Vol.2015-GN-96 No.18
2015/10/3

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

小野謙太朗[†1]　立澤祐樹[†2]　岡誠[†3]　森博彦[†1]

近年，Web 上には膨大な量の Web ページがある．ユーザが検索を行うと多くの Web ページがヒットしてしまい,適切な検索キーワードをユーザが選択するのは困難である.その際にユーザが表示した検索結果の中でユーザが閲覧しなかったリンクの Web ページにユーザの必要としている検索キーワードとなる単語が存在する可能性がある．そこで，本研究ではユーザの検索意図に関する単語であり，ユーザ自身が未確認の単語の推薦を行うために，ユーザの検索意図に関する単語を未読の Web ページから推薦する.その方法として,ユーザの閲覧行動から抽出した検索意図に関係する単語の特徴語と共起する単語を未読の Web ページから抽出してキーワード推薦を行う手法を提案する. 提案システムを用いて評価を行なった結果, 推薦キーワードはある程度ユーザにとって役に立つキーワードであったと示す事が出来,未読の Web ページにユーザの必要とする単語が存在する可能性を示した.

Keyword recommendation from the unread pages using words that co-occur with the characteristic words of the Web

KENTAROU ONO[†1]　YUKI TACHIZAWA[†2]
MAKOTO OKA[†3]　HIROHIKO MORI[†1]

Recently. There is a vast amount of web pages on the web. In order to the search results hit too many web pages, it is difficult for user to select the search keywords that appropriate search results to hit.In that case, there is a possibility that appropriate search key words exist on the web page which has not been accessed by the user. Therefore, in order to recommend words of yourself unconfirmed on the user's search intention, this study recommend the word about the search user's intention from unread web pages. As the method, we propose a method for recommending keywords to extract words which co-occur with the characteristic words which related to the search intention from the viewing behavior of the user from the unread web pages. The results were evaluated using the proposed system, the recommendation keyword can be shown keywords useful for the user. We show a possible word which requires the user to unread web pages.

1. はじめに

近年, パソコンやスマートフォンの普及により Web での情報検索が頻繁に行われている．情報検索ではユーザが適切なキーワードを選択することが重要となる．

また，ユーザが入力した検索キーワードによって表示された検索結果において目的の情報が見つからない場合，ユーザは新たなキーワードで検索を行う．その場合，様々な Web ページから情報を取得し，その情報の中から次の検索キーワードとなる単語の発見は自分の欲しい情報が明確でないほど難しい．さらに，ユーザは検索結果のタイトル・スニペットを確認して検索結果にあるリンクの Web ページがどのようなものか判断を行うため，検索結果に表示されたリンクの中でユーザが未読のリンク先の Web ページにはタイトル・スニペット以外のユーザが未確認の情報が多く存在する．

そこで，本研究ではユーザの検索意図に関する単語であり，ユーザ自身が未確認の単語の推薦を行うために，ユーザの検索意図に関する単語を未読の Web ページから抽出してキーワード推薦を行う手法を提案する．

2. 関連研究

検索キーワードを探すことを支援する研究として望月ら[1]の研究がある．この研究はユーザが検索結果に表示されている Web ページから適合・不適合ページを選択し，検索結果におけるランキング変動に着目することで，判定した Web ページから推薦キーワードを提示するシステムを提案している．しかし，この研究ではユーザが既読の Web ページから推薦キーワードを提示しているため，ユーザが既に確認しているキーワードを再度提示してしまう可能性がある．このような既読の Web ページから検索支援を行う研究[2] [3]は多数存在する．しかし，ユーザが未読の Web ページから検索支援を行う研究は少ない．

3. 研究目的

ユーザが検索を行う際に，検索結果の一覧に表示されている上位の Web ページしか見ない場合，検索結果の上位ではない閲覧しなかった Web ページにユーザの必要としている検索キーワードとなる単語が存在する可能性がある．また，ユーザは検索結果のタイトルとスニペットを確認し

†1 東京都市大学大学院工学研究科
Graduate School of Engineering, Tokyo City University Graduate School
†2 (株)プラスアルファ・コンサルティング
Plus Alpha Consulting
†3 東京都市大学知識工学部
Faculty of Knowledge Engineering, Tokyo City University Undergraduate Division

