# 潜在的興味の表示による検索結果の選択性の向上

関　　峰† 北山　大輔†

† 工学院大学大学院工学研究科情報学専攻　〒 163-8677 東京都新宿区西新宿 1 丁目 24-2
E-mail: †em22006@ns.kogakuin.ac.jp, ††kitayama@cc.kogakuin.ac.jp

**あらまし**　現在，人々はさまざまな検索エンジンや推薦システムを使っている．しかし，これらのシステムを利用する場合，通常，ユーザは提示されたコンテンツリストから，自身の興味に合致するかどうかを判断し，コンテンツを選択する．この時，ユーザのコンテンツの選択性を向上させるために，ユーザの興味と検索結果の関連性を表示するシステムが有効だと考える．そこで，本研究では，SentenceBERT を用いたユーザの興味ベクトルと検索結果ベクトルの類似度に基づく関連性を可視化するシステムを提案し，実行結果を示す．

**キーワード**　検索エンジン，類似性分析，ユーザ支援

## 1　はじめに

近年，インターネットやモバイル，スマートフォンの普及に伴い，人々は様々な検索システムを利用して問題を解決することが多くなっている．そのような中で，ユーザに対する情報検索，情報推薦システムが特に重要となってきている．生活の中では，自分が分からない分野のコンテンツに出会うことが多く，検索結果等の断方的な情報では，ユーザが興味との合致を判断することは困難である．

その時ユーザが現在の検索ドメインとは直接的には関係のない，潜在的な興味との関係が示されると，検索結果を選択しやすくなると考えられる．すなわちユーザのコンテンツの選択性を向上させるために，ユーザの興味と検索結果の関連性を表示することが有効と考える．図 1 を例に説明する．あるユーザの Web 閲覧履歴にポケモンや旅行に関連するページが多く含まれており，このユーザはポケモンや旅行に興味があると仮定する．このユーザが閲覧履歴に対して無関係なコンテンツを取得した場合（例えば，コーヒーを検索した場合）．ポケモンに関連する Web サイトを強調表示すると，ユーザに閲覧を促すことができる．

そこで，本稿では，ユーザが閲覧したページに基づいて，ユーザが知っている情報や嗜好を判断し，嗜好に応じた強調表示を行うことで，ユーザのコンテンツの選択性を向上する方法を提案する．

## 2　関連研究

本研究では，ユーザの嗜好を判定及び，分類し，ユーザのコンテンツの選択性を向上させるために，潜在的興味と検索結果の関連性表示するシステムを提案する．そこで，関連研究として，ユーザ情報の嗜好分析と，コンテンツへの情報付加の 2 つの観点の研究を紹介する．

### 2.1　ユーザの嗜好分析

潘ら [1] はユーザの未知なスポットに対する理解を支援するために，訪問したことがある観光スポットの特徴を用いて，未訪問エリアの観光スポットを説明する方法を提案した．提案手法の方がより詳細な情報を提示しているため，ユーザが未訪問スポットをイメージしやすいことを確認した．

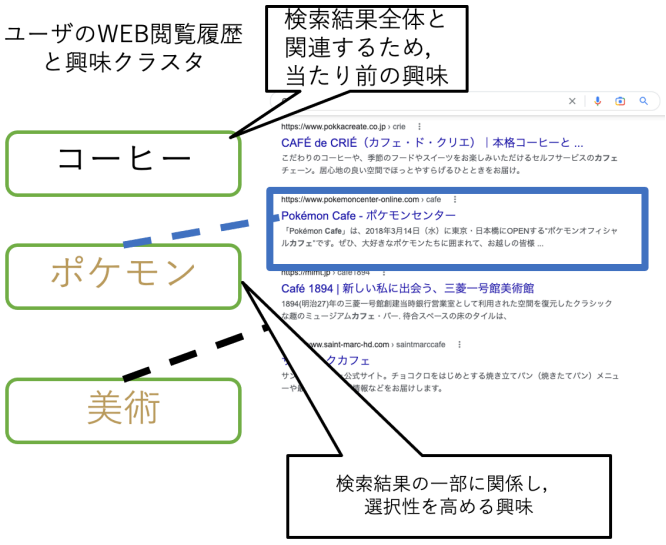


図 1　検索結果に含まれる潜在的興味の強調の例

山田ら [2] は，検索結果内の未訪問ページの内容に対し，含まれる情報と訪問済みページとの比較情報として既知度をユーザの閲覧判断基準として提示することで，訪問済みページの語集合から含まれるトピックを階層的クラスタリングを用いて推定し，未訪問ページのスニペットの語集合からその内容を類似度で予想することで，その既知度合いを算出する手法を提案した．

