ユーザの未訪問ページ予測のための拡張スニペットによる検索支援手法

山田 純平[1,a) 北山 大輔[1,b)

概要：近年，インターネットや情報端末の普及により，検索エンジンを利用した Web 検索の機会が増えている．これまでにも，Web 検索をより便利にするための研究は数多くされてきた．しかしながら，効率の良い Web 検索をできないことは多々ある．そこで我々は，検索結果から無造作に閲覧していると訪問済みページと同じような内容で効率よく新しい知識や必要な知識が得られない問題に着目した．この問題に対して，検索結果内の未訪問ページに対し，その内容を予想できるようなキーワードを提示し必要な情報が記載されていそうな Web ページを選択しやすくすることで解決する．本稿では各検索結果に対しそのスニペット文の中から 2 種類の特徴語を抽出する．1 つ目は Web ページのトピックを表す詳細語である．2 つ目は，Web ページを見ることで新たに得られそうな知識を表す未習得語である．それぞれ，単語の分散表現を用いたスニペット文中の単語のクラスタリング，訪問済みページ中の単語のクラスタリングに基づいて抽出する．これらの特徴語を各検索結果に提示する拡張スニペットを提案する．

1. はじめに

近年，インターネットや情報端末の普及に伴い，情報検索がより身近なものになった．中でも Web 検索は情報検索の代表例である．Web 検索では一つの検索クエリから多くの検索結果が提示される．その中でユーザは数多くの検索結果から必要な情報を効率よく閲覧したい．しかしながら実際は，すでに閲覧したページ（訪問済みページ）の内容と同じようなページや，ユーザにとって新しい情報，欲しい情報がないページを閲覧してしまい検索時間や労力を無駄にしてしまう．

そこで我々はこの問題に対し，検索結果内の各スニペット文から 2 種類の特徴語を抽出し提示することで解決する．本稿では，その Web ページの内容を表す "詳細語" と，ユーザがその Web ページからまだ習得できそうな内容を表す "未習得語" を提示する手法を提案する．これら 2 種類の特徴語を検索結果ページに提示することでユーザは，Web ページを開いて確認する前に内容を予想しやすくなり，自分の目的に合った結果ページを選択しやすくなると考える．

本稿では実際に拡張スニペットを実装し検証実験を行った．被験者は提示されたタスクに従って探索的検索を行い，その検索行動ログから拡張スニペットがあることで効率的に必要としている情報が載った Web ページを選択できているか調査する．

本稿の構成は，以下の通りである．2 節ではこれまでの検索支援に関する研究と，本研究との相違点について触れる．3 節では提案手法の詳しい説明を述べる．4 節では実際にシステムを実装し検証した結果から，本手法の特徴について議論する．最後に 5 節ではまとめと今後の展望について述べる．

2. 関連研究

これまで Web 検索行動における支援手法について多くの研究がされてきた．その中にはユーザに検索クエリを提示する研究 [1][2] や，関連キーワードを提示する研究 [3] などがある．渡辺ら [4] は閲覧中の Web ページに関連する情報の検索を支援する研究をしている．閲覧ページの関連情報を得たい時にキーワード検索やニュース検索，Wikipedia 検索の機能を用いて数回のタッチ操作で調べることができる手法を提案している．望月ら [5] は検索結果におけるランキング変動に着目し，ユーザの求めるキーワードを提示する手法を提案している．このように，推薦キーワードを提示して検索支援を行う研究は多く存在するが，提案手法のように検索結果の選択を支援する目的の研究は少ない．

検索結果をリランキングする研究 [6] も存在するが，ユーザはリランキングされた結果から選択する必要があるのは変わらない．本研究では，提示された検索結果からユーザ

[1] 工学院大学大学院工学研究科情報学専攻
a) em19024@ns.kogakuin.ac.jp
b) kitayama@cc.kogakuin.ac.jp

