト設計時の工夫点や用いた手法などを説明する．

4.2.1 情報提示要求リストの更新

情報提示要求の「情報の必要性」と「検索可能性」の推定値などは，議論参加者の発言を受ける度に変化することが考えられる．そのため，プロンプトには議論参加者による最新の発言以前では未解決であった情報提示要求のリストを含み，最新の発言を受けてこれらの値を更新するタスクも同時に行わせるようにした．例えば，ある議論参加者からの疑問に対し，他の参加者が疑問に答えたとすると，「情報の必要性」の値は小さくなる．反対に他の参加者も答えられず，「わからない」などの発言をした場合はその値は大きくなる．このように発言を受ける度に情報提示要求リストを更新していくことで，発言 1 つのみに着目するのではなく，議論全体の流れを考慮した情報推薦のタイミング推定が可能となる．

また，「recommendAction」の値が True だった情報提示要求が，最新の発言を受けて False に変更されたとき，この要求はリストから削除されるようにした．さらに，False が 5 回続いた要求に対しては議論が流れ，情報推薦が必要なくなったと判断し，リストから削除されるようにした．このように，情報推薦が行われない要求が溜まることで情報提示要求リストが多くなり，プロンプトの文字数が増加しすぎることを防いだ．

4.2.2 プロンプトの手法

本研究では，プロンプト設計時に様々な手法を用いて出力の精度を高めることを試みた．ここでは用いた 3 つの手法について説明する．

1 つ目は Step-back Prompting[4]である．これは，LLM の推論能力を向上させる手法の 1 つである．複雑なタスクを行う際，人間が時々行うように，一歩後退し，前提に関す

る抽象的な質問を最初に提示したうえで，その前提に基づいて回答を生成する．これにより，効率的かつ正確な推論が可能になる．

本研究では情報提示要求の抽出や，リストの更新といったタスクを行う前に着目すべき要素について質問をし，その質問と回答の一部をプロンプトに含めることでシステムの推論能力の向上を図った．先に述べた，情報提示要求の尺度である「支援の必要性」と「検索可能性」はこの質問と回答から得られたものである．

2 つ目は Chain-of-Thought Prompting [5] である．これも LLM の推論能力を向上させる手法の 1 つである．問題に対する答えのみを出力するのではなく，答えに至るまでの中間推論ステップも出力させることで，推論能力が向上する．本研究では，表 1 のように「情報提示要求」や「情報の必要性」，「検索可能性」のみを出力させるのではなく，「情報提示要求」を抽出した理由や，2 つの尺度の値を推定した理由も出力することでシステムの推論能力の向上を図った．

3 つ目は Few-shot Prompting [6] である．これは少数の例を与えることでタスクを効果的に解決できるようにする手法である．本研究では議論参加者の投稿と，それに対する情報提示要求の抽出，情報提示要求リストの更新の例をプロンプトに含めることでモデルの理解を高めることを図った．

4.3 ファクトチェック

情報推薦の適切なタイミングを推定し，情報推薦をする情報提示要求を決定したら，必要な情報を web 上から探してくる必要がある．しかし，web サイトの情報は誤情報や古い情報を含んでいる可能性がある．また，本研究で開発したシステムは GPT-4 Turbo により web サイトの情報をまとめ，情報推薦を行うが，その発言にもハルシネーションを含む可能性がある．そこで木下ら[2]は web サイトの情報とシステムの発言の両方に対してファクトチェックを行い，信頼度が閾値を超えるまで発言を再生成させるようにした．これにより，根拠に基づいた情報推薦ができるようになると考えられる．

5. 実験

5.1 実験方法

本研究では 10 名に対して議論実験を行った．この 10 名を 5 人毎にグループ 1 とグループ 2 に分けた．それぞれのグループで今回開発した情報推薦のタイミング推定機能を含む関連情報推薦エージェントであるシステム A と，ベースライン手法の時間間隔で関連情報推薦エージェントが動作するシステム B を 1 回ずつ使用し，2 回の議論を行った．グループ 1 は初めにシステム A を使用し，グループ 2 は初めにシステム B を使用した． 2 つの議論のトピックは以下に示す．

