トピック理解のためのより貪欲な情報探索を促進する問いかけ文の提示

齊藤　史明†　　山本　祐輔†

† 静岡大学大学院総合科学技術研究科　〒 432–8011 静岡県浜松市中区城北 3–5–1
E-mail: †saito@design.inf.shizuoka.ac.jp, ††yamamoto@inf.shizuoka.ac.jp

あらまし　本稿では，より貪欲なウェブ情報探索を促進するために，トピックへの理解に欠かせない知識をもっと調べたくなる問いかけ文をウェブから発見し，ウェブ検索中のユーザに提示する手法について提案を行う．本稿では，公的機関のウェブサイトで掲載されている Q&A リストを利用し，Yahoo！知恵袋に投稿された質問の中から問いかけに適した質問文を選択する学習器を構築する．提案手法によって選択された問いかけ文をウェブ検索中のユーザに提示することで，その後のウェブ検索行動に与える影響を分析するために，健康・医療トピックに関するウェブ検索タスク実験を行った．

キーワード　行動変容，ウェブ情報探索，QA データ処理，ソーシャルビッグデータ

1　はじめに

ウェブ情報の信憑性について，人々が精査行動をしないことが問題となっている．情報源を意識せずにウェブ情報探索を行なうユーザや，情報の信憑性について疑問を抱いたことがないユーザが多く存在することが報告されている [1] [2] [3]．情報の信憑性の精査を積極的に行わなければ，真偽の確認できない情報を信用して誤った意思決定を行う可能性や，信憑性の低い情報の拡散に加担してしまう可能性がある．また，情報の精査不足は情報の信憑性の誤判断以外にも問題がある．不十分な状態で情報探索を終えてしまうことで，情報探索しているトピックに対して不正確な理解や偏った情報の拡散を行ってしまう危険性がある．

十分な情報精査が行われない理由の 1 つに，認知バイアスの影響が挙げられる．ウェブ探索ユーザは，認知バイアスと呼ばれるある種の先入観をもってウェブ探索を行うことが知られている．たとえば，自身の信念を支持する情報を優先的に集めてしまう確証バイアス [4]，特定のドメインのウェブページを信用するドメインバイアス [5]，検索結果一覧 (以下，SERP) の上位に掲載されているウェブサイトを優先的に閲覧するポジションバイアス [6] の存在が知られている．検索アルゴリズムに対する過大評価も情報精査を消極的にする要因の 1 つである [3]．ウェブ検索を通じて確実な意思決定を行うためには，先入観を排して必要となる情報を貪欲に収集する必要がある．

より貪欲な情報探索を促すシステムとして，ユーザの検索行動の量や時間を内省させるシステム [7]，ウェブ検索中にクエリ推薦を行う際に，そのクエリで検索した際に得られる情報量を表示するシステムなどが提案されている [8]．これらの提案システムでは検索行動の量についてユーザに注目させている．そのため，ユーザが十分な検索を行ったかどうかが，理解ではなく数値的な行動指標が満たされたかどうかに置き換わっている．

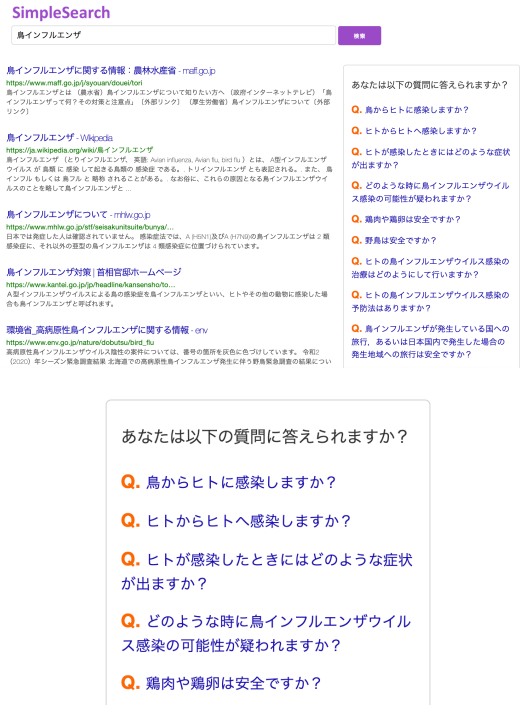

図 1　提案システムの動作イメージと問いかけの拡大図．

行動指標を満たしたとしても，それが検索対象の理解に繋がっていない場合，さらなる情報探索が求められる．探索時間や情報量といった行動指標ではなく，検索トピックについての理解を問うインタラクションを行うことで，より効果的に情報探索を促すことができると考える．

本研究では，問いかけられたユーザが調べたくなる，調べているトピックへの理解に欠かせない知識を問う形式の問いかけを表示することで，より貪欲な情報探索の促進を狙う．たとえば，鳥インフルエンザについてその危険性や症状をウェブ検索しているユーザに対して，図 1 のように，「鳥からヒトに感染し

ますか？」というように，検索トピックについて調べるべきサブトピックを，ユーザが気になるような問いかけ形式の文で表示する．この問いかけにより，ユーザが知らなかった危険性を知るためのより貪欲な情報探索が促進されることが期待される．

提示する問いかけ文の選択には，Q&A サイトの Yahoo！知恵袋を利用する．Yahoo！知恵袋に投稿された質問を，調べているトピックへの理解に欠かせない知識を問うものであるか評価するアルゴリズムを提案する．提案アルゴリズムは，文章内容やその質問への回答数，同じトピックで他に投稿された質問群との類似度を特徴量として質問の評価を行う．そして，学習器の評価が高かった質問を調べたくなる問いかけとして Web 検索結果一覧ページ (SERP) に表示する．提案システムにより，情報探索が促進され，ユーザの意思決定に良い影響を与えることが期待される．