が選択する際に助けとなるような Web 検索支援手法を検討する．

Web 検索が日常化する中で検索効率向上のための研究も行われている．阿部ら [7] は，検索結果に出てきにくい Web ページなどに必要な情報がある場合，複数リンクをたどるなど余計に時間がかかる問題に対して，Web ページ内のリンク数などの HTML 構造を利用したクラスタリングを行い，対象とする検索クエリについてスニペットを自動生成する手法を提案している．白川ら [8] は検索スニペットに形態素解析を用い，テキスト解析時に同一格フレームに存在するか否かを判別することにより，関係抽出の精度や網羅性向上を目的とした手法を提案している．この研究ではスニペットを用いることで，Web ページから判別するより処理時間を減らすことができた．本研究でもスムーズな検索環境を維持した上で検索結果に抽出した特徴語を提示したいことから，スニペットを利用した内容予測をしている．

早乙女ら [9] はウェブブラウザにおける既読コンテンツ検出とその表示手法について研究している．この研究では既読コンテンツ検出にコサイン類似度を用いているが，本研究では蓄積した訪問済みページ内でクラスタリングを行い，含まれるトピックを考慮している点で異なる．

ユーザと検索タスクの満足度の関係性に着目し分析した研究なども存在する．梅本ら [10] は検索専門性と事前知識と検索タスクの満足度について研究している．評価実験では，検索専門性と事前知識があるユーザには必要な情報だけをより判断しやすい付加情報を添えて提示し，逆に検索専門性と事前知識がないユーザには多くの検索結果を提示することが満足度向上につながると確認できた．本研究では検索結果を提示しつつ，検索結果に対して付加情報として特徴語を提示することから，両者の満足度向上が望めると考える．

3. 拡張スニペット

拡張スニペットは各検索結果に対してスニペット文から抽出した 2 種類の特徴語で構成される．これら “詳細語”と“未習得語”の 2 種類の特徴語について，その抽出手法を交えながら詳しく説明する．なお，すべての語は Word2Vec 等による単語の分散表現が得られているものとする．

3.1 特徴語抽出

本稿ではスニペット文や Web ページ本文から特徴語を抽出する際に，Dinghan ら [11] が提案している文章特徴ベクトル表現手法である SWEM を応用している．

SWEM の 1 つである SWEM_max は文中の単語に対し，Word2Vec 等で分散表現を得て，その単語群の各次元の最大値を文章のベクトルとする手法である．この最大値を採用する方法では，各次元の値がどの単語由来のものなのか特定することが可能である．本稿では負の値も特徴を表すと考え，絶対値が最大になるものを採用している．そこで我々は，多くの次元で採用された単語が，文章の特徴語になると考え，閾値以上の採用次元数となるものを特徴語とした．

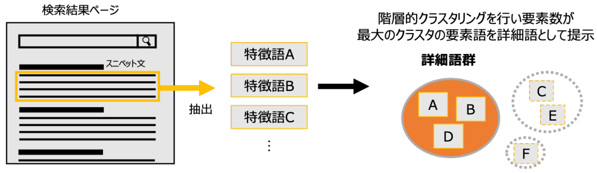

図 1　詳細語抽出手法

3.2 スニペット文からの内容予測のための詳細語

本節では，Web ページの内容を表す “詳細語”を抽出する手法について説明する．本研究ではスニペット文から特徴語を取得する．これは Web ページの本文を直接解析できることが望ましいが，検索結果すべてにアクセスするのは処理時間として現実的ではない．そのため，スニペット文から内容を表す特徴語を抽出することにした．なお，本稿では Google 検索で生成されるスニペットを用いた．

我々はスニペット文に含まれる単語集合に対しクラスタリングを行なった結果生成されたクラスタが，それぞれ文章のトピックを表すと考えた．また要素数が最大となるクラスタを，ある文章の中心トピックを表すと考えた．詳細語抽出手法の概要を図 1 に示す．