て Web ページの内容を予測していくが，スニペットには検索キーワードが書かれている文章が中心に表示されており，文章の要約が含まれているわけではない．よって，ユーザの閲覧しなかった Web ページのタイトルとスニペット以外の文章にユーザが必要としている未確認のキーワードが存在する可能性がある．

そこで，本研究では閲覧行動から抽出した検索意図に関係する単語を未読の Web ページから抽出してキーワード推薦を行うことを目的とする．

4. 提案手法

本研究では何度も検索を繰り返し，様々な Web ページを閲覧していく様子を観察できる調査学習型検索タスクを対象とする．

ユーザの動作と提案システムの動作を図 1 に示す．システムはユーザが Web ページを確認して検索結果に戻る際に動作する．システムは既読の Web ページからユーザの検索意図を抽出し，ユーザが未確認の情報が存在する可能性のある未読の Web ページから検索意図に関するキーワードを推薦する．ここからはシステムの詳細な説明を行う．

4.1 ユーザの検索意図に関する特徴語抽出

4.1.1 タイトル・スニペットからの特徴語抽出

ユーザは検索結果のタイトル・スニペットを確認して Web ページの内容を予測するため，ユーザがクリックしたタイトル・スニペットにはユーザの検索意図が含まれていると考えられる．そこで，ユーザがクリックした動作から検索意図を取得するために，ユーザが既読の Web ページのタイトル，スニペットから単語を抽出する．タイトル・スニペットからの単語抽出方法を図 2 に示す．システムは実線で囲まれているユーザがクリックしたリンクのタイトルとスニペットの単語の抽出を行う．そして，その単語の中から点線で囲まれているクリックしていないリンクのタイトルとスニペットの単語を除いて，単語と出現頻度をデータベースにある単語と出現頻度に加えて保存する．

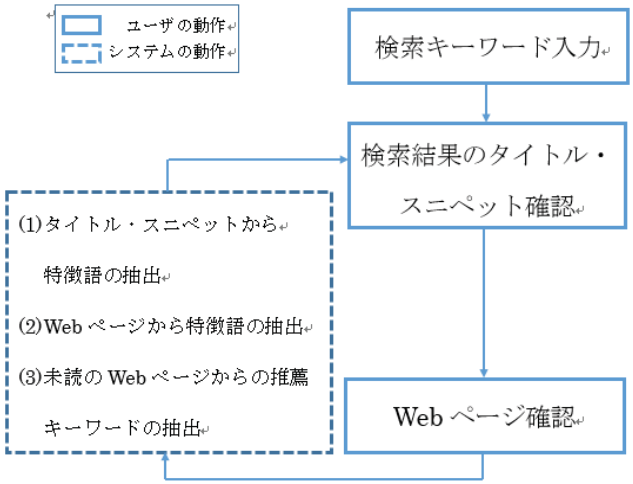

図 1　システム構成

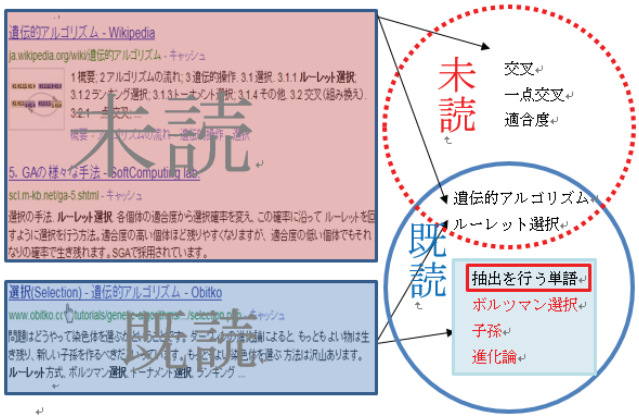

図 2　タイトル・スニペットから単語抽出

4.1.2 Web ページからの特徴語抽出

ユーザの Web ページ閲覧行動から検索意図を取得するために，4.3.で述べる有益判定アルゴリズムを用いて既読の Web ページの中でシステムがユーザにとって有益だと判断したリンク先の Web ページのテキストデータから単語を抽出し，単語と出現頻度をデータベースにある単語と出現頻度に加えて保存する．そして，データベースに保存した単語の中で出現頻度が上位の名詞と英数字を推薦キーワード抽出のためのユーザの検索意図である特徴語とする．