2 関 連 研 究

2.1 Search as Learning

情報検索プロセスにおけるトピックの学習過程の分析や，学習のための情報検索システムを対象とした研究分野として，Search as Learning と呼ばれる分野が存在する [9]．

Rieh らは，学習する知識の種類によって学習プロセスや検索行動が異なることを報告しており，その違いを考慮してケースに合った検索支援を行うことの重要性を説いている [10]．

Collins らは，SERP 画面に表示する情報の豊富さによるユーザのウェブ検索行動の違いを分析した [11]．検索ログと学習成果を分析した結果，ユーザの予想する学習成果と実際の成果が一致しており，検索結果の豊富さにより高い成果を残すことを明らかにした．

2.2 ウェブ探索行動を促進するインタラクション

Umemoto らはウェブ検索中にクエリ推薦を行う際に，今までに得た情報量とそのクエリで検索することで得られる情報量を表示するシステムを提案した [8]．提案システムは，ユーザにまだ得られる情報があることを提示することで，網羅的なウェブ探索行動の促す．

Saito らは，情報の出典が曖昧な表現をハイライト表示して情報探索を促す手法を提案した [12]．ユーザ実験の結果，提案手法によってウェブブラウジング中に訪れるページ数の増加や，ブラウジングに掛ける時間の増加が見られた．

以上の手法では，ユーザはトピック理解のためのウェブ情報探索を行っていない可能性が考えられる．理由として，表示されている情報量を十分な値にするためにウェブページを閲覧したり，ハイライト表示が少ない文献を探したりすることが目的になっている場合が考えられる．本研究では，情報量といった行動指標ではなく，直接そのトピックについての理解を問うことで，ユーザにウェブ探索の必要性を説明する．

2.3 ウェブ探索の支援

Collins らは，クエリの推薦方法の違いがウェブ探索を用いた学習効率に与える影響を調べた [11]．多くのサブトピックについてウェブページを閲覧できるクエリを推薦し，SERP 画面にウェブページのリンクと共に関連するクエリを表示することで，トピックに対して高い学習成果をもたらすことが明らかになった．

Harvery らは，欲しい情報が得られるような良いクエリ例をウェブ探索ユーザに提示することで，クエリ生成スキルを向上させる手法を提案した [13]．

本研究は，以上の研究と同様にウェブ探索の支援を目的としている．本研究では，トピックへの理解に欠かせない知識をもっと調べたくなる文を問いかけることで，ユーザに調べるべきサブトピックを知らせると同時に，検索を促す手法を提案する．

3 提 案 手 法

本章では，Yahoo！知恵袋に投稿された質問から，任意のトピックの理解に欠かせない知識をもっと調べたくなる問いかけ型質問文を発見する手法について述べる．また，ウェブ情報探索中のユーザにより貪欲な情報探索を促すために，問いかけ型質問文を提示するシステムについて述べる．

3.1 貪欲な情報探索を促す問いかけの性質

粟津は，良い質問とは問われた人が思わず答えたくなる，新しい気づきを与えてくれる質問であると定義している [14]．思わず答えたくなるという性質には，言い回しや表現が問いかけの受け手にとって適切であるかという要素や，不快にさせるような問いかけ内容ではないという要素が挙げられる．本研究では，対象とする「問いかけ型質問文」は，検索しているトピックに関して，押さえるべきサブトピックに関して問うものであり，問いに思わず答えたくなる，考えたくなるような言い回しを持つ文であると仮定する．

3.2 問いかけの発見方法

本稿では，疑問文形式の文（質問文）がトピックへの理解に欠かせない知識をもっと調べたくなる文であるかを判定する問題を，2 値分類問題として定式化する．提案システムは，入力となる文をベクトル化し，訓練済みのモデルを使用し判定する．以下で，(1) 学習・評価用データ，(2) ラベルデータの付与方法，(3) 特徴量，(4) 学習器，(5) 分類性能について述べる．

3.2.1 学習・評価用データ

学習器の学習・評価用データとして，ヤフー株式会社が提供している国立情報学研究所の情報学研究データリポジトリ「Yahoo！知恵袋データ (第 2 版)」を利用する [1]．このデータセットから，筆者が設定したトピックについての質問文を抽出する．トピックの設定には，医療・健康分野から，3.2.2 節で述べる公的機関の Q&A 集があるものとした．データセットのうち，トピック名が質問タイトルに含まれており，文末が”?”，”？”である質問文を抽出する．また，3.2.3 節で述べる LexRank の計算量のため，5000 件を上限として抽出した．表 1 に，設定したトピック名と各トピックで使用した文章数を示す．

1：https://www.nii.ac.jp/dsc/idr/yahoo/chiebkr2/Y_chiebukuro.html

抽出した質問文に対して，トピックへの理解に欠かせない知識をもっと調べたくなる文であるか否かを示すラベルを付与（3.2.2 節参照）し，正例と負例を各トピックで 100 件づつ抽出し 1200 件のデータセットを作成した．

3.2.2 ラベルの付与方法

抽出した質問文に対して，トピックへの理解に欠かせない知識をもっと調べたくなる文であるか否かを示すラベルを付与するために，公的機関が用意した Q&A 集を利用する．あるトピックについて，公的機関が特に周知したい事柄や，人々が疑問に思う事柄について Q&A 方式でウェブサイトに公開している場合がある．

公的機関の用意した Q&A 集は，知るべきサブトピックについて丁寧な表現で質問が用意されており，3.1 節で述べた貪欲な情報探索を促す問いかけの性質に合ったものであると考えられる．