塩川ら [3] は，ウェブ上で学術用語を解説するページ群を対象として，それらのページ群における用語解説の分かり易さおよび見易さを自動評定する手法を提案した．理工系学術分野を対象とし，学術用語を検索クエリとする検索上位ウェブページを収集し，各ページに対し人手により参照用データの作成を行う．

甲谷ら [4] は QA コミュニティの活性化を目的とし，QA コミュニティの回答者の知識，興味を推定し，質問を推薦する手法を提案した．ユーザの知識を推定する際にコンテンツフィル

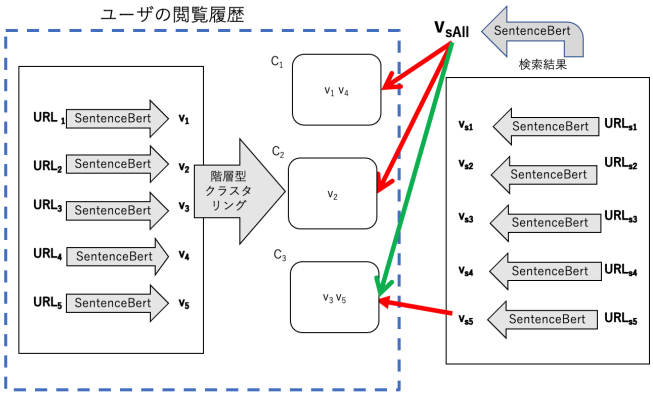


図 2　システム概要

タリングを行ったと協調フィルタリングでテキストを使わずにユーザの知識と興味を推定する．

### 2.2　コンテンツへの情報付加

kyriakidi ら [5] は異なる推薦要素を容易に捕らえ，異種情報空間のつながりを探索し，アルゴリズムの設計と評価を容易にし，拡張性の高い推薦エンジンにつながる，適切な小さな抽象化セットを持つ統一モデルを目標していた．研究結果はデータモデル，すなわちドメインの要素をモデル化するためのグラフと，計算モデル，すなわち異なる推薦を表現するためのパス演算子の小さなセットを持つ，統一された方法論を提案した．

真野ら [6] は検索結果の状態を可視化する手法を提案する．検索クエリ変更前後の状態変化を比較できるシステムを提案することで，検索クエリの設定能力を育成する手法を提案した．検索結果の状態を把握するために，検索結果の話題を表す語句を提示し，語句の判別に基づいて検索結果の状態を可視化する機能を有する．検索クエリ設定能力の育成に効果がある可能性が示唆された．

池田ら [7] は Twitter の反応を利用しニュースの全体像の理解支援を行うための可視化手法を提案した．Twitter で投稿されたニュースに対する反応としてリプライ，引用リツイートを用い，ニュース自体の特徴語と反応の特徴語を抽出する．抽出した特徴語を利用して，ニュースや反応の特徴および他のニュースとの関連性をわかりやすく可視化する．

片岡ら [8] はユーザに提示される情報に偏りが生じていた場合，ユーザが接触している情報がどの程度他のユーザと乖離しているのかを気付かせるシステムを提案した．ユーザがフィルターバブルの存在を自覚することで，情報の探索領域を広げ，自らフィルターバブル問題を解決するように促すシステムの提案を行った．

## 3　潜在的興味の抽出と関係性の判定

### 3.1　システムの概要

本研究では，ユーザの興味と検索結果コンテンツの関連性を検索結果中に示すステムを提案する．システムの流れを図 2 に示す．全ての Web ページは SentenceBERT による文書の特徴ベクトルで表現され，ユーザの興味も同様に特徴ベクトルで表現されるものとする．赤い線は類似しているベクトル同士を表している，緑の線は類似していないベクトル同士を表している．クラスタ $C_3$は，検索結果全体である $V_{s_{All}}$と類似しておらず，今回の検索に無関係な潜在的興味と考える．そして，その $C_3$と類似する $URL_{s_5}$は，潜在的興味を含む検索結果と考えられるため，これを強調表示する．

システムは以下の流れになる．

- ユーザの Web 閲覧履歴を取得し．ユーザの興味を抽出する．
- 階層的クラスタリングを使用して，閲覧履歴をクラスタリングする，得られたクラスタを興味とする．
- 今の Web 検索結果全体の特徴ベクトルと類似しない興味クラスタを潜在的興味とする．潜在的興味に類似する，検索結果中の個別 Web ページを強調表示する．