4.2 未読の Web ページからの推薦キーワード抽出

ユーザが未確認の検索意図に関する単語の推薦を行うために，検索結果の中でユーザが未読の Web ページのテキストデータから特徴語と共起する単語を推薦キーワードとする．特徴語を推薦キーワードに含まない理由としてはユーザがリンククリックした際に，既にユーザが確認していると考えられるからである．

4.3 Web ページ判定処理

提案手法の 4.1.2.で使用する Web ページがユーザにとって有益であるかどうかの判定を行うアルゴリズムの提案とログのパラメータの最適値の調査実験を行った．

実験では検索課題を行ってもらい，5 種類のログ（操作停止時間， Web ページ滞在時間，なぞり読みしていた時間，マウス操作時間，スクロールバー操作時間）を取得し，それぞれ閲覧した Web ページが被験者にとって有益であるかどうか被験者に判定を行ってもらった．5 種類のログはユーザの閲覧行動を使用した南ら[3]の研究で収集されたデータから x^2 値による属性選択を行い，Web ページがユーザにとって有益かどうか判定するのに向いているとされたため採用した．

そして，取得した 5 種類のログのデータを属性とし，{有益，無益}の 2 種類を分類するクラスとして C4.5 アルゴリズムを用いた．その結果，図 3 のような決定木となった．この決定木から以下のような 3 種類のクラスに分かれ，いずれかのクラスを通ることで有益であると判断する有益判定アルゴリズムを提案する．

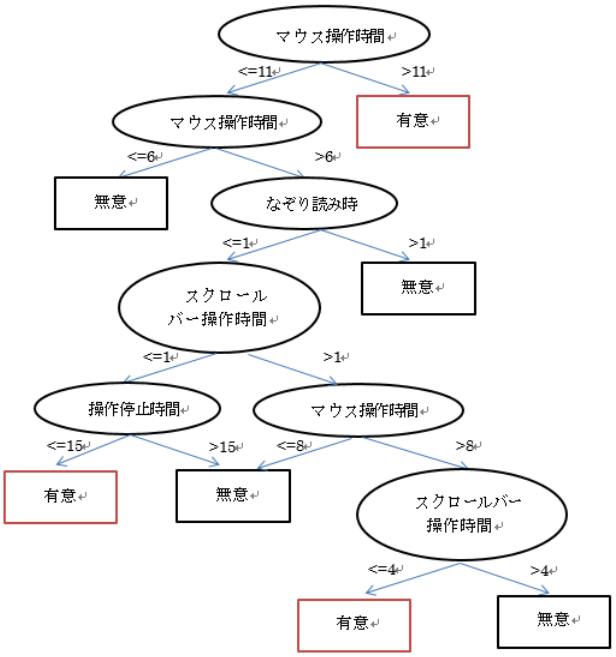

図 3　決定木

クラス 1 : マウス操作時間が長いページを取得するため，ユーザが詳しく見るページの取得を目的としている．

クラス 2 : マウス操作が長く，なぞり読みが短く，操作停止時間が長くなく，スクロールバー操作時間が短いページを取得するため，被験者が見ている Web ページの最初の方に目的の情報があるページの取得を目的としている．

クラス 3 : マウス操作時間が長く，なぞり読みが短く，スクロールバー操作が行われているページを取得しているため，ユーザがページ全体を閲覧して目的の情報を見つけられるページの取得を目的としている．

5. 実験

提案したシステムの評価実験を行った．

5.1 実験方法

被験者には検索課題として，定義の検索質問と理由の検索質問をそれぞれ 2 題ずつ行ってもらった． また，推薦キーワードを表示する状態と表示しない状態ではそれぞれ違う質問で実験を行った．

実験では推薦キーワードを実際に使用した場合の変化を見るために，推薦キーワードを表示した状態を 15 人，推薦キーワードを表示しない状態を 15 人の計 30 人の大学生・大学院生の被験者に行ってもらった．今回は推薦キーワードとして表示するキーワード数を未読 Web ページから抽出した単語の頻度が多い順に最大 10 個までに制限した．

被験者には少しでも役に立った Web ページをお気に入り機能に登録してもらった．また，それぞれの検索課題の解答の記入を行ってもらった．

5.2 評価方法

推薦システムによって提示された推薦キーワードを以下の 7 つで評価を行なった．

1. 有益な Web ページがヒットする推薦キーワード
2. 推薦キーワード表示／未表示状態の比較
3. 推薦キーワード使用後の推薦キーワードの変化
4. 推薦回数による推薦キーワードの比較
5. 推薦キーワードの使用率
6. 推薦キーワードによって新規の web ページがどの程度表示されるか
7. 推薦キーワード候補数と検索回数の推移

1〜7 を評価するために推薦システムによって提示された未読ページから抽出した推薦キーワードの評価を行った．

実験後，本研究者は推薦キーワードが表示された時点での検索キーワードとなった単語に推薦キーワードとして表示された単語を加えて&検索を行った．また，今回は&検索の検索結果上位 10 件を評価対象とした．さらに，検索結果の中で有益な Web ページかどうかの判断はユーザに実験中にお気に入りに登録してもらった Web ページを少しでも役に立った Web ページとし，その中でも実験の解答時に Web ページを開いて解答に使用されていた Web ページを確実に役に立った Web ページとした．

1.の評価を行なうために役に立った Web ページが表示出来る推薦キーワードがどれだけ存在するか適合率の計算を行った． また，2~5 の評価のための適合率の計算も同様に行った． 各推薦キーワードの評価のための適合率を（1）のようにして求めた．

6.の評価を行うために推薦キーワードの使用率を（2）のようにして求めた．

7.の評価を行うために新規の Web ページの表示率を（3）のようにして求めた．

5.3 結果

1.有益な Web ページがヒットする推薦キーワードの適合率の計算を行った結果を表 2 に示す．表 2 より，ユーザにとって役に立つ推薦キーワードが表示出来ていることが分かる．

(1) 各推薦キーワードの適合率=

$$\frac{\text{有益な Web ページが検索結果に表示されるキーワード数}}{\text{推薦キーワード総数}}$$

(2) 推薦キーワードの使用率=

$$\frac{\text{推薦キーワードの総使用回数}}{\text{推薦キーワードの総表示回数}}$$

(3) 新規の Web ページの表示率=

$$\frac{\text{ユーザが未確認の Web ページ数}}{\text{推薦キーワードによって表示された Web ページ数}}$$

2.推薦キーワード表示／未表示の適合率は表 3 のようになった．表 3 では少しでも役に立った Web ページでは推薦キーワード未表示状態よりも推薦キーワード表示状態の方が適合率は上がっている事が分かる．

3.推薦キーワード使用後の適合率は表 4 のようになった．表 4 では確実に役に立った Web ページの適合率が全体の推薦キーワードよりも高い事が分かる．

4. 推薦回数による推薦キーワードの比較を行うために推薦キーワードの最初のリストに表示される推薦キーワードと 2 回目以降に表示される推薦キーワードの適合率を表 5 に示す．推薦キーワードのリストの表示は一つの検索課題の中で最大 3 回までしか今回の実験では確認されなかった．2~3 回目の推薦キーワードの方が 1 回目の推薦キーワードよりも良い結果が出ていることが分かる．

5.推薦キーワード使用率は 24.6%と低かった．

6.推薦キーワードによって新規の Web ページがどの程度表示されるかの新規の Web ページの表示率を表 6 に示す．表 6 から推薦キーワードを使用することにより新規の Web ページが 7 割程出てくることが分かる．

推薦キーワード候補の合計数が調べて行く毎にどのような変化していくか調べるために 7.推薦キーワード候補数と検索回数の推移を図 4 に示す．図 4 より推薦キーワード表示・非表示状態共に検索毎に推薦キーワード候補の合計数が減少していることが分かる．

