

視野の広い情報取得のための閲覧ページに対する他者観点の提示方法 A Method of Presenting Other People's Viewpoints on Web Pages for Acquiring Information from a Broader Perspective

小竹 神¹⁾ 北山 大輔¹⁾
Jin Kotake Daisuke Kitayama

1 背景

検索エンジンの発展によって、ウェブ検索を用いて情報収集することが多くなっている。検索ユーザは、自身の興味・観点に基づいて検索をする。また、複数人で同様のトピックを検索し、目的を達成する検索もある。それは、協調検索と呼ばれる。例えば、複数人で旅行計画を立てている際に、観光地について調べるときに協力して行われる検索である。このような背景から、情報ニーズ自体は、他者と類似していることも多いが、同様の情報ニーズであっても、ユーザによって閲覧する情報が異なる場合は、そのユーザにとって共有された情報は有用であると考えられる。関連研究として Pickens ら [1] は、協調検索の仲介をするシステムを構築し、評価し、検索クエリを共有する検索エンジンを使用した場合に、ユニークな関連文書の発見数等が向上することを示した。また、Bhuiyan ら [2] は、他者の YouTube の推薦リストを提示することによる動画視聴行動の違いについて検証した。

そこで本研究では、他者のモデルを OpenAI が公開した GPT にユーザのクエリ履歴を入力することで作成し、それによって検索ユーザが閲覧したページ中の部分コンテンツの判定を行い、その判定結果に基づいてその部分コンテンツを強調表示するシステムの実装を行った。本研究の目的は、ユーザ実験を行い、ユーザの閲覧ページ中の一部を他者の観点で強調表示することで、閲覧されるページが多様化するかを確かめることである。本稿では、ユーザ実験を実施するためのシステムについて述べる。

2 提案手法

提案システムの処理の概要を図 1 に示す。提案システムは、処理 1 から処理 4 の流れで成り立つ。まず、図 1 の処理 1 「仮想他者の検索クエリ生成」について述べる。この処理では、現在検索しているユーザの検索クエリを用いて仮想他者の検索クエリ生成を行う。本稿では、Google が提供する Google Suggest API¹⁾ にユーザの検索クエリの 1 語目を投げ、その検索クエリに対するサジェストワードを取得する。検索クエリの 1 語目を使用する理由は、検索トピックを代表する可能性が高いためである。次に、Google Suggest API から得たサジェストワードとユーザの検索クエリの 2 語目との類似度を算出し、最も値が低いサジェストワードを仮想他者の 1 番目の検索クエリとする。ここでは、ユーザの検索クエリとサジェストワードを fasttext を用いてベクトル化し、コサイン類似度を用いて算出した。

そして、仮想他者の 2 番目以降の検索クエリ生成について説明する。仮想他者の検索クエリにも、実際のユーザの検索クエリのように一貫性があることが求められる。そこで、仮想他者の 2 番目以降の検索クエリの

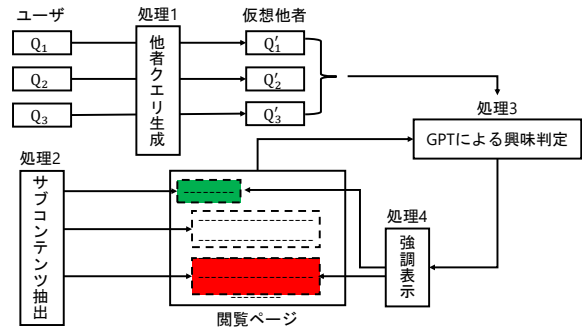


図 1 提案システムの処理の流れ

条件は「ユーザの 2 番目の検索クエリとは類似しなく、仮想他者の直前のクエリと類似する」と設定した。それは、式 (1) で表現した。式 (1) について、 $Score(c)$ の $\text{sim}(q_{i,2}, c)$ はユーザの i 番目の 2 語目とサジェストワード c との類似度、 $(1 - \text{sim}(p_{i-1,2}, c))$ は仮想他者の $i-1$ 番目の 2 語目とサジェストワード c との類似度を示す。また、類似度は、コサイン類似度により算出する。式 (1) より求めたスコア c が最も低いサジェストワードを仮想他者の検索クエリとする。

$$score(c) = \text{sim}(q_{i,2}, c) \cdot (1 - \text{sim}(p_{i-1,2}, c)) \quad (1)$$