### 3.2　文書の特徴ベクトルの生成

本研究では，SentenceBERT により文書を特徴ベクトルで表現する．SentenceBERT [9]（以下は SBERT）は，文のエンコーディングや分類などの自然言語処理タスクを実行するために学習させた言語モデルの一種である．Hugging Face の研究者によって開発され，BERT [10][1] アーキテクチャに基づいている．sentenceBERT はモデルの上限を超える (一般に 512 トークン) 長い文章を扱えない．実際に使用するのは，長い文章であることが多いため，我々は長いテキストを 800 字程度で分割する．まず URL ページからすべてのテキスト情報を抽出し，その後分割されたテキストの平均ベクトルを出し，文書ベクトルとする．

### 3.3　ユーザの興味抽出

ユーザの興味抽出には，階層型クラスタリングを使用した．階層型クラスタリングは，データを階層的にグループ化するためのクラスタリングアルゴリズムである．本研究ではデータ間距離にコサイン，クラスタ間距離に群平均法を使用して階層的にクラスタリングを行う．

本稿では閾値の決定に 8 カテゴリに分かれることを想定した閲覧履歴を用意した予備実験を行った．閾値が 0.6 の場合に階層型クラスタリングの結果が想定したものに最も近かったためそれを用いる．

### 3.4　潜在的興味の判定

潜在的興味を，今の関心事に対しての直接的な関係のない，ユーザが日常的に持っている関心事と定義する．そこで，今の関心事を現在の検索対象とすると，検索結果ページ全体とは類似しない興味クラスタを潜在的興味と考えることができる．そこで，検索結果に対し，検索結果ページベクトル $v_{sall}$ と，個々のページベクトル $v_{si}$ を作成し，潜在的興味およびそれを含むページを判定する．

まず，個々のページベクトル $v_{si}$ に関しては，sentenceBERT

1：BERT：Bidirectional Encoder Representations from Transformers

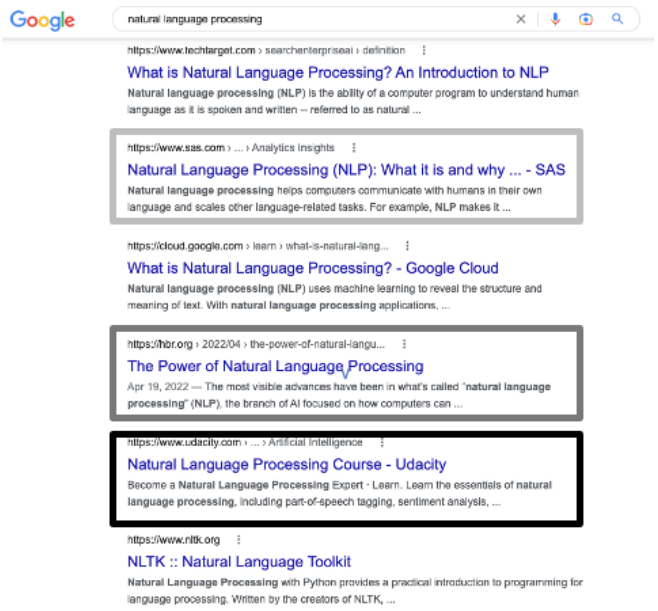


図 3　潜在的興味の判定例

を用いて，検索結果のタイトルとスニペットを結合した文をベクトルに変換したものを用いる．次に，検索結果全体のベクトル $v_{sall}$ に関しては，全ての $v_{si}$ の平均ベクトルを用いる．

クラスタのベクトルはクラスタに属する閲覧履歴ベクトルの平均ベクトルである．$v_{sall}$ に対して，類似度が閾値 0.5 未満となるクラスタを潜在的興味クラスタとし，潜在的興味クラスタと類似度が閾値 0.5 以上となる個々のページをそれを含むページとし，強調表示する．強調度合いは，類似度に基づいて行う．図 3 に強調表示の例を示す．類似度がより高いものは濃い色の四角で囲み，類似度が低くなるにつれて薄い色の四角で囲む．このことにより，ユーザは潜在的興味との関係を直感的に把握することが可能となる．

## 4　実 行 結 果

被験者から得たデータによる実行例を示す．データの収集は以下のように行った．

（1） 閲覧履歴：興味があるトピックを複数個回答し，それに関連する Web ページ 20 件を回答する．

（2） 検索キーワード：興味があるトピックを 1 つ回答する．(1. のもの以外)

（3） 正解ページ：興味にあう Web ページを 0 件から 3 件回答する．この時，2. のトピックによる検索結果中から選び，かつ 1. で入力したトピックも考慮して判断する．

