

推薦論文

検索質問と検索結果の変化を利用した先読み検索

藤田 遼 治^{†1} 太田 学^{†1}

我々は、検索結果の推移を用いてユーザの検索意図を推測し、ユーザの代わりに検索質問を生成して検索する先読み検索を提案する。本研究では、検索結果の推移に加え、ユーザが入力した検索質問の変化パターンを利用してユーザの検索意図を推測する。本稿では実装したプロトタイプシステムを、擬似適合性フィードバックによる検索、および Google が示す検索キーワード候補による検索と比較することにより評価を行った。さらに、先読み検索において検索質問変化パターンを考慮することの効果を実量的に評価した。

A Prediction Search Using Changes of Queries and Search Results

RYOJI FUJITA^{†1} and MANABU OHTA^{†1}

We propose a prediction search which infers a user's search intention based on a change in the search results and searches for an automatically generated query on behalf of the user. In addition to the change of search results, we also utilize a change of the user's queries for inferring its search intention. In this paper, we evaluated our implemented prototype system by comparing it with pseudo-relevance feedback retrieval and the retrieval of Google suggested queries. We also quantitatively evaluated the effect of utilizing a user's query change patterns for the prediction search.

^{†1} 岡山大学大学院自然科学研究科

Graduate School of Natural Science and Technology, Okayama University

1. はじめに

Google^{*1}に代表されるサーチエンジンは Web 上の情報とユーザとの接点として重要な役割を果たしている。しかし、Web 上のコンテンツは膨大であり、1 度検索しただけでユーザが求める情報をいつも得られるとは限らない。そのためユーザは通常、求める情報が得られるまで検索質問を修正し再検索を繰り返す。サーチエンジンの高速化により、1 度の検索時間は短縮されたが、適切な検索質問が思いつかなければ必要な情報を得るまでに思わぬ時間がかかってしまうことも少なくない。

そこで本研究では、ユーザの連続した検索に着目し、ユーザの検索の先を見越し、検索意図に沿った検索質問を自動生成して検索する先読み検索を提案する。本研究では、検索結果中に現れる特徴的な語の出現頻度の変化と検索質問の変化からユーザの検索意図を推測し、その検索意図に沿った適切な検索質問を自動生成して検索を行い、その検索結果を提示する。

本稿ではまず、2 章で関連研究について述べる。3 章で検索質問の変化を考慮しない先読み検索、4 章で検索質問の変化を考慮した先読み検索について説明し、5 章でプロトタイプシステムの実装について述べる。さらに、6 章で評価実験を行い、7 章でまとめと今後の課題について述べる。

2. 関連研究

本章では、本研究に関連する主な研究として、2.1 節で検索結果からの検索語候補の取得に関する研究、2.2 節で検索過程を用いた検索語想起支援に関する研究、2.3 節で適合性フィードバックを利用した検索結果の再ランキングに関する研究について述べる。

2.1 検索結果からの検索語候補の取得

検索結果から後の検索に役立つであろう検索語候補を取得し、ユーザに推薦する研究¹⁾が行われている。検索結果から語の出現頻度に基づくスコアを用いて検索語候補を取得し、ユーザに提示する。ユーザは提示された語から任意の語を選択することにより、検索質問を拡張することができる。

また、単語の出現する文どうしの距離から関連単語を抽出し、検索質問拡張に利用する研究²⁾もある。検索結果中の単語の出現頻度、および検索語が含まれる文とその単語が含まれる文との距離からスコアを計算し、スコアが上位の単語を検索質問に追加することで検索

*1 <http://www.google.co.jp/>

質問拡張を行う．本研究でも検索結果を用いて検索質問を拡張するが，検索結果の推移を利用して検索質問を拡張する点が異なる．

2.2 検索過程を利用した検索語想起支援

検索結果だけでなく，検索の全体像をデータフロー図として表示することで，検索過程の処理をユーザに分かりやすく提示する試み³⁾も行われている．データフロー図に基づいた視覚的なインタフェースを用いて検索条件や検索結果を提示することにより，ユーザが検索過程や処理の結果を理解しやすく，またデータフロー図を直接操作することで，検索条件の修正などが簡単に行えるようになっている．中岡ら⁴⁾は取得した検索語候補をグラフで表示することにより，視覚的に検索を進める手法を提案した．検索語 q の検索結果から「 q や s 」, 「 q の t 」というフレーズを用いて検索語候補を取得する．検索語と取得した語をノードとしたグラフを提示し，ノードを選択したときにそのノードの語を検索語として再検索し，グラフを広げていくことで検索過程の“見える化”を行っている．本研究でも検索過程を利用するが，検索質問生成をユーザに任せるのではなく，自動生成し検索を行う点が異なる．

