4Q-01

Webでの特徴語と共起する単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†　　岡　誠‡　　森　博彦†

東京都市大学大学院工学研究科†　　東京都市大学知識工学部‡

1. はじめに

近年，Web での情報検索が頻繁に行われている．情報検索は適切なキーワードを選択することが重要となる．ユーザが検索を行っ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って様々な Web ページから情報を取得する．しかし，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．特に繰り返し検索を行わなければならない検索目的の場合，検索キーワードとなる単語の発見はさらに難しくなってしまう．

そこで，検索の際にユーザが確認していない未確認の単語が検索キーワード，検索キーワードのヒントとなる可能性があるため重要となる．未確認の単語はユーザが未読の Web ページに多く含まれていると考えられる．

2. 関連研究

情報検索を支援する研究として望月ら[1]の研究がある．この研究は検索結果におけるランキング変動に着目することで，ユーザが適合・不適合を判定した 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にとって既読の 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．逆に，ユーザにとって未読の Web ページにはユーザが未確認のキーワードが多く存在することが考えられる．しかし，未読の Web ページに存在する単語の中からユーザの知りたい情報にあったキーワードを推薦することは既読の Web ページから推薦することよりも難しい．なぜなら，未読の Web ページはユーザが Web ページを閲覧していないためユーザの検索意図を推測する手がかりが無いからである．

ユーザの未読の Web ページから検索支援を行う研究として小野ら[2]の研究がある．小野ら[2]はユーザの検索意図に関係する単語を特徴語として抽出した．ユーザの検索意図とは「ユーザが知りたい情報」である．この特徴語の抽出はユーザが既読の Web ページからユーザの閲覧行動を用いて抽出した．そして，この特徴語を用いて未読の Web ページからユーザの検索意図に関係し、ユーザが未確認の単語を抽出して推薦キーワードとして提示する手法を提案した．この研究結果から未読の Web ページからユーザにとって役に立つ推薦キーワードをユーザに表示することが可能だと示した．しかし，問題点として推薦キーワードを使用した際，検索結果にユーザにとって有効な Web ページの表示率が低いことが分かった．

3. 研究目的

ユーザの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い，ユーザが推薦キーワードを使用することでユーザにとって有効な Web ページを多く表示出来ることを目的とする．

4. 推薦キーワードの評価

4.1.有効な Web ページの表示検証実験

小野ら[2]の研究で推薦キーワードによってユーザにとって有効な Web ページの表示率が低いことが分かった．さらに詳しく調べるため，小野ら[2]の提案した推薦キーワードによってユーザにとって有効な Web ページの表示が検索回数毎にどのようになるか確認するために実験を行う．

推薦キーワードを使用した際とユーザ自身が考えた検索キーワードを使用した際の検索結果の比較を行う．そのため，小野ら[2]の実験で扱った検索課題を被験者に複数のキーワードを用いた&検索で行ってもらった．

4.1.1. 評価方法

推薦キーワードを使用することでユーザが未確認の検索課題に関する情報が見つかるか確認する．実験でユーザが考えた検索キーワードと推薦キーワードで&検索を行った検索結果を評価した．評価方法は検索回数毎の新規の Web ページの表示率と新規の Web ページの中でユーザに有効な Web ページの表示率で評価を行った．

4.1.2. 結果

検索回数毎の新規の Web ページの表示率と新規の Web ページの中で検索課題に関する Web ページの表示率で比較を行った結果を示す．推薦キーワードを使用することで新規の Web ページの全体の表示率は 0.47 で検索が進む毎に表示率は下がっていく傾向が見

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

† Kentarou Ono, Hirohiko Mori
Graduate School of Engineering, Tokyo City University Graduate School
‡Makoto Oka
Faculty of Knowledge Engineering, Tokyo City University Undergraduate Division

Copyright ©2017 Information Processing Society of Japan.
All Rights Reserved.

られた. 検索課題に関する Web ページの全体の表示率は 0.36 と低く, 検索が進む毎に徐々に下がっていく傾向が見られた.

4.1.3. 考察

結果から各検索回数で推薦キーワードを使用することによって, 検索課題に関係のないリンクを多く表示してしまう問題点があることが分かった. これは検索が進んだ際でも, 推薦キーワードにユーザの検索意図が反映出来ていない可能性が考えられる. 小野ら[2]は検索が進む毎に推薦キーワードにユーザの検索意図を反映させる提案をしている. 推薦キーワードにユーザの検索意図が含まれるかどうかはユーザの閲覧行動によって推薦キーワードが抽出されるまでの過程が重要となる. そこで, 実験結果から推薦キーワードが抽出されるまでの過程の中でユーザの検索意図を抽出する箇所に問題点がある可能性が考えられる.