スニペット文とそのタイトルから形態素解析を行い，一般名詞，固有名詞を抽出する．さらに，3.1 節で説明した手法で特徴語を抽出した．この特徴語集合に対し用意した単語ベクトルモデルを用いてコサイン距離で階層的クラスタリングを行なう．このクラスタリング結果の中から要素数が最大のクラスタに属している特徴語を “詳細語”とした．

3.3 訪問済みページを用いて予測する未習得語

本節では，検索結果の各スニペット文から，ユーザの訪問済みページ本文に出現する特徴語を用い，Web ページを見ることでまだ習得できそうな内容を表す “未習得語”を抽出する手法について説明する．

我々は訪問済みページの特徴語集合に対してクラスタリングで生成された各結果が，ユーザの習得しているトピック分布を表していると考えた．そして，各クラスタの要素数をトピックの習得量と考え，習得量が少ないトピック，すなわち要素数が少ないクラスタにユーザがまだ習得していない内容を表す特徴語があると考えた．未習得語抽出手法の概要を図 2 に示す．

訪問済みページ本文からスニペット文から抽出するのと同様の 3.1 節の手法で特徴語を抽出した．今回用いている絶対値が最大値となるものを採用する SWEM は文章長が

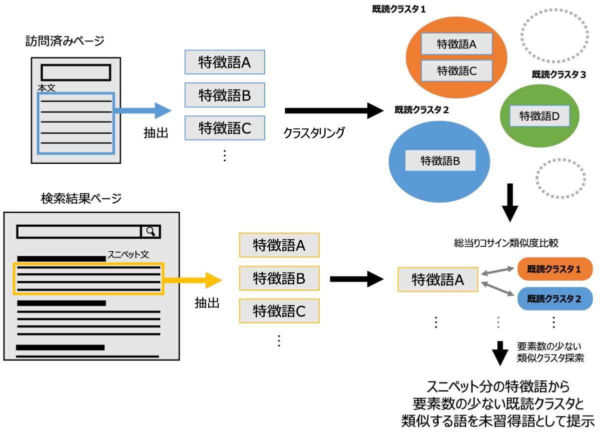

図 2　未習得語抽出手法

図 3　検索結果ページ提示例

長くなるにつれ，ノイズを拾いやすくなることが考えられる．そこで，句点で区切った画分で算出し，採用次元数の平均値を用いて特徴語を抽出した．次に抽出した特徴語について階層的クラスタリングを行い，その結果で得られたクラスタ群を既読クラスタとした．この既読クラスタとスニペット文の各特徴語を総当りで類似度比較した．そして，要素数が少ない既読クラスタとコサイン類似度が高いスニペット文の特徴語を“未習得語”とした．ここでの類似度比較とは，クラスタリングでも用いた単語ベクトルモデルを使って求めたコサイン類似度の平均値で比較している．

3.4　提示例

前述の方法で得られた“詳細語”と“未習得語”は検索結果ページに提示する．実際の提示例を図 3 に示す．

この Web ページは景色について内容を含んでいるため、青枠で示した「景勝地」や「眺め」、「景色」といった“詳細語”が提示されている．また，訪問済みページからユーザにとって歴史やゆかりなどのトピックに関する内容をもっと知ることができるとオレンジ色の枠で示した“未習得語”からわかる．さらに，両方の意味を併せ持つ遊覧船や四季といった内容は青枠で下線だけオレンジ色に示し，ユーザにとってまだ習得できそうなこのページの中心的な内容であることを示している．

4.　評価実験

今回の実験では実装した拡張スニペットが提示される検索結果ページと提示されないページを用意し，クラウドソーシングサービスで被験者を募った．被験者はどちらかのシステムを使い，我々が用意した検索タスクに従って探索的検索を行う．実際に提示した検索タスクを図 4 に示す．