Yahoo！知恵袋の質問文が公的機関 Q&A の質問群と類似していれば，その質問文はトピックへの理解に欠かせない知識をもっと調べたくなる質問であると仮定する．

以下，ラベルの判定方法を述べる．まず，Yahoo！知恵袋に投稿された 1 質問文に対し，公的機関の Q&A 集の各質問文ごとに類似度を測定する．この類似度群の最大値を Yahoo！知恵袋の 1 質問文の評価値とする．この作業を，学習・評価用データのすべての質問文に行う．得られた評価値群の中央値を閾値として，閾値以上の質問文を正例，閾値未満の質問文を負例とした．

類似度の測定には，文中の単語について分散表現を得る BERT を利用して文のベクトル化を行い，ベクトル間のコサイン類似度を計算した．BERT とは，Devlin らが提案した自然言語処理モデルであり，文脈をベクトル表現に組み込むことが可能になっている [15]．文のベクトル化には，BERT 事前学習済みモデルを使い，文中の各単語について 768 次元の分散表現を得た [2]．そして，単語毎に得た分散表現から 1 文を意味内容を表現するベクトルを得るために，平均プーリングを行い，得られたベクトルを文のベクトルとした．

本研究では公的機関の条件として，URL のトップレベルドメインおよびセカンドレベルドメインが，日本の政府機関を示す「go.jp」であるウェブサイトの Q&A 集を対象とした．

3.2.3 特 徴 量

提案手法では Yahoo！知恵袋に投稿された質問文の評価のために，以下の特徴量を使用し，質問文のベクトル化を行う．

BERT: 文に BERT 事前学習済みモデルを適用し，768 次元のベクトル表現を得た [2]．

Bag-of-Words: 教師データ中の文に対して形態素解析を行い，特徴語辞書を作成し，文中の単語が作成した辞書に含まれているかどうかを 2 値ベクトルで表現する．文の分かち書きには、形態素解析器 MeCab を使用した [3]．また，トピック特有の単語による影響を受けないようにするため，名詞以外の品詞を抽出対象とした．

質問への回答数*:* 質問に対して多く回答が寄せられるということが，質問文の内容的な特徴として表現できると考えられる．質問に対して投稿された回答数をベクトル化した．

LexRank: Yahoo！知恵袋に投稿された質問の中には，個人的な相談や多く閲覧してもらうことを狙った投稿など，問いかけに適さない質問も存在する．そのような質問を省くために，質問文群の中で普遍的な質問かを表現する特徴量が必要である．そこで，データセットの質問文群内での相対的な類似度を特徴量として利用する．この特徴量の計算のために，文章要約に使われる LexRank を利用する [16]．LexRank は，文章群の中で，多くの文章と似ている文章を重要な文章であると定義して重要度を計算する．文をノードとし，BERT ベクトル化した文間のコサイン類似度をエッジとするグラフ構造を作成する．そして，このグラフの各ノードの滞在確率を計算し，特徴量としてベクトル化した．

3.2.4 学 習 器

分類器の構築には，機械学習アルゴリズムの GBDT を用いた [17]．GBDT の実装には，LightGBM を利用した [18]．パラメータは，Python の LightGBM ライブラリのデフォルトパラメータを用いた [4]．

3.2.5 評 価

3.2.1 節で用意したデータを用いて，分類性能の評価を行った．評価では，3.2.3 節で述べた特徴量を組み合わせたパターンを複数作成し比較を行った．評価指標として，適合率，再現率，F1 値，AUC の 4 つを使用した．5 分割のクロスバリデーションを行った結果を表 2 に示す．

表 1 使用したトピック名と質問の数．

トピック	対象の質問数
ダイエット	5000
ノロウイルス	413
鳥インフルエンザ	302
アスベスト	259
健康食品	258
BSE	219

表 2 各特徴量パターンの分類性能．各指標で最も高かった数値を太字で示す．

	再現率	適合率	F1	AUC
質問への回答数， LexRank	0.548	0.508	0.507	0.514
質問への回答数， LexRank， Bag-of-Words	0.628	0.601	0.614	0.643
質問への回答数， LexRank， BERT	0.780	0.786	0.782	**0.874**
BERT， Bag-of-Words	0.775	0.778	0.776	0.872
質問への回答数， BERT， Bag-of-Words	0.775	0.778	0.776	0.872
LexRank， BERT， Bag-of-Words	**0.783**	**0.786**	**0.784**	0.869
質問への回答数， LexRank， BERT， Bag-of-Words	**0.783**	0.785	0.783	0.869

表 2 に示すように， LexRank，BERT，Bag-of-Words の特徴量を用いた学習器が再現率，適合率，F1 値で一番高い性能を示した．また，BERT 特徴量が特に性能向上に寄与すること

2：東北大学の乾研究室が構築した事前学習済みの BERT モデルを利用した．(https://github.com/cl-tohoku/bert-japanese)

3：MeCab: http://taku910.github.io/mecab/

4：https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMClassifier.html

が分かった．

また，性能の違いによって学習器が実際にどのような質問を正例と判定したか，一番性能の高かった「質問への回答数, BERT, Bag-of-Words」と一番低かった「質問への回答数, LexRank」の 2 つの特徴量パターンで比較した．用意した 6 つのトピックからトピック「アスベスト」以外の 5 トピックを学習データとして学習器を作成し，トピック「アスベスト」の質問について評価を行った．表 3 と表 4 に，2 つの学習器が正例である確率が高いと判断した質問の上位 3 件を示す．

表 3　性能の高かった特徴量パターン「質問への回答数, BERT, Bag-of-Words」で正例である確率が高いと判断した質問の上位 3 件

・アスベストと診断されたら、余命は短いのでしょうか？
・アスベストの含まれた空気を何回呼吸すれば病気になるのでしょうか？
・アスベストが原因の肺がん発症の可能性の有無が分かる方法は有りますか？

表 4　性能の低かった特徴量パターン「質問への回答数, LexRank」で正例である確率が高いと判断した質問の上位 3 件

・新築を機に IH クッキングヒーターに変えようと考えていますが母が第二のアスベストで何十年先に子や孫に癌が発生するからと反対です。本当でしょうか？
・空気中に漂うアスベストはまったく見えないそうですが、どうすればいいのですか？
・アスベストの被害って、役人の怠慢による人災ですよね？

3.3　ウェブ検索中に問いかけ型質問文を提示するインタラクション

提案システムは，ウェブ検索時に検索トピックに関する問いかけを SERP 画面に表示する．システム稼動例を図 1 に示す．

トピック「ノロウィルス」を例にシステムの動作を以下に示す:

（1） ユーザが「ノロウィルス」と検索クエリを入力すると，システムはそのクエリを取得する．

（2） 取得したクエリがタイトルに含まれる，文末が “?”, “？”である Yahoo！知恵袋の質問を検索する．

（3） システムが事前に構築した分類器を用いて各文がトピックへの理解に欠かせない知識をもっと調べたくなる文であるか判定する．

（4） 分類器が調べたくなる文である確率が高いと判定した順に上位 10 件を SERP 画面に表示する．

（5） 問いかけと共に,「あなたは以下の質問に答えられますか？」と問いかけ文の説明を表示する．

問いかけ文表示の説明を表示することで，質問文自体に意識を向けさせる効果を狙う．問いかけ文に注目がされなかった場合，問いかけの効果が発揮されないため，上記の文言で問いかけの重要性を説明する．

3.4　仮　　説

提案システムが問いかけを行うことで，ユーザにはウェブ探索行動の促進が期待される．そこで本稿では，以下の仮説を設定し検証を行う．

H1　提案システムにより，ウェブ探索にかける時間が長くなる．

H2　提案システムにより，ウェブ探索でクエリの発行回数が増える．

H3　提案システムにより，訪問するウェブページの件数が増える．

H4　提案システムにより，SERP 画面のより下部まで閲覧するようになる．

また提案システムの，トピックへの理解に欠かせない知識をもっと調べたくなる文を問いかけるという性質から，次の仮説を設定する．

H5　H1，H2，H3，H4 の効果は，ユーザの検索トピックについての事前知識に影響される．

4　実　　験

本章では，提案システムの効果を分析するためのユーザ実験について述べる．

4.1　被　験　者

クラウドソーシングサービス Lancers を用いて，200 名の被験者を募集した．被験者のうち，タスクの遂行に不備のあった被験者を除外した．その結果，残った 180 名の被験者を分析の対象とした．各被験者には，実験参加の報酬として 110 円を支払った (平均タスク時間 7 分 53 秒)．

4.2　タ　ス　ク

被験者には，ウェブ検索エンジンを模したシステムを用いた検索タスクを行なった．検索トピックには，3.2.1 節にてデータセット作成のために使用した中から，ノロウイルス，鳥インフルエンザ，アスベストの 3 つのトピックを設定した．医療・健康分野は，身近なテーマかつ慎重な意思決定が求められるため，検索トピックとして選択した．

被験者は指定した検索トピックの概要を知るために，指定した検索システムを用いてウェブ情報探索を行った．被験者が十分に探索を行なったと感じた時点で検索タスクを終了してもらった．検索タスクの終了後に，概要についてまとめた文章を入力してもらった．また，被験者には各検索タスクの事前に，検索トピックについて事前知識の程度を自己申告で回答してもらった．事前知識の程度は，5 段階のリッカート尺度で回答してもらった (1:まったくそう思わない，5:かなりそう思う)．

4.3　実験システム

実験システムには，検索と SERP 画面の表示を行うシステムを構築した．実験システムは，4.5 節で述べる行動指標を測定するために，タスク中の被験者の行動ログを記録した．被験者が検索クエリを入力すると，検索システムが検索を行い，クエリに対する検索結果を SERP 画面に表示する．検索システム

は，Bing Web Search API[5]を利用した．

SERP 画面に表示される検索結果の要素は，ウェブサイトのタイトル，スニペット文，URL の 3 つである．また，SERP 画面に表示されるウェブサイトは最大で 100 件とした．本実験では，3 種類の SERP 画面インタフェースを用意した．1 つ目は， 3 節で述べた，問いかけを表示する **questioning** である (図 1 参照)．2 つ目は， **questioning** 条件の問いかけ文から複数の単語を抽出したものを，関連検索キーワードとして表示する **term_suggestion** である．3 つ目は，問いかけや関連キーワードが表示されない **plain** である．インタフェース間で問いかけの表示以外に機能の違いは無い．学習器の精度による影響が考えられるため，今回は **questioning** 条件で表示する問いかけ文の生成には 3 章で述べた学習器を用いる手法は使用しない．良い問いかけを行うことによる効果を測定することが目的であるため，筆者が任意で質問を選択する方法をとった．**questioning** 条件で表示する問いかけは，3. 2. 2 節で述べた条件を満たす公的機関が公開している Q&A 集から，10 件を筆者が選択した．また，**term_suggestion** 条件で表示する単語も，**questioning** 条件で表示する 10 件の問いかけ文それぞれで任意に文を構成する単語を選択した．