検索結果中に出現する特徴語の出現頻度の変化を用いて検索語候補を取得する研究^{5),6)}もある．そこでは，特徴語の出現頻度の変化を「増加」, 「減少」, 「高い値での推移」の3種類に分類し，出現頻度の変化がその特徴語への興味の変化と考え，興味の増加した語，減少した語，興味を維持している語が検索結果から抽出されている．三浦⁵⁾は出現頻度の変化が似ている語を類似した語だと考え，クラスタリングを行い，クラスタの代表語をユーザに提示した．また長畑ら⁶⁾は，興味の増減した語および興味が維持されている語の可視化手法を提案している．本研究でも同様に出現頻度の変化を興味の変化と見なして特徴語を抽出する．しかし，本研究では特徴語の抽出に検索質問の変化も考慮する．さらに，検索語の想起支援ではなく，先読み検索を自動で行う点も異なる．また，検索質問の変化を利用する研究には，検索質問の変化とクリックログからのフィードバックによって検索結果の再ランキングを行うものもある⁷⁾．本研究は，検索結果の再ランキングではなく先読み検索を行う点で異なる．

2.3 適合判定を利用した再ランキング

サポートベクタマシンを用いて，文書を適合，不適合に分類し，分類した文書を適合性フィードバックに利用する手法⁸⁾が提案されている．検索結果の上位 N 件をユーザに適合か，不適合か判定してもらう．その後，判定された文書を用いて学習された SVM で検索結果の全文書から適合文書を分離する．分離した適合文書集合のみを対象として再度検索を行うことで，最終的な検索結果を得る．

一方，検索結果を直接編集することで再ランキングを行うシステム^{9)–11)}も提案されている．検索結果中の語に対して直接，強調操作，削除操作をすることで，その語を含む文書のランキングを移動させる．また，ページをノードとし，ページ間に類似度を基にエッジを張りページランクアルゴリズムを適用することで，検索意図に適合する確率の高いページをランキングの上位に表示している．強調操作，削除操作を行うごとにページ間の類似度に重みを与え，遷移確率行列を計算する．この遷移確率行列により，複数回の操作を再ランキングに反映させることができる．

本研究は，ユーザにこのような検索結果や特徴語の適合判定といった負荷はかけない．また，1度得られた検索結果の再ランキングではなく，得られた情報から自動で再検索を行う点が異なる．

3. 検索質問の変化を考慮しない先読み検索

本章では，検索質問の変化を考慮しない先読み検索手法について述べる．本研究では，ユーザが目的とする検索結果を1度の検索で得られず，連続して検索を行う場面を想定して支援を行う．ここで検索行動とは，検索質問の入力とし，一連の検索行動が n 回からなるとき， i 回目の検索をそのまま「 i 回目の検索」, i 回目の検索の検索質問を「 i 回目のクエリ」と表現する．また，ユーザからの検索質問入力による検索を「通常検索」, ユーザの検索意図を推測して生成した検索質問を用いてシステムが自動で行う検索を「先読み検索」と呼ぶ．以下に先読み検索の流れを示す．

- (1) 通常検索を行い検索結果を取得
- (2) 検索結果から特徴語とその出現頻度を取得
- (3) 前回の検索結果と今回の検索結果から変化を計算
- (4) 先読み検索用の検索質問を生成
- (5) 先読み検索を実行
- (6) (1)に戻る

検索結果の取得には Yahoo! ウェブ検索 Web API^{*1}を用いる．

3.1 特徴語の抽出

特徴語の抽出は長畑ら⁶⁾の手法を参考にした．取得した検索結果のタイトルおよびサマ

*1 <http://developer.yahoo.co.jp/>

りから形態素解析器 Sen^{*1}により形態素を取得する．得られた形態素のうち以下のいずれかの条件を満たす形態素を，特徴語を構成する要素として抽出する．ただし，除外形態素リスト中の語は除く．

- (1) カタカナのみから構成される形態素
- (2) 名詞（形容詞語幹，副詞可能，非自立，代名詞は除く）
- (3) アルファベットのみで構成される未知語
- (4) 数字
- (5) 漢字のみで構成される未知語
- (6) 連体助詞の「の」
- (7) 名前区切りの記号

上記の形態素が連続して現れた場合は連結する（表 1）．ただし，(6)，(7) は語頭にならない．連結処理により得られた語は長くなりがちであるため，接頭詞，接尾詞による分割処理を行う．具体的には接頭詞の直前と接尾詞の直後で分割する．たとえば，連結処理により得られた「新幹線新倉敷駅」は「新幹線」と「新倉敷駅」に分割される．これは，表 2 に示すように，形態素「新」が接頭詞であるため，「新」の直前で分割されている．

以上の処理で得られた語から，以下の語を取り除く．

- (1) 除外語リストに含まれる語

表 1 連結の例

Table 1 Examples of concatenation.

形態素	連結後
窓，の，杜	窓の杜
バラク，・，オバマ	バラク・オバマ

表 2 分割の例

Table 2 An example of division.

形態素	品詞情報	分割後
新幹線	名詞-一般	新幹線
新	接頭詞-名詞接続	新倉敷駅
倉敷	名詞-固有名詞-地名-一般	
駅	名詞-接尾-地域	

*1 <https://sen.dev.java.net/>

- (2) ひらがