視線情報と単語の分散表現を利用したWebページ関連語提示システムの提案

水野 翔太[1,a)]　寺田 実[1,b)]

概要：Web ページ閲覧によって情報探索を行っているユーザに対して，Web 上での情報探索を支援するために，Web ページ内の単語の頻出度と情報探索時の視線計測結果から重要単語を特定し，それを入力単語として分散表現を利用して類似単語を求め，ユーザが興味のありそうな単語を提示するシステムを試作した．

1. はじめに

ユーザは Web ページの情報探索を，理解が曖昧な状態で行うことがある．このような状況において，ユーザは求めている情報を得るまでに情報の探索と検索を繰り返し，徐々に探索意図と適合する情報を得る．このような方法では，ユーザが探索する情報の取捨選択を行うコストが発生する．

このような状況を改善するためには，ユーザがどのような意図で検索を行っているかを自動で推定し，必ずしも現在閲覧している Web ページに含まれていない単語をキーワードとしてユーザへ提示することで情報探索の支援を行うシステムが求められる．

例えば，プログラミング言語である「python」という単語に着目しているユーザがある Web ページを探索している場合を想定する．このとき，「python」に関連しており，Web ページ内には具体的に存在していない単語を何らかの手法により自動でユーザへと知らせることで情報探索の手助けにならないかと考えた．

2. 目的

Web ページ閲覧により情報探索を行っているユーザに視線情報やページ内の頻出単語情報から判断した「Web ページ内重要語」を割り出し，それを利用して Web ページ内には存在していない情報を含む関連語の提示を行うことにより，自動でユーザの情報探索の支援を行うシステムの構築が目的となる．

[1] 電気通信大学
[a)] m1931135@edu.cc.uec.ac.jp
[b)] terada.minoru@uec.ac.jp

3. 関連研究

3.1 ユーザの Web ページ閲覧と情報探索

ユーザの Web ページ閲覧時の視線情報と検索意図に関する研究が行われている．

Huang らの研究 [1] では，Web 検索時のユーザの視線位置とマウスカーソル位置の関連性に関して疑問が呈されており，ユーザの Web 閲覧時における着目箇所の特定のためにはユーザの視線情報の利用が重要であることが示唆されている．

戸田らの研究 [2] では，実験により，Web ページから情報を探索しているユーザは得たい情報のある近傍を探索し，そこが目的の情報であるか，もしくは目的の情報に近づけるかどうかの判断を繰り返し行い，ユーザの目的とする情報と一致するかの判断時に視線の停留時間が長くなることが示された．

これにより，Web ページ閲覧時の視線情報がユーザの興味の抽出に有益であり，視線の停留時間によってユーザの興味の強い情報の取得可能性が示唆されている．

3.2 視線情報からの注目語抽出に基づく検索意図のリアルタイム推定 [3]

「Web ページ中でユーザがよく見ている語は，ユーザの検索意図に対する適合度が高い」と仮定し，Web 検索時におけるユーザの注目語からの検索意図のリアルタイム推定を行った．評価の結果，Web 中に何度も出現していて，かつユーザの注目頻度が高い単語に対して高い注目度の値を与えたモデルの性能が良いことが示された．

本研究ではこの先行研究による注目語抽出手法により求めた単語に対して，分散表現から類似単語集合を求めている．このとき，注目箇所近傍においても有益な情報が存在

しているとの考えから，単語単位ではなく，ある文章集合を単位として視線計測を行い，得られた注目語に新たに分散表現を適用してユーザの興味の強い単語候補の提示を行う．

4. 提案システム

4.1 システムの概要

提案システムはユーザが Web ページ閲覧に利用する Web ブラウザ上での処理を行うフロントエンドと視線情報の計測および Web ページ内のテキストデータの解析を行うバックエンドで構成される．

ユーザの Web ページ閲覧を動的に解析する前に，予めコーパスを用意し，単語の分散表現を作成する．

ユーザの Web ページ閲覧が開始されると，最初にフロントエンドからバックエンドへ現在閲覧している Web ページのテキスト情報が送られてくる．また，バックエンドで視線情報取得が開始され，リアルタイムにフロントエンドへと送られてくる．この視線情報により現在ユーザが着目している Web ページの箇所を特定し，着目時間や着目箇所情報を逐次バックエンドへと送信する．バックエンドでは，送られてきた解析データから，「Web ページに出現する頻度」および「Web ページを閲覧しているユーザの注目頻度」が高い「Web ページ内重要単語」を算出する．ユーザはキー押下により Web ページ閲覧終了をシステムに知らせる．

ユーザの Web ページ閲覧が終了すると，「Web ページ内重要単語」と後述するモデルによって最終的な提示候補となるスコア上位の単語をバックエンドで算出し，フロントエンドへと結果を送信して Web ページへと埋め込む．