次に、図 1 の処理 2 「閲覧ページ中からサブコンテンツの取得」について述べる。閲覧中のページ（以下、閲覧ページ）のサブコンテンツ抽出は、以下の条件を満たすものとした。

- そのタグのテキスト文字数が N 文字以上であるかつそのタグの直接の子タグが存在しない場合：そのタグ中のテキストを N 文字以下になるように分割する
- そのタグのテキスト文字数が M 文字 N 文字未満である場合：そのタグ中のテキストを取得する

最後に、図 1 の処理 3 「GPT を利用した仮想他者の興味判定」と処理 4 「閲覧ページに閲覧興味判定結果の提示」について述べる。まず、処理 3 について述べる。処理 3 は図 1 の処理 2 で得られたユーザの閲覧ページ中のサブコンテンツと処理 1 で得られた仮想他者のクエリ履歴とユーザのクエリ履歴による説明文との類似度判定を行う。

まず、ユーザ/仮想他者の各クエリ履歴を OpenAI が提供している GPT の API²⁾ を用いてそのユーザのクエリ履歴のトピックを表す説明文生成を行う。表 1 は、ユーザ/仮想他者の各クエリ履歴からそのトピックを表す説明文を GPT により生成する際の各項目の設定を示す。Model は、gpt-3.5-turbo を使用した。出力する単語の創造性を 0 から 1 で表現する temperature は 0.0 とした。0.0 と設

1) 工学院大学大学院工学研究科情報学専攻

1) <https://developers.google.com/cloud-search/docs/reference/rest/v1/query/suggest?hl=ja>

2) <https://platform.openai.com/docs/api-reference/chat>

表 1 GPT による説明文生成のための設定

System	ユーザ A が以下の検索クエリの履歴を持っています。そのときに、ユーザ A が興味を持っているトピックを説明する文章を 200 字程度で作成してください。
User	クエリ履歴「○○」、「○○」、「○○」

表 2 GPT による説明文生成結果

【ユーザ】: ユーザ A は、東京での食事に興味を持っているようです。特に、ディナーやダイニングバー、そして肉料理に注目しているようです。東京は、美味しい食事が豊富にあることで知られており、ユーザ A はその中でも特に高級感のある場所や、こだわりのある肉料理を求めているのかもしれませんが。
【仮想他者】: ユーザ A は東京に興味を持っていることがわかります。特に、東京の代表的な観光スポットである東京タワーや、交通手段である東京メトロについても関心があるようです。東京都についても検索していることから、東京全体に興味を持っていると考えられます。

定した理由は、出力される説明文をプロンプトに対して固定化するためである。生成する文章の最大単語数を表す max_tokens は、200 とした。そして、プロンプトの System と User は、表 1 の通りである。System は、User のクエリ履歴からそのトピックを説明するための文章を作成させるためのプロンプトである。User は、ユーザ/仮想他者のクエリ履歴を入力させるためのプロンプトである。

次に、生成したユーザ/仮想他者のクエリ履歴のトピックを表す説明文と閲覧ページ中のサブコンテンツとの類似度算出を行う。fasttext を用いた文書ベクトルによるコサイン類似度を用いた。この類似度を興味度として扱い、ユーザと仮想他者の興味度を降順にソートし、上位 N 件を強調表示する対象とする。

最後に、処理 4「閲覧ページに閲覧興味判定結果の提示」について述べる。この処理は、処理 3 で得られたユーザ/仮想他者が興味を持つであろうサブコンテンツを閲覧ページ中にて強調表示する処理である。具体的には、閲覧ページ中で該当するサブコンテンツの背景色をユーザ/仮想他者によって適応した。