被験者の閲覧履歴データは検索行動ログとして保存し，検索時間，クエリ入力回数，閲覧ページ数を比較し，拡張スニペットの効果を評価する．

4.1　実験システム

今回の実験ではユーザのブラウジング時に提案手法である拡張スニペットを生成するサーバを中継させて実現した．フロントエンド部とサーバサイド部に分けて説明する．

4.1.1　フロントエンド部

フロントエンドはブラウザで Google Chrome 環境を想定している．被験者は実験開始ページにユーザを認識するための ID を入力し，最初の検索キーワードを入れて実験を開始する．この実験開始ページから遷移した検索結果ページ，閲覧ページはリンクがすべて実験サーバへリダイレクトするように書き換えてあり，常に Web ページの URL やユーザ ID，手法（提案手法あり，または提案手法なし）が送信されている．そして，サーバサイドから返信される書き換えられた HTML 文を表示している．また，今回「戻る」ボタンによるページ遷移の際に検索結果ページが再読み込みされず，“未習得語”が表示されない問題に対して，10 秒ごとに自動更新するように設定した．

初回の検索結果ページでは訪問済みページが存在しないため，各検索結果にはスニペット文のみで抽出できる“詳細語”のみが提示される．

4.1.2　サーバサイド部

サーバサイドではユーザがリンクをクリックするごとに URL とユーザ ID，手法を受信する．検索結果ページでは，“詳細語”と“未習得語”が生成され，各検索結果のタイトルとスニペット文の間に挿入した HTML 文を返信している．各特徴語の抽出手法については 3 節で述べた手法で抽出する．

閲覧ページでは，既読クラスタ生成のために訪問済みページ本文として p タグと meta タグ内の description を取得する．これらを基に既読クラスタを生成する．また，訪問済みページ数が増えるとクラスタリングの対象となる取得した特徴語も増えることから，実装上の処理時間の都合上，各単語平均採用次元数の多い順に並べ上位 500 語をクラスタリング生成元の特徴語集合とした．また，要素数が少なすぎるクラスタについてはノイズであることが多いことから，要素数が 1 のクラスタは除外している．未習得語はスニペット文の特徴語とコサイン類似度が高く要素数が少ないクラスタであるため，今回はクラスタ要素数 10 以下でコサイン類似度 0.3 以上のクラスタから“未習得語”を採用している．

本実験では，学習済みの単語の分散表現として，Wikipedia

1. 検索タスク

今あなたは年末旅行に行く計画をしています。行ってやりたいこと、見てみたいことなど目的を5つ探してください。また、その時の旅行先を回答ください。（例：旅行先「松島」、目的「松島さかな市場で穴子丼を食べる」など）

図 4　検索タスク

表 1　実験結果

手法	平均検索時間 (m:s)	平均クエリ入力回数 (回)	平均閲覧ページ数 (ページ)
拡張スニペットなし	08:54	5.8	7.8
拡張スニペットあり	08:54	2.8	3.0

ダンプを，形態素解析器である MeCab に辞書「mecab-ipadic-NEologd」用いて，分かち書きし，fasttext で学習したものを用いた*1．また，SWEM での平均採用次元数数 11 以上を特徴語とした．階層的クラスタリングは群平均法を用い，コサイン距離は 0.7 を採用した．

4.2　検索行動ログの分析

拡張スニペットなしと拡張スニペットありでの検索時間，クエリ入力回数，閲覧ページ数の各平均値をそれぞれ比較した評価実験の結果を表 1 に示す．また，今回の実験では 1 分以上検索を行った検索行動ログを基に平均値を算出した．1 分以上検索を行った被験者数は拡張スニペットなしが 21 名，拡張スニペットありが 24 名であった．