最後に，最終的な提示結果を Web ページ上でユーザが確認する．

図 1 にシステムの全体像を示す．

4.2 単語の分散表現の取得

主にコンピュータ技術に関する記事が一般ユーザによって投稿され共有されている Qiita[*1] を対象に，Qiita に投稿されている 2019 年 12 月 31 日以前の全記事を API により取得し，その本文のみを抽出し整形してコーパスとする．本文データのサイズは 1GB 弱ほどになった．これをデータセットとして gensim[*2] に実装されている word2vec[4] を利用して単語の分散表現を作成する．

作成した分散表現を利用することにより，例えば単語間で cos 類似度を計算することによって単語間の関連性を計算することが出来るようになる．すなわち，入力単語に似ている単語群が結果として求まる．

表 1 に類似語を求めた例を示す．

[*1] https://qiita.com
[*2] https://radimrehurek.com/gensim/

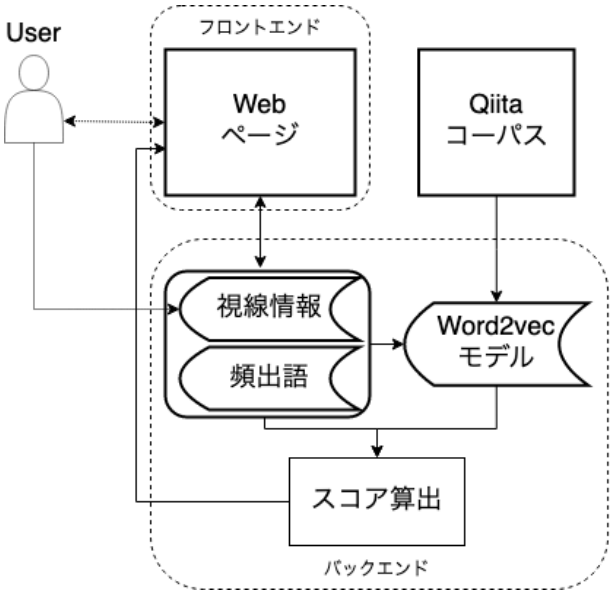

図 1 提案システムの全体像

表 1 word2vec を利用した単語の類似度計算の例

入力単語	出力単語	類似度
python	pip	0.7039
	venv	0.6364
	anaconda	0.6313
	distutils	0.6257
	virtualenv	0.6093

4.3 Web ページ上のテキストの解析

DOM で表されたツリー状の Web ページ構造において，XPath でテキストノードを指定し全てのテキストノードを要素として取得する．このテキストノードは親ノードとして p 要素などを持っており，それを着目単位とする．

図 2 の例では，中央下の p 要素の子要素であるテキストノードは同一の親ノードを保持しているために，合わせて 1 つの着目単位としている．

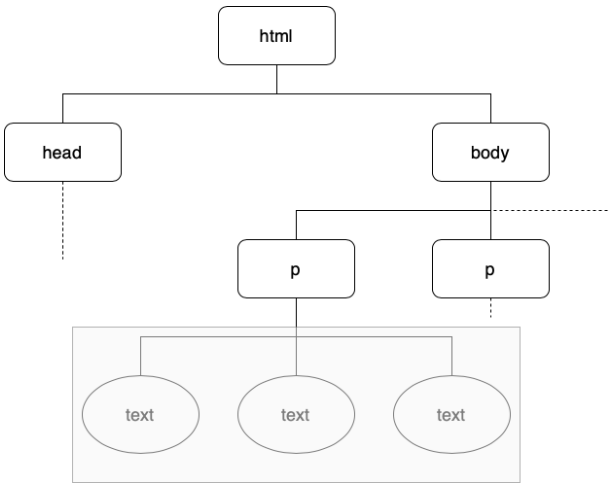

図 2 着目単位の決定イメージ

4.4 Web ページ閲覧時の視線情報の取得と要素選択

Tobii pro ナノを利用しバックエンドで視線情報を計測する．WebSocket により双方向通信経路を確保し，フロントエンドに視線情報を逐次送信する．Tobii により計測した視線情報はディスプレイ領域の左上が (0, 0), 右下が (1, 1) となる座標系に正規化されて取得されているので，Web ブラウザの画面上における位置やドキュメント表示領域等を考慮し Web ブラウザ上の座標系に変換する．

図 3 では，ディスプレイ全体の中に Web ブラウザのウィンドウが表示されており，その中でディスプレイ基準の視線座標 (0.3, 0.4) から，実際に Web ページが表示される Document 領域基準の視線座標 (300, 250) へと変換する例を示している．