- 遺伝子編集と倫理問題：遺伝子編集が可能にする医療上の利点と，それに伴う倫理的な問題はどのように調整すべきか？
- 仮想通貨と金融規制：仮想通貨とブロックチェーン技術が従来の金融システムに及ぼす影響と，それに対する適切な規制の形はどうあるべきか？

システムの処理時間が長く，短期集中型の議論では処理が追い付かないことが考えられる．そのため，1 回の議論は約 1 時間程度と長時間にし，緩やかなペースで議論を行ってもらった．また，議論後には以下の 8 つのアンケート項目に答えてもらった．

1. Bot による情報推薦の頻度は適切でしたか？
2. 情報が欲しいタイミングで情報推薦が行われていましたか？
3. 情報推薦により発言しやすくなりましたか？
4. Bot の発言や参考文献を読んだことで投稿した自分の発言はいくつありましたか？
5. 推薦した情報(参考文献)は議論内容に関連していましたか？
6. 推薦した情報(参考文献)は議論の役に立ちましたか？

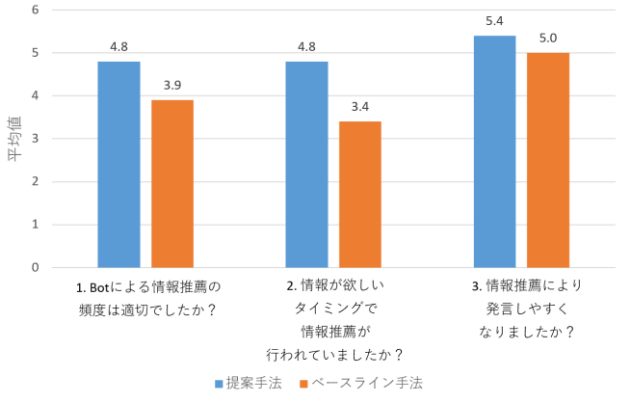

図 2　アンケート項目 1~3 の結果

7. トピックに対する自分なりの結論を簡単に記述して下さい
8. 実験を通して気づいたことや思ったことがあったらご自由にお書きください

項目 1，2，3，5，6 は 1~7 の 7 段階評価である．

5.2 実験結果

システムの情報推薦のタイミングや頻度に関する項目 1~3 の結果は提案手法のシステム A とベースライン手法のシステム B で比較した．結果を図 2 に示す．1~3 の 3 項目のいずれにおいても今回開発したシステム A のほうが平均値は高くなった．そのため，提案手法により情報を適切なタイミングで推薦でき，ユーザの議論への参加が容易になる傾向があることがわかった．

推薦した情報の議論への関連度と有益度に関する項目 5, 6 に関しては，システム A を用いたときの結果と，先行研究[1]の議論実験時の結果を比較した．これらの項目は議論テーマの難易度や理解度の差によって評価が大きく異なることが予想されたので，グループ 1 の 5 人の参加者とグループ 2 の 5 人の参加者で別々にアンケート結果を集計した．以下の表 2 に結果を示す．

表 2　グループ毎のアンケート項目 5，6 の結果と先行研究との比較

	グループ 1	グループ 2	先行研究
項目 5	6.4	4.6	5.4
項目 6	6.4	5.4	5.1

グループ 2 の項目 5 以外の平均値は先行研究のシステムより高い値になっているのがわかる．有意水準 5%でマン・ホイットニーのU検定を行った結果を以下の表 3 に示す．

表 3　先行研究と比較した U 検定の結果

項目	グループ	統計検定量(U)	棄却限界値 (α=0.05)	有意差
項目 5	1	11.5	8	無し
項目 5	2	17	8	無し
項目 6	1	8	8	あり
項目 6	2	22.5	8	無し

グループ 1 の項目 6 では統計的に有意な差が見られたことがわかった．よって提案手法によりグループ 1 では議論への関連度と有益度が高い情報を推薦できていた傾向があることがわかった．

発言数の増加に関する質問 4 に関して，システムの発言や参考文献を読んだことで投稿された投稿は，グループ 1 では 22 個の参加者の発言に対し 6 個，グループ 2 では 17 個の参加者の発言に対して 5 個あったことがわかった．よって，グループ 1 では 37.5%, グループ 2 では 41.7%発言数が増えたことがわかる．