3 出力例

ここでは、1 つの検索シナリオに対して 2 章で述べた処理 1 から処理 3 を適応した出力結果について考察する。使用した検索シナリオは、「あなたは東京でディナーをするための飲食店を探しています。今から Web 検索をして東京でディナーに適する飲食店を検索してください」である。またユーザ役は、著者が行い、結果および考察は、処理ごとに述べる。

まず、処理 1「仮想他者の検索クエリ生成」について述べる。ユーザの入力したクエリは、「東京 ディナー」、「東京 ダイニングバー」、「東京 肉料理」である。そして、このユーザのクエリ履歴から生成された仮想他者のクエリは、「東京都」、「東京タワー」、「東京メトロ」である。生成された仮想他者のクエリは、ユーザのクエリと似ていないかつ自身の直前のクエリと似ていることを目的として生成したが、検索シナリオとの関連度が低いワードとなった。この原因は仮想他者のクエリ生成には、Google Suggest API から得たサジェストワード中から選択しているためである。今後は、サジェストワードのバリエーションが豊富になるように改善する。

次に、処理 2「閲覧ページ中からサブコンテンツの取得」について述べる。ユーザが閲覧したページは、シナリオに基づいて「東京にある ディナーにおすすめレストラン 20 選」³⁾とした。このページ中から取得したサブコンテンツの一部は、以下の通りである。この結果から、ページ中の各スポットの説明文が取得できていると考える。

- Peter / ザ・ペニンシュラ東京日比谷 / ステーキ & グリル 4.20 (232 件) - 良い 146,000 円~7,999 円 20,000 円~29,999 円最上階の 24 階に位置し、どこまでも続く美しい眺望とモダンなインテリアでゆったりとした時間をお過ごしいただけます。

最後に、処理 3「GPT を利用した検索/仮想他者の興味判定」について述べる。ユーザ/仮想他者のクエリ履歴のトピックを表す説明文の生成結果を、表 2 に示す。表 2 のユーザ/仮想他者の説明文の内容は、上記で述べたクエリ履歴のトピックを説明する文章になっていると考える。次に、生成されたユーザ/仮想他者の説明文とサブコンテンツとの興味度判定の結果について述べる。ユーザ/仮想他者の説明文と最も類似度が高いサブコンテンツは、上記の Peter / ザ・ペニンシュラ東京日比谷であった。この結果より、ユーザが興味を持つサブコンテンツとしては妥当だと考える。理由は、説明文は肉料理について言及しているためである。一方で、仮想他者が興味を持つサブコンテンツとしては適切でないと考えられる。理由は、説明文は、東京について言及しておりサブコンテンツとの関連度が低いと考えたためである。以上の結果から、今後は、ユーザ/仮想他者の説明文生成およびサブコンテンツのベクトル化手法を改善する。

4 まとめ

本稿では、他者のユーザのモデルを OpenAI が公開した GPT にユーザのクエリ履歴を入力することで作成し、それによって検索ユーザが閲覧したページ中の部分コンテンツの判定を行い、その判定結果に基づいてその部分コンテンツを強調表示するシステムの実装について述べた。今後は、出力例で出た課題の改善およびユーザ実験を実施する。

謝辞

本研究の一部は、2023 年度科研費基盤研究 (C)(課題番号: 21K12147) によるものです。ここに記して謝意を表すものとします。

参考文献

- [1] Jeremy Pickens, Gene Golovchinsky, Chirag Shah, Pernilla Qvarfordt, and Maribeth Back. Algorithmic mediation for collaborative exploratory search. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pp. 315–322, 2008.
- [2] Md Momen Bhuiyan, Carlos Augusto Bautista Isaza, Tanushree Mitra, and Sang Won Lee. Othertube: Facilitating content discovery and reflection by exchanging youtube recommendations with strangers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pp. 1–17, 2022.

3) <https://restaurant.ikyuu.com/area/tokyo/05/ranking/>