表 1　全体の推薦キーワードの評価

	少しでも役に立った Web ページ	解答に確実に役に立った Web ページ
適合率	0.580	0.502

表 2　推薦キーワード表示/未表示の結果

	少しでも役に立った Web ページ	解答に確実に役に立った Web ページ
キーワード表示	0.592	0.493
キーワード未表示	0.555	0.526

表 3　推薦キーワード使用後に表示された評価

	少しでも役に立った Web ページ	解答に確実に役に立った Web ページ
適合率	0.580	0.580

表 4　推薦回数の結果

	少しでも役に立った Web ページ	解答に確実に役に立った Web ページ
1 回目の検索	0.550	0.471
2～3 回目の検索	0.643	0.571

表 5　新規の Web ページの評価

	キーワード表示	キーワード非表示
新規の Web ページの表示率	0.746	0.771

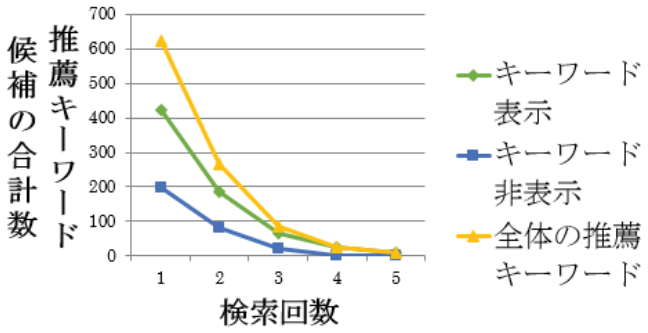

図 4　検索数と推薦キーワード合計数の推移

6. 考察

全体の推薦キーワードを評価した表 1，表 5 より，ユーザにとって役に立つ Web ページを検索結果一覧の上位に表示することが出来る推薦キーワードを表示できたことを示していると考えられる． しかし，表示された推薦キーワードにはユーザにとって役に立たないと感じるキーワードも表示されてしまっていることが分かる．また，表 2，表 3 の結果から推薦キーワード表示状態で推薦キーワードを使用することで，推薦キーワードの使えない未表示状態よりも目的の情報に少しでも役に立つ Web ページが見つけられることが考えられる．

表 4 の結果からは 1 回目の推薦キーワードより 2～3 回目の推薦キーワードの方がユーザに役に立つ Web ページを表示出来ることが分かる．これはユーザがより多くの Web ページを見ることにより，推薦キーワードの抽出に必要となる特徴語がユーザの検索意図に近い単語になっていっているのではないかと考えられる．さらに，図 4 から推薦キーワード表示・非表示状態共に検索毎に推薦キーワード候補の合計数が減少していることが分かるため，最初の 1，2 回目の検索の方が様々な推薦キーワードを表示できる可能性が高いと考えられる．

1~7 の評価による実験結果より，未読ページから抽出した推薦キーワードはユーザにとってある程度役に立つキーワードが含まれていることが分かったことから未読の Web ページにはユーザの必要としている単語が存在する可能性を示すことが出来たと考えられる．

7. おわりに

本研究では検索意図に関する単語と共起する推薦キーワードを未読ページから抽出することにより，ユーザが未確認で検索意図に関するキーワードの抽出を行う提案をし

た．また，キーワード推薦システムを実装し，評価を行なった．その結果，未読ページから抽出した推薦キーワードはある程度ユーザにとって役に立つキーワードであったと示す事が出来，未読の Web ページにユーザの必要とする単語が存在する可能性を示す事が出来た．

今後の課題として今回，推薦キーワード表示した際の使用率が低かった．これは，未読の Web ページから抽出した単語の頻度の多い上位 10 件を推薦キーワードとして表示したことが一つの原因だと考えられる．今後は頻度だけではなく他のパラメータを追加して順位付けしていく必要があると考えられる．

参考文献

1) 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,全国大会講演論文集 第 70 回平成 20 年(1), "1-493"-"1-494",(2008)
2) 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.
3) 南 翔太郎, 岡 誠, 吉村 宏樹, 森 博彦:閲覧行動モニタリングに基づく検索意図の抽出と検索支援, SICE SSI2011, pp96-101, 2011