本実験では，25 名の被験者のデータを用いた．システムによる強調ページと被験者による正解ページが 1 件以上一致した被験者は 0 名であった．

25 名の被験者の閲覧履歴をクタスタリングした結果がどの程度回答した興味トピックと一致するか，以下の確認を行った．まず，被験者が回答した興味トピックに対して，閲覧履歴を割り当てて，それを正解クラスタとする．次に，システムの生成したクラスタと正解クラスタで共通ページが最も多いクラスタをその正解に対応するクラスタとする．最後に，対応するクラスタ内で正解クラスタであるページを「入力トピックとクラスタが一致したページ」とし，これが多いほどクラスタリングによる興味抽出が成功していると考える．表 1 に結果をまとめる．

5 番目や 11 番目の被験者のように，クラスタリングが適切で，ユーザの興味が抽出できたと考えられる結果においても，被験者の正解ページとシステムの強調ページが一致しないことから，強調ページ判定のアルゴリズムに問題があることが考えられる．これらの被験者のデータを用いて，検索結果全体との類似度および個別ページとの類似度の閾値等を検討する必要がある．

入力されたトピック数に対し，システムのクラスタ数が少ない場合がある．現在，ページ単位でクラスタリングを行っているが，意味のある内容で分割し，粒度を細かくすることで，ユーザの考えるトピックの粒度にそろえる必要があると考えられる．また，トピックの内容が非常に近い場合，分類の正確性が低下することもある．例えば，第 25 の例での英語と留学のような関連性の強い内容では，システムはそれらを同じカテゴリに分類することがある．これは，今回のクラスタリングの正解を判定する単位を被験者が入力したトピック単位としていることに起因しており，事前に類似するトピックを集約するなどしてから評価する必要があったと考える．これらのことから，興味抽出に関してはクラスタリング対象の粒度を細かくし，表現能力を向上することが必要と考える．また，強調ページの判定に関しては，検索結果と興味クラスタの関係を分析し，閾値のみならず判定アルゴリズムを整理する必要がある．

## 5　ま　と　め

本研究はユーザのコンテンツの選択性を向上させるために，ユーザが閲覧したページに基づいてユーザの興味と検索結果の関連性を表示するシステムを提案した．本研究は SentenceBERT 手法でユーザの閲覧履歴と検索結果の文書の特徴ベクトルで表現し，長いテキスト分割されたテキストのベクトルの平均を出した．その後，群平均法階層型クラスタリングを使用し，ユーザの閲覧履歴に分類を行った．現行の検索されたキーワードが以前に検索した履歴のクラスタリングの平均ベクトルのコサイン類似度が類似していない場合，各検索結果のベクトルの他のクラスタリングの平均ベクトルに対するコサイン類似度を個別に計算し，ユーザの潜在的関心対象であると判断した．実験の結果から，入力されるトピックの数が多い場合や，トピックの内容が密接に関連している場合には，分類の正確性が低下することがわかった．また，関連性の強いトピックでは，システムがそれらを同じカテゴリに分類する傾向があった．今後の課題はシステムの改善と強調表示による選択性向上の効果の検証として考えている．