今回取得できた検索行動ログでは，クエリ入力時から最後のクエリを入力するまでの平均検索時間はどちらも 8 分 54 秒だった．平均クエリ入力回数は拡張スニペットなしでは 5.8 回，拡張スニペットありでは 2.8 回であった．平均閲覧ページ数は拡張スニペットなしでは 7.8 ページ，拡張スニペットありでは 3.0 ページであった．よって，拡張スニペットなしに比べ拡張スニペットありの方が，クエリ入力回数と閲覧ページ数は減少傾向であることが確認できた．これは検索結果ページに拡張スニペットとして提示された “詳細語”と “未習得語”を見ることで，Web ページを閲覧せずに内容を把握できたと感じて終えてしまっていることが考えられる．また，従来のスニペット文より各検索結果ページに含まれるトピックごとの違いがわかりやすいことから，多くのトピックを網羅的に知ることができたと感じやすいことも考えられる．今回の実験では，同じような内容のページを閲覧するようなログは見られなかったが，閲覧ページ数が減少傾向にあったことからこの点についての影響の考察はできないと考えた．

今回の実装上の限界として，読み込み未完了時の行動ログが記録されないことがある．そのため，実際に被験者が素早く閲覧し内容を把握している場合でも，閲覧ページや検索結果ページにおいて読み込み完了時にログが記録されるためカウントされない．この解決策としては検索行動ログをサーバサイド部で記録せずにフロントエンド部で記録することや検索中の画面をキャプチャすることが考えられるが，今回は読み込み未完了の遷移については未閲覧として考えた．

4.3　アンケートによる有用性の調査

本実験では，拡張スニペットありの各被験者に対して検索終了後に，アンケート調査を実施した．具体的には，青枠で示した “詳細語”，オレンジ枠で示した “未習得語”，青枠にオレンジの下線で示した両方の特徴を持つ語についてそれぞれ，「適切だった」，「やや適切だった」，「不適切だった」の三段階の選択肢を設けた．また，拡張スニペットが助けになったかどうかについても被験者に尋ね，こちらは「助けになった」，「やや助けになった」，「助けにならなかった」の三段階の選択肢を設けた．

各アンケート回答の三段階評価を 1, 0, -1 に置き換え，平均値を算出した結果を表 2 に示す．結果として，青枠で提示された “詳細語”は平均 0.71 となり，適切だと感じた被験者が多かったことが確認できた．オレンジ枠で提示された “未習得語”や，青枠とオレンジ枠で表示された両方の特徴を持つ語についてはどちらも 0.4 以上で過半数以上の被験者が適切だと感じたことが確認できた．拡張スニペットが閲覧する Web ページを選ぶ助けになったかという質問に対しては，平均回答が-0.05 で若干ネガティブな印象を持たれたことがわかった．原因としては実装上の問題として一部の Web ページではページが表示されないエラーや，通常の検索時よりも読み込みに時間がかかりストレスを感じさせてしまったことが自由記述のコメント欄より確認できた．今後の検証実験では実装の完成度を更にあげ，より普段通りの Web 検索をしてもらうことが必要である．今回の調査については比較対象である拡張スニペットなしでも同じ実装環境であるため比較が可能であると判断した．

5.　おわりに

本研究では，ユーザが多くの検索結果から閲覧ページを選ぶ際に，訪問済みページの内容と同じようなページなどを閲覧することで時間や労力を無駄にしてしまう問題に対し，閲覧判断基準となる “詳細語”と “未習得語”の 2 種類

*1　https://qiita.com/Hironsan/items/513b9f93752ecee9e670

表 2 アンケート結果

質問	平均回答
青の枠で表示された単語は適切でしたか？	0.71
オレンジ色の枠で表示された単語は適切でしたか？	0.43
青とオレンジ色の枠で表示された単語は適切でしたか？	0.48
青やオレンジの枠で表示された単語は閲覧する Web ページを選ぶのに助けになりましたか？	-0.05