4. 4 手　　順

実験はウェブを介して行なった．各被験者は，提案インタフェースを使う **questioning** 群，ベースラインインタフェースを使う **term_suggestion** 群，**plain** 群の 3 群のいずれかに無作為に割り当てた．被験者はすべてのタスクで同じ UI を使用した．取り組むタスクの順番の影響を取り除くために，タスクの順番は被験者ごとにランダムに設定した．被験者に実験参加への同意を確認し，実験用ウェブサイトに移動してもらった．その後，検索タスクを 3 件行ってもらった．各検索タスクの事前に，その検索トピックについて事前知識の程度の回答を求めた．事前知識の回答後に，検索タスクの導入文として以下の文章を表示した．

> **以下の状況を想定してください:**
> **あなたは，Twitter や Yahoo！といったウェブサイトで,「アスベスト」が話題になっていることを知りました.「アスベスト」に興味を持ったあなたは，ウェブ検索を行い情報を集めることにしました．**
> **この文章の下の「検索結果を表示」ボタンを押して，こちらで用意したウェブ検索システムを利用して「アスベスト」について調べてください.「アスベスト」についてあなたなりに十分に情報を得られたら，ウェブ検索を終えてください．ウェブ検索を終えましたら，得た情報をまとめて「調べた結果のまとめ」の欄に入力してください．**

検索タスクごとに上記の導入文の「ノロウイルス」の箇所がトピック名になる．被験者は説明を読んだ後,「検索を開始する」ボタンをクリックすると，最初にトピック名をクエリとした際の SERP 画面が表示される．SERP 画面に表示されたウェブページにアクセスするか，別の検索クエリを入力することができる．また，**questioning** 群と **term_suggestion** 群は，SERP 画面に表示される問いかけや単語をクリックすることで，それをクエリとした SERP 画面に遷移する．そして，十分に情報を得られたと判断し次第ウェブ探索を終了した．これを 3 タスク繰り返し，実験を完了した．

4. 5 測 定 項 目

3. 4 節で述べた仮説について検証するために，以下の検索タスク中の被験者の行動を測定する．

タスクの所要時間　各タスクで被験者がタスク全体に費やした時間を測定する．貪欲なウェブ探索が促進されることで，所要時間が長くなると考えられる．この変数を用いて，仮説 **H1** および仮説 **H5** の検証を行う．

クエリ発行数　各タスクで被験者が発行したクエリ数を測定する．貪欲なウェブ探索が促進されることで，クエリ発行数が増えると考えられる．この変数を用いて，仮説 **H2** および仮説 **H5** の検証を行う．

SERP 画面の閲覧時間　各タスクで被験者が SERP 画面を閲覧した時間を測定する．貪欲なウェブ探索が促進されることで，閲覧時間が長くなると考えられる．この変数を用いて，仮説 **H1** および仮説 **H5** の検証を行う．

記事ページの訪問件数　各タスク中に被験者が表示したウェブページの数を測定する．貪欲なウェブ探索が促進されることで，閲覧するウェブページの件数の増加が考えられる．この変数を用いて，仮説 **H3** および仮説 **H5** の検証を行う．

検索結果の最大クリック深度　各タスクで被験者が訪問したウェブページの，検索結果順位の最大値を測定する．貪欲なウェブ探索が促進されることで，SERP 画面を注意深く閲覧するようになり，検索結果の下位のウェブページまで閲覧するようになると考えられる．閲覧した記事ページよりも深く SERP 画面を閲覧している可能性が考えられるが，最低限確実に閲覧した場所を深度として選択した．この変数を用いて，仮説 **H4** および仮説 **H5** の検証を行う．

5 結　　果

本章では，4 章で述べたユーザ実験の結果について記す．180 名の被験者のうち，**plain** 群が 62 名，**term_suggestion** 群が 61 名，**questioning** 群が 57 名となった．

5. 1 分 析 方 法

本稿では，データの分析に一般化線形混合モデル (GLMM) を使用した [19][6]．これは，ベイズ統計モデリングを行い，目的変数となる各行動指標について，説明変数となる UI 条件と事前知識のパラメータの事後確率分布を推定する手法である．

5：https://docs.microsoft.com/ja-jp/azure/cognitive-services/bing-web-search/

6：GLMM には，R パッケージの brms を利用した [20].

GLMM は，UI などの固定効果に加えて，被験者間の個人差やタスクの違いによる差をランダム効果として考慮してモデリングを行うことが可能になる．また，古典的な仮説検定手法と違い，データ数が異なる場合や，等分散性や正規性が確認できない場合でも分析することが可能である．

GLMM 分析には，説明変数として筆者が設定した「UI 条件」と事前アンケートで調査した被験者の「事前知識」を設定した．目的変数には，4.5 節で述べた「タスクの所要時間」,「クエリ発行数」,「SERP 画面の閲覧時間」,「記事ページの訪問件数」,「検索結果の最大クリック深度」を設定した．また，UI 条件とトピックへの知識量を固定効果に，被験者と検索タスクをランダム効果として設定した. [21] に従い，ユーザ実験で測定した各項目を下記式でモデリングを行った．

$$\begin{aligned} \mathrm{Y} \sim\ & \mathrm{UI} + \mathrm{Topic} + \mathrm{UI} : \mathrm{Topic} + (1|\mathrm{Task}) \\ & + (1 + \mathrm{UI} + \mathrm{Topic} + \mathrm{UI} : \mathrm{Topic}|\mathrm{Participant}) , \end{aligned}$$

上式において，`Y` は目的変数，`UI` は各参加者の UI 条件を表すバイナリ値，`TOPIC` は検索トピックに対する知識量を意味する．また，`(x|y)` は `y` が `x` のランダム効果になっていることを意味する．

GLMM は目的変数の確率分布について仮定することが可能である．Liu らは，ウェブページの閲覧時間がワイブル分布に従うことを報告した [22]．そのため，目的変数の「タスクの所要時間」と「SERP 画面の閲覧時間」はワイブル分布に従うと仮定した．また,「クエリ発行数」については，0 回であったデータが非常に多かったため，ゼロ過剰ポアソン分布に従うと仮定した.「記事ページの訪問件数」,「検索結果の最大クリック深度」については，ポアソン分布に従うと仮定した．

本稿では，各パラメータの最高密度区間 (HDI) を用いて仮説を検証する．HDI は，ある確率でパラメータが存在する区間を示したものである．パラメータの 95%HDI に 0 が含まれない場合は，頻度論的統計の有意性 ($p > .05$) に相当する．また，UI 条件は **questioning** 条件を基準として，**plain** 条件と **term_suggestion** 条件で差を分析した．

5.2 分析結果

5.2.1 タスクの所要時間

仮説 H1 および H5 の検証のために，1 タスクの所要時間について分析した．タスクの所要時間について GLMM の結果を表 5 に示す．所要時間は，タスクの説明・回答入力画面が表示されてから，回答を送信した時間までの間とした．分析の結果，各係数の 90%HDI に 0 が含まれていた．これは，頻度論的統計で $p > 0.1$ に相当し，帰無仮説を棄却できなかったことを意味する．よって，表 5 の 90%HDI から見て，問いかけ UI がタスクの所要時間に影響を与えると主張できない．

5.2.2 クエリ発行数

仮説 H2 および H5 の検証のために，クエリ発行数について分析した．1 タスク中に発行したクエリ発行数について GLMM の結果を表 6 に示す．実験システムが最初に検索を行うので，ユーザが検索ワードを 1 度入力し直した場合に初めてクエリ発行数が 1 となる．分析の結果，各係数の 90%HDI に 0 が含まれていた．これは，頻度論的統計で $p > 0.1$ に相当し，帰無仮説を棄却できなかったことを意味する．よって，表 6 の 90%HDI から見て，問いかけ UI がクエリ発行数に影響を与えると主張できない．

表 5　タスクの所要時間の GLMM の結果．各独立変数の係数の平均，標準誤差，90% HDI，95% HDI.

変数名	平均	標準誤差	90% HDI	95% HDI
切片	3.33	4.45	[-4.41, 6.07]	[-4.41, 6.15]
UI 条件 (**plain**)	-0.28	0.55	[-1.16, 0.28]	[-1.16, 0.38]
UI 条件 (**term_suggestion**)	-0.37	0.81	[-1.73, 0.35]	[-1.73, 0.44]
事前知識	0.49	0.77	[-0.02, 1.83]	[-0.05, 1.83]
UI 条件 (**plain**) * 事前知識	0.16	0.32	[-0.11, 0.70]	[-0.14, 0.70]
UI 条件 (**term_suggestion**) * 事前知識	0.17	0.39	[-0.15, 0.83]	[-0.18, 0.83]

表 6　クエリ発行数の GLMM の結果．各独立変数の係数の平均，標準誤差，90% HDI，95% HDI.

変数名	平均	標準誤差	90% HDI	95% HDI
切片	-1.25	0.86	[-2.60, 0.18]	[-2.96, 0.41]
UI 条件 (**plain**)	-1.98	1.57	[-4.59, 0.55]	[-5.08, 1.05]
UI 条件 (**term_suggestion**)	-0.81	1.39	[-3.18, 1.40]	[-3.64, 1.90]
事前知識	-0.04	0.25	[-0.45, 0.39]	[-0.53, 0.46]
UI 条件 (**plain**) * 事前知識	-0.17	0.62	[-1.14, 0.85]	[-1.41, 1.04]
UI 条件 (**term_suggestion**) * 事前知識	-0.55	0.67	[-1.57, 0.57]	[-1.92, 0.67]

5.2.3 SERP 画面の閲覧時間

仮説 H1 および H5 の検証のために，SERP 画面の閲覧時間について分析した．1 タスク中の SERP 画面の閲覧時間について GLMM の結果を表 7 に示す．分析の結果，各係数の 90%HDI に 0 が含まれていた．これは，頻度論的統計で $p > 0.1$ に相当し，帰無仮説を棄却できなかったことを意味する．よって，表 5 の 90%HDI から見て，問いかけ UI が SERP 画面の閲覧時間に影響を与えると主張できない．

5.2.4 記事ページの訪問件数

仮説 H3 および H5 の検証のために，記事ページの訪問件数について分析した．1 タスク中に SERP 画面からクリックした記事ページの数について GLMM の結果を表 8 に示す．分析の結果，各係数の 90%HDI に 0 が含まれていた．これは，頻度論的統計で $p > 0.1$ に相当し，帰無仮説を棄却できなかったことを意味する．よって，表 8 の 90%HDI から見て，問いかけ UI が記事ページの訪問件数に影響を与えると主張できない．

5.2.5 検索結果の最大クリック深度

仮説 H4 および H5 の検証のために，検索結果の最大クリック深度について分析した．1 タスク中に訪れた記事ページの検索順位について GLMM の結果を表 9 に示す．分析の結果，各係数の 90%HDI に 0 が含まれていた．これは，頻度論的統計で $p > 0.1$ に相当し，帰無仮説を棄却できなかったことを意味する．よって，表 9 の 90%HDI から見て，問いかけ UI が検索結果の最大クリック深度に影響を与えると主張できない．

表 7　SERP 画面の閲覧時間の GLMM の結果．各独立変数の係数の平均，標準誤差，90% HDI，95% HDI.