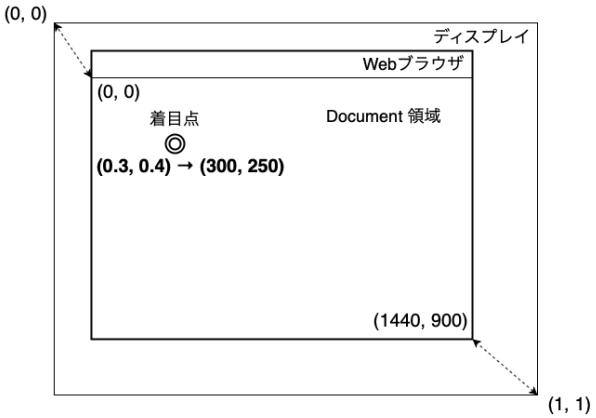

図 3 Web ブラウザ上での視線情報の補正

Web ページ上で現在着目している視線ポインタを前述の変換により求め，要素の選択を行う．この際にテキストノードが p 要素などの子要素として DOM で表されている場合，通常では一番上の親要素だけしか取得できないので，その親要素にぶら下がっているテキストノードの表示領域に現在の視線ポインタが含まれているかをそれぞれ確認し，どのテキストノードを選択しているのかを特定する処理を行っている．これにより，より詳細なテキストノード処理を行うことが可能となる．

図 4 に現在選択されているテキストノードの決定例を示す．着目点は Text Node 1 の領域内の座標であるため，この場合は Text Node 1 を選択する．

4.5 スコア算出

「ページに頻出の単語かつユーザが着目している単語はユーザの検索意図との適合度が高い」という考えを反映させたスコアモデル $SCORE_{MGT}$[3] がある．

$$SCORE_{MGT} = gf \cdot tf \tag{1}$$

- gf
  ユーザの注視頻度を表しており，この値が高いほど頻繁にユーザがその近傍で情報を探索している．
- tf
  閲覧している Web ページ内でのストップワードを除いた単語の頻出度である．

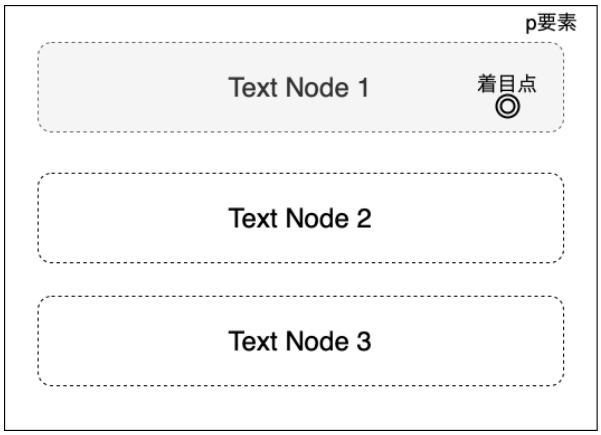

図 4 着目しているテキストノード の決定

このモデルにより算出されたスコア上位の単語群を入力として，それらの分散表現における類似度の高い単語を求める．そして，「ページに頻出でありユーザの着目している単語と似ている語はユーザの興味が強い」と仮定し，新たに以下のスコアモデル $SCORE_{SIM}$ を定義する．

$$SCORE_{SIM} = SCORE_{MGT} \cdot st \tag{2}$$

- st
  分散表現により求めた類似度を表している．この値が高いほど，同一の文脈で利用されている．

この $SCORE_{SIM}$ モデルを使用して，「単語の出現頻度」，「ユーザの着目度合い」，「分散表現による類似度」をパラメータとしたスコアを算出し，スコア上位単語群を最終的なユーザへの提示候補とする．

4.6 結果の提示

今回のシステムでは (入力単語, 類似単語) が結果として送られてくるが，入力単語は Web ページ内に存在する重要単語である．そこで，提案システムで推薦された重要単語が含まれた着目単位に印をつけてユーザページ内の重要部位を確認出来るようにする．ユーザがその重要部位をマウスオーバーすることにより，重要単語により拡張された類似単語をツールチップで確認できるようにする．

図 5 の提示例では，提示候補となる拡張された類似単語を求めた「ページ内重要単語」が含まれている着目単位全体が赤で着色されており，ツールチップで着目単位付近をマウスオーバーすることにより拡張単語が表示されていることが確認できる．

図 5　ツールチップによる提示語の表示 [5]

5. おわりに

本研究ではユーザの Web ページ閲覧を想定した視線情報と分散表現を利用した情報探索補助システムを提案した．今後は，実際にどれほどユーザの意図に適合した関連語の出力が出来ているかを頻出度のみを根拠に拡張した単語群との比較等を行い評価していきたい．

参考文献

[1] Jeff Huang, Ryen White and Georg Buscher.: User See, User Point: Gaze and Cursor Alignment in Web Search, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1341–1350, 2012.
[2] 戸田航史，中道上，島和之，大平雅雄，阪井誠，松本健一：Web ページ閲覧者の視線に基づいた情報探索モデルの提案，情報処理学会研究報告．HI，ヒューマンインタフェース研究会報告，Vol．2005, No．52，pp．35–42，2005．
[3] 梅本和俊，山本岳洋，中村聡史，田中克己：視線情報からの注目語抽出に基づく検索意図のリアルタイム推定，情報処理学会論文誌データベース（TOD），Vol．6，No．3，pp．120–131，2013．
[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado and Jeff Dean.: Distributed representations of words and phrases and their compositionality, Advances in Neural Information Processing Systems, Vol. 26, pp. 3111–3119, 2013.
[5] Python - Wikipedia. https://ja.wikipedia.org/wiki/Python