の特徴語で構成された拡張スニペットを提示することで解決する手法を提案した．

拡張スニペットを実装し，実際に拡張スニペットありとなしにおいて，検索行動ログをクラウドソーシングサービスで収集した．その検索行動ログから，拡張スニペットありでは拡張スニペットなしと比べ，クエリ入力回数と閲覧ページ数が減少する傾向にあることを確認した．また，検索終了時に実施したアンケート調査からは，全体的に被験者は "詳細語"に対して適切に提示されたと感じ，"未習得語"と両方の特徴を持つ語"については過半数以上の被験者が適切に提示されたと感じたことを確認した．拡張スニペットが検索結果ページから閲覧する Web ページを選ぶのに助けになるかという質問に対しては今回の実装環境では考察に至らなかった．

今後は，閾値の決定のための検証を重ね，最適な閾値設定を考えていく．また，今回処理時間の理由から訪問済みページから取得する特徴語数に制限をかけていたが，この制限方法についてもユーザの知識習得量を正確に示せるような手法を考えて行きたいと考えている．拡張スニペットの Web 検索におけるユーザの行動への影響や，その効果の検証についてもより実用的なシステムとして実装し，実施していきたいと考える．

謝辞

本研究の一部は，2019 年度科研費基盤研究 (C)(課題番号：18K11551) 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] 鈴木永史郎, 杉本徹. 意外性のある検索クエリの推薦方法の提案. 第 78 回 情報処理学会 全国大会講演論文集, No. 1, pp. 503–504, 2016.

[2] 大石哲也, 倉元俊介, 峯恒憲, 長谷川隆三, 藤田博, 越村三幸, 堀憲太郎. 関連単語抽出アルゴリズムを用いた web 検索クエリの生成. 情報処理学会研究報告, No. 56(2008-DBS-145), 2008.

[3] 小野謙太朗, 立澤祐樹, 岡誠, 森博彦. Web での特徴語と共起する語を用いた未読ページからのキーワード推薦. 第 79 回 情報処理学会 全国大会講演論文集, No. 18, pp. 1 – 5, 2015.

[4] 渡辺奈夕子, 岡本昌之, 菊池匡晃, 飯田貴之, 佐々木健太, 堀内健介, 山崎智弘, 大村寿美, 服部正典. 閲覧 web ページからの第 1 検索キーワード抽出に基づく検索支援. 情報処理学会論文誌, Vol. 53, No. 7, pp. 1783–1796, 2012.

[5] 望月祐臣, 東基衛. Web 検索結果におけるランキング変動に着目したキーワード支援システム. 第 70 回 情報処理学会 全国大会講演論文集, pp. 493–494, 2008.

[6] 浩二倉門, 哲也大石, 隆三長谷川, 博藤田, 三幸越村. Wikipedia のリンク共起とカテゴリに基づくリランキング手法. 研究報告データベースシステム（DBS）, Vol. 2010, No. 12, pp. 1–8, jul 2010.

[7] 寛之阿部, 雅文松原, Chakraborty Goutam, 浩司馬淵. Html 文書構造を利用した web 検索結果クラスタリング手法の有効性について. 第 79 回全国大会講演論文集, 第 2017 巻, pp. 547–548, mar 2017.

[8] 謙太朗小野, 誠岡, 博彦森. Web での特徴語と共起する語を用いたキーワード推薦. 第 79 回全国大会講演論文集, 第 2017 巻, pp. 543–544, mar 2017.

[9] 高大早乙女, 仁相田. ウェブブラウザにおける既読コンテンツの検出・表示手法の検討. 研究報告情報基礎とアクセス技術（IFAT）, 第 2017-DC-104 巻, pp. 1—8, mar 2017.

[10] 和俊梅本, 岳洋山本, 克己田中. 検索専門性と事前知識に着目した検索行動とタスク満足度の関係性分析. 情報処理学会論文誌データベース（TOD）, Vol. 7, No. 4, pp. 13–28, 2014.

[11] Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. *CoRR*, Vol. abs/1805.09843, , 2018.