変数名	平均	標準誤差	90% HDI	95% HDI
切片	1.58	3.47	[-4.41, 3.90]	[-4.41, 4.02]
UI 条件 (**plain**)	-0.38	0.53	[-1.16, 0.24]	[-1.16, 0.37]
UI 条件 (**term_suggestion**)	-0.50	0.77	[-1.73, 0.27]	[-1.73, 0.40]
事前知識	0.43	0.81	[-0.14, 1.83]	[-0.18, 1.83]
UI 条件 (**plain**) * 事前知識	0.17	0.32	[-0.14, 0.70]	[-0.19, 0.70]
UI 条件 (**term_suggestion**) * 事前知識	0.21	0.37	[-0.14, 0.83]	[-0.18, 0.83]

表 8　記事ページの訪問件数の GLMM の結果．各独立変数の係数の平均，標準誤差，90% HDI，95% HDI.

変数名	平均	標準誤差	90% HDI	95% HDI
切片	1.19	0.29	[0.80, 1.58]	[0.70, 1.71]
UI 条件 (**plain**)	-0.03	0.25	[-0.45, 0.38]	[-0.54, 0.46]
UI 条件 (**term_suggestion**)	-0.21	0.26	[-0.62, 0.22]	[-0.70, 0.30]
事前知識	-0.07	0.07	[-0.19, 0.05]	[-0.22, 0.07]
UI 条件 (**plain**) * 事前知識	0.00	0.10	[-0.15, 0.17]	[-0.18, 0.20]
UI 条件 (**term_suggestion**) * 事前知識	0.05	0.10	[-0.12, 0.21]	[-0.15, 0.25]

表 9　検索結果の最大クリック深度の GLMM の結果．各独立変数の係数の平均，標準誤差，90% HDI，95% HDI.

変数名	平均	標準誤差	90% HDI	95% HDI
切片	1.73	0.29	[1.30, 2.17]	[1.21, 2.28]
UI 条件 (**plain**)	-0.01	0.33	[-0.55, 0.52]	[-0.67, 0.60]
UI 条件 (**term_suggestion**)	0.08	0.33	[-0.45, 0.62]	[-0.53, 0.78]
事前知識	-0.07	0.09	[-0.21, 0.08]	[-0.23, 0.12]
UI 条件 (**plain**) * 事前知識	0.03	0.13	[-0.18, 0.24]	[-0.22, 0.27]
UI 条件 (**term_suggestion**) * 事前知識	-0.04	0.13	[-0.24, 0.17]	[-0.29, 0.21]

6　考　　察

6.1　貪欲な情報探索を促す問いかけの学習器の性能

3.2.5 節で述べたように，本稿で作成した貪欲な情報探索を促す問いかけの学習器の中で，LexRank, BERT, Bag-of-Words の特徴量を使用した学習器が再現率，適合率，F1 値でそれぞれ一番高い性能を示した．表 3 と表 4 が示すように，性能の違いにより良い質問である確率が高いと判断する文章に違いが見られた．

性能の高い学習器では，公的機関の Q&A 集に近い内容を問う質問が選択されているのに対し，性能の低い学習器では個人的な質問や問いかけとして相応しくない質問が選択されていた．性能の高い学習器には BERT を特徴量抽出器として使用しているため，公的機関の Q&A 集と意味的に近い質問がより選択されるようになったと考えられる．Q&A サイト特有の特徴量として作成した質問への回答数や LexRank については，BERT と Bag-of-Words 特徴量と組み合わせた際に，質問の回答数のみを使用した際に BERT, Bag-of-Words 特徴量と性能が変わらなかった．対して，LexRank 特徴量を組み合わせた際に性能向上が見られたことから，LexRank 特徴量が有効であると考えられる．今回使用したトピックにおいては，質問者の多くが同じような質問をするトピックである可能性が考えられるため，他に個人的な質問が多くなるトピックが存在した場合に，LexRank 特徴量が機能するかが課題である．また，今回構築した学習器について，BERT 特徴量の Attention weight を分析することで，良い問いかけの言語的特徴の考察が可能になると考えられる．

6.2　問いかけ文の提示効果

ユーザ実験の結果，タスクの所要時間，クエリの入力回数,SERP 画面の閲覧時間，記事ページの訪問件数，検索結果の最大クリック深度について，UI 条件と事前知識量，およびその相互作用による効果は確認されなかった．これにより，仮説 H1，H2，H3，H4，H5 は支持されなかった．

効果が確認されなかった理由として，検索結果の設計が考えられる．医療・健康情報の分野で検索をした際に，ユーザ実験で表示した問いかけ文の抽出元である公的機関の Q&A 集が検索結果上位に含まれていた．そのため，検索開始直後に問いかけ内容の答えとなる情報がすぐに発見できる状態になってしまい，ユーザがすぐにタスクを終えてしまった可能性が考えられる．被験者の行動ログを分析した結果，鳥インフルエンザでは 24 名, ノロウイルスでは 84 名，アスベストでは 98 名の被験者が，システムの問いかけ文として使用した公的機関の Q&A サイトを訪問していた．これらの被験者は，表示される問いかけと同じ内容の文や単語群を閲覧している可能性が考えられる．一般的に検索結果の上位のウェブサイトを訪問しやすい傾向が報告されていることからも，公的機関の Q&A 集が訪問されやすく問いかけの効果が少なくなったことが考えられる [23].