5. テキスト量を考慮した特徴語のパラメータ調整

推薦キーワード評価実験の問題点から推薦キーワードにユーザの検索意図を反映させる新たな提案を行う.

小野ら[2]の提案手法で推薦キーワードはユーザの閲覧行動から抽出される特徴語を使って抽出を行う. よって特徴語は推薦キーワードのユーザの検索意図となるため重要となる. 特徴語の抽出は小野ら[2]の提案手法ではユーザがクリックしたリンクのタイトル・スニペットとリンクのテキストデータから抽出した. しかし, クリックしたタイトル・スニペットとリンクの Web ページのテキストデータではテキスト量の違いがある. また, 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可能性が高い. これらからクリックしたリンクのタイトル・スニペットから抽出する単語の頻度の重みを上げる. このようにすることで検索意図が含まれる可能性が高いタイトル・スニペットから抽出される単語の頻度の重みが上がり, テキスト量の違いにも対応出来ると考えている.

6. 特徴語の提案を用いた特徴語・推薦キーワードの評価

6.1.特徴語の評価

提案手法によって特徴語がユーザの検索意図として抽出出来ているか評価した. 使用したデータは推薦キーワード評価実験のデータを用いた. 評価方法は検索課題に関する特徴語が含まれているか小野ら[2]の提案手法と比較した.

その結果, 小野らの[2]提案手法の特徴語の中で検索課題に関する単語の表示率は 0.17, 提案手法の特徴語の中で検索課題に関する単語の表示率は 0.25 だった. この結果から提案手法を用いることでユーザの検索意図が特徴語に反映することが出来ると考えられる. また, 特徴語を用いて抽出される推薦キーワードに対してもユーザの検索意図をさらに反映出来る可能性があると考えられる.

6.2. 推薦キーワードの評価

特徴語の提案を用いた推薦キーワードを使用することでユーザが検索課題に関する未確認の情報が見つかるか確認する. 実際に被験者に検索課題を行ってもらい推薦キーワードを評価した. 評価実験でユーザが考えた検索キーワードと推薦キーワードで&検索を行った検索結果を評価した. 評価方法は新規の Web ページの中でユーザに有効な Web ページの表示率の評価を行い, 小野ら[2]の提案手法と比較した.

その結果, 小野ら[2]の提案手法を用いた推薦キーワードの表示率は 0.37, 特徴語の提案を用いた推薦キーワードの表示率は 0.39 という大きく値が上がることは確認できなかった. この原因として, 今回推薦キーワードの抽出は単純に頻度が多い単語を抽出した. そのため、検索課題に関係しない推薦キーワードには一般的に使われる単語が多く含まれてしまったことが考えられる.

7. 推薦キーワードの絞込み

特徴語の提案を用いた推薦キーワードの評価の問題点から一般的な単語を取り除く提案を行う.

小野ら[2]の提案手法で推薦キーワードは特徴語と同じ文章内に存在する単語の頻度が多い単語を推薦キーワードとして提示していた. しかし, 単純に頻度が多い単語を推薦キーワードとした場合一般的に使われる単語が多く含んでしまう可能性がある. そこで, 推薦キーワードを抽出する際に tf･idf を用いることを提案する. 多くの文章に存在する単語は一般的に使われる単語が多いと考えられるため, 特定の文章に出現する単語を抽出出来る tf･idf を用いることで一般的に使われる単語を取り除くことが出来, 推薦キーワードを絞り込むことが出来る可能性があると考えられる.

8. おわりに

本研究はユーザの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手法を提案した. 有効な Web ページの表示検証実験を行った結果からテキスト量を考慮した特徴語のパラメータ調整と推薦キーワードの絞込みが必要だと分かった. 今後はテキスト量を考慮した特徴語のパラメータ調整と推薦キーワードの絞込みの両方を行った推薦キーワードの評価を行うことを考えている.

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] 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キーワード推薦, 研究報告グループウェアとネットワークサービス (GN),2015-GN-96(18),1-5(2015-09-25) , 2188-8744

Copyright ©2017 Information Processing Society of Japan.
All Rights Reserved.