5.3 考察

情報推薦のタイミングと頻度に関して，提案手法の方がベースライン手法よりも優れている傾向にあることがわかった．しかし，アンケート結果を個別に見ると低評価の者も存在した．これには 2 つの理由が考えられる．1 つ目は投稿に返信する形で直接情報推薦をされなかった参加者が存在するからである．今回の議論実験では，1 回の議論につき参加者は 5 人で，システムは平均 7 回の情報推薦を行っていた．複数回情報推薦が行われた人，反対に 1 回も行われなかった人もいることが確認できた．そのような参加者から見たら情報推薦のタイミングが適切とは感じにくく，評価値が低くなったのではないかと考えた．反対に，直接情報推薦が行われた参加者は高い評価値になったと考えられる．2 つ目はシステムの処理時間が長いためである．これにより，議論に追い付けず，最新の投稿よりもいくつも前の投稿に対して返信をすることがあったため，評価値があまりよくならなかったと考えた．実際にアンケート項目 8 の自由記述欄には，「BOT の投稿がもう少し早くなればうれしいと感じました」というコメントがあった．

推薦情報の議論への関連度と有益度の評価はグループ 1 とグループ 2 で差があった．これは設定した議論トピックの難易度と背景知識が影響したと考えられる．今回グループ 2 は議論トピックに対して十分な背景知識を持っていた人が一人もいなく，情報が推薦されてもあまり理解できず，議論が進まなかったと考えられる．グループ 2 ではシステムによって発言が増加しているが，議論全体の投稿数が 17 であり，グループ 1 と比べて議論が活発でなかったことが確認できる．実際に実験の感想を記述する質問の回答には，「議論トピックが難しすぎて発言できなかった」や「システムから単語などを説明されても理解できなかった」という意見があった．

以上より，本システムは今回の実験のグループ 1 のように，複雑な議論トピックであるが背景知識を持った参加者が数人おり，システムがなくてもある程度は議論が進むような状況においては有効であることがわかった．

6. まとめと今後の課題

本研究では，web 議論において LLM が適切なタイミングで情報推薦を行うエージェントの開発を行った．このエージェントは情報を適切なタイミングで推薦でき，議論参加者の議論への参加を容易にする傾向があることがわかった．また，ファクトチェック機能を取り入れたことから LLM が生成する発言の信頼性が向上したことがわかった．しかし，推薦情報の議論への関連度と有益度は，議論トピックや参加者の背景知識の影響を受けると考えられる．

今後はシステムの処理時間の短縮に取り組みたい．このエージェントは 1 回投稿を行うまでに GPT-4 Turbo を何度も使う設計となっている．そのため，処理時間が長く，短時間の集中的な議論では使用できない．Fine-tuning を行ったり，別の LLM モデルを試したりすることでこの問題の解決を試みたい．また，議論トピックの難易度によって生じる問題について詳しく調べたい．これにより，状況に応じて適切な情報推薦ができるようになると考えている．さらに，被験者数を増やして実験することで，このエージェントの評価をより正しく行いたいと考えている．

謝辞

本研究の一部は，JST CREST (JPMJCR20D1) および NEDO (JPNP20006) の支援を受けた．

参考文献

[1] Ryosuke Kinoshita, Shun Shiramatsu, “Agent for Recommending Information Relevant to Web-based Discussion by Generating Query Terms using GPT-3” Proceedings of the 6th IEEE International Conference on Agents, 2022.
[2] 木下良輔, 櫻井崇貴, 白松俊, "LLM を用いたファクトチェック機能の試作と Web 議論における関連情報推薦システムへの応用",人工知能学会第二種研究会資料, CCI-012 号, pp.41-44 (2024)
[3] Takaaki Sugiyama, Kotaro Funakoshi, Mikio Nakano, Kazunori Komatani, “Estimating response obligation in multi-party human-robot dialogues”, 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids), 2015.
[4] H. S. Zheng et al., “Take a step back: Evoking reasoning via abstraction in large language models”, arXiv, 2310.06117, cs.LG, 2024.
[5] Jason Wei et al., “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”, Advances in Neural Information Processing Systems, volume 35, pages 24824--24837, 2022.
[6] Tom B. Brown et al., “Language Models are Few-Shot Learners”, Advances in neural information processing systems, volume 33, pages 1877--1901, 2020.