今後の実験では，公的機関の Q&A 集を問いかける場合にそのウェブサイトを SERP から除外するか，公的機関の Q&A 集が SERP の上位にならないトピックをタスクとする実験設計にすることが必要だと考える．また，分析方法として訪問した記事ページの内容や滞在時間についても分析する必要があると考える．

提案システムの評価として，行動に注目したユーザ実験を行ったが，提案システムを使ったことによる態度面の評価を行う必要があった．行動面で貪欲な情報探索を促進すること以外にも，ユーザが情報探索を行おうとする意識に提案システムが影響しているか調査する必要があった．これは，普段の検索意識や，実験後に UI としての満足度や検索意識を問うことで調査することができる．また，問いかけを見た際のユーザの反応や意思決定のプロセスを知るためにも，事後インタビューを行い検証する必要がある．

7　ま　と　め

本稿では，より貪欲なウェブ情報探索を促進するために，トピックへの理解に欠かせない知識をもっと調べたくなる問いかけ文をウェブから発見し，ウェブ検索中のユーザに提示する手法を提案した．問いかけ文の作成には，Q&A サイトの Yahoo！知恵袋に投稿された質問の中から問いかけに適した質問を選択する．質問の選択には，公的機関のウェブサイトに掲載されている Q&A リストと，Yahoo！知恵袋に投稿された質問との類似度を予測する学習器を構築し，学習器による評価を行う．評

価の結果，学習器は BERT 特徴量を用いることで精度向上が見られ，精度は F1 値が 0.784 であった．精度向上のために，今後は多くのトピックをデータセットとする必要がある．また，構築した学習器の特徴量について，BERT 特徴量の Attention weight を分析し，良い問いかけの言語的特徴の考察が必要である．

ユーザ実験を行った結果，トピックの理解に欠かせない知識をもっと調べたくなる問いかけによるウェブ探索行動の影響は認められなかった．これは，問いかけ文自体やその回答がすぐに見つけられるトピックをタスクに設定したことが要因の１つと考えられる．今後は，表示する問いかけを，すぐに閲覧できるウェブページの内容と違うことを条件として理想の問いかけを行う実験や，問いかけによる態度面への影響にも注目した実験を行う必要がある．

謝　　辞

本研究は JSPS 科研費 JP18KT0097，JP18H03243，JP18H03494，C18H032440 および課題設定による先導的人文学・社会科学研究推進事業の助成を受けたものです．ここに記して謝意を表します．

文　　献

[1] Miriam J Metzger, Andrew J Flanagin, and Lara Zwarun. College student web use, perceptions of information credibility, and verification behavior. *Computers & Education*, Vol. 41, No. 3, pp. 271–290, 2003.

[2] Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. What do people ask their social networks, and why? a survey study of status message q&a behavior. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 1739–1748, 2010.

[3] Satoshi Nakamura, Shinji Konishi, Adam Jatowt, Hiroaki Ohshima, Hiroyuki Kondo, Taro Tezuka, Satoshi Oyama, and Katsumi Tanaka. Trustworthiness analysis of web search results. In *Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2007)*, pp. 38–49, 2007.

[4] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, Vol. 2, No. 2, pp. 175–220, 1998.

[5] Samuel Ieong, Nina Mishra, Eldar Sadikov, and Li Zhang. Domain bias in web search. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pp. 413–422, 2012.

[6] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*, pp. 87–94, 2008.

[7] Scott Bateman, Jaime Teevan, and Ryen W. White. The search dashboard: How reflection and comparison impact search behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2012)*, pp. 1785–1794, 2012.

[8] Kazutoshi Umemoto, Takehiro Yamamoto, and Katsumi Tanaka. Scentbar: A query suggestion interface visualizing the amount of missed relevant information for intrinsically diverse search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 405–414, 2016.

[9] Jacek Gwizdka, Preben Hansen, Claudia Hauff, Jiyin He, and Noriko Kando. Search as learning (sal) workshop 2016. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 1249–1250, 2016.

[10] Soo Young Rieh, Kevyn Collins-Thompson, Preben Hansen, and Hye-Jung Lee. Towards searching as a learning process: A review of current perspectives and future directions. *Journal of Information Science*, Vol. 42, No. 1, pp. 19–34, 2016.

[11] Kevyn Collins-Thompson, Soo Young Rieh, Carl C Haynes, and Rohail Syed. Assessing learning outcomes in web search: A comparison of tasks and query strategies. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pp. 163–172, 2016.

[12] Fumiaki Saito, Yoshiyuki Shoji, and Yusuke Yamamoto. Highlighting weasel sentences for promoting critical information seeking on the web. In *International Conference on Web Information Systems Engineering*, pp. 424–440. Springer, 2020.

[13] Morgan Harvey, Claudia Hauff, and David Elsweiler. Learning by example: training users with high-quality query suggestions. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 133–142, 2015.

[14] 粟津恭一郎. 「良い質問」をする技術. ダイヤモンド社, 2016.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2019)*, pp. 4171–4186, 2019.

[16] Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, Vol. 22, pp. 457–479, 2004.

[17] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.

[18] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, pp. 3146–3154, 2017.

[19] Jisoo Lee, Erin Walker, Winslow Burleson, Matthew Kay, Matthew Buman, and Eric B. Hekler. Self-experimentation for behavior change: Design and formative evaluation of two approaches. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI 2017)*, pp. 6837–6849, 2017.

[20] P. Christian Bürkner. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, Vol. 80, No. 1, pp. 1–8, 2017.

[21] D. Barr, R. Levy, C. Scheepers, and H. Tily. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, Vol. 68, No. 3, pp. 255–278, 2013.

[22] Chao Liu, Ryen W. White, and Susan Dumais. Understanding web browsing behaviors through weibull analysis of dwell time. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010)*, pp. 379–386, 2010.

[23] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)*, Vol. 25, No. 2, pp. 7–es, 2007.