## 謝　　辞

本研究の一部は，2022 年度科研費基盤研究 (C)(課題番号：

表 1 実 行 結 果

| 番号 | 入力したトピック数 | クラスタ数 | 入力トピックと<br>クラスタが一致した数 | 入力トピック |
|---|---|---|---|---|
| 1 | 3 | 4 | 11/20 | トルコ地震、東京に数年ぶりの大雪、露のパリ五輪出場禁止 |
| 2 | 4 | 3 | 14/20 | ChatGPT, WBC, ルフィ, 三苫薫 |
| 3 | 10 | 5 | 6/20 | プロ野球キャンプ, オンラインゲーム, ホロライブ, 海外ドラマ, 映画,<br>M リーグ, 転職, マスク緩和, 飲食店迷惑行為, トルコ大地震 |
| 4 | 6 | 3 | 8/20 | 無農薬野菜，野菜の栄養，株投資，EV，テスラ，chatgpt |
| 5 | 4 | 4 | 19/20 | 進化論, フランス革命,WBC, 英語の発音 |
| 6 | 4 | 2 | 9/19 | マイナポイント,aupay, 日銀総裁,u-next |
| 7 | 3 | 4 | 15/18 | マウンテンバイク, 日銀総裁, 第 5 類 |
| 8 | 5 | 2 | 8/16 | チョ・グク元法務部長官に懲役 2 年, 中国気球撃墜, 卒業式マスクなし,<br>ウクライナへの戦闘機供与, ロシアとベラルーシのパリ五輪出場禁止を要請 |
| 9 | 4 | 4 | 8/20 | King Gnu,Vocaloid,Cocco, フジファブリック |
| 10 | 5 | 6 | 11/20 | 海洋生物、星座、育児、小学生勉強、北海道科学館 |
| 11 | 3 | 3 | 20/20 | 釣り、値上げ、サイクリング |
| 12 | 4 | 3 | 10/20 | k-pop、韓国、t-pop、タイ |
| 13 | 10 | 6 | 13/19 | サッカー, お菓子, パソコン, アイドル, アニメ,<br>イヤホン, 編集ソフト, 副業, 農業, 外国 |
| 14 | 3 | 2 | 13/18 | マイナポイント, 値上げ, 節約術 |
| 15 | 4 | 3 | 15/20 | fx, ファスティング, 手抜き料理, 砧公園 |
| 16 | 5 | 2 | 6/20 | iPhone, 携帯, ツイッター, 機種変更,Mac |
| 17 | 6 | 7 | 11/20 | 地震, 寄付, 節約, バレンタイン, 映画, オンラインショッピング |
| 18 | 5 | 4 | 13/20 | puipui モルカー、ブルーハムハム、ヤバ T、鍋料理、ちいかわ |
| 19 | 4 | 4 | 15/19 | 脱マスク、ザッケローニ氏が自宅で転倒、<br>ロシアのウクライナ侵攻、トルコ大地震 |
| 20 | 8 | 4 | 11/20 | 仮想通貨、メタバース、コロナウイルス、PS5、<br>日経平均株価、ダウ平均、ウクライナ、ロシア |
| 21 | 10 | 5 | 11/19 | 中華料理，宅配ピザ，回転寿司，オンラインゲーム，エミー賞,<br>アカデミー賞，新作アニメ，ノートパソコン，季節家電，節電 |
| 22 | 4 | 4 | 12/20 | 喫茶店，廃墟，ホラー，和菓子 |
| 23 | 10 | 5 | 8/20 | 出前，とんかつ，鍋料理，スニーカー，筋トレ，手抜き料理,<br>ガジェット，宇宙旅行，クラフトビール，引っ越し |
| 24 | 4 | 3 | 8/16 | 面白い動画,Twitter, 怖い話, 値上がり |
| 25 | 2 | 1 | 14/20 | 英語, 留学 |

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．

## 文　　献


[1] 潘健太, 北山大輔. 説明性向上のためのユーザレビューを用いた観光スポットの対応付け手法. 情報処理学会論文誌データベース (TOD), Vol. 13, No. 1, pp. 1–7, 2020.

[2] 山田純平, 北山大輔. Web 検索結果における閲覧効率化のための分散表現を用いた既知度予測手法. 研究報告情報基礎とアクセス技術（IFAT), No. 17, 2019.

[3] 塩川隼人, 岡田心太朗, 韓炳材, 廣花智遥, 宇津呂武仁, 河田容英, 神門典子. 深層学習を用いた学術用語解説ウェブページの分かり易さ・見易さの自動評定. 第 11 回 DEIM フォーラム論文集, Vol. 2019, , 2019.

[4] 甲谷優, 岩田具治, 塩原寿子, 藤村考. Qa コミュニティにおける複数情報源を用いた効果的な質問推薦. 情報処理学会論文誌データベース（TOD）, Vol. 3, No. 4, pp. 34–47, dec 2010.

[5] Marialena Kyriakidi, Georgia Koutrika, and Yannis Ioannidis. Recommendations as graph explorations. In *Fourteenth ACM Conference on Recommender Systems*, pp. 289–298, 2020.

[6] 真野宗和, 小尻智子. 検索リテラシ育成のための検索結果可視化システム. 関西大学インフォメーションテクノロジーセンター年報: IT センター年報, No. 11, pp. 13–25, 2020.

[7] 池田将, 牛尼剛聡. Twitter の反応を用いたニュース全体像の理解支援のための可視化手法. 研究報告データベースシステム (DBS), Vol. 2019, No. 5, pp. 1–6, 2019.

[8] 片岡雅裕, 橋山智訓, 田野俊一. 情報推薦システムにおいて閲覧する情報の偏りを気付かせる ui の設計. 日本知能情報ファジィ学会 ファジィ システム シンポジウム 講演論文集, Vol. 31, pp. 350–353, 2015.

[9] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, Vol. abs/1908.10084, , 2019.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, Vol. abs/1810.04805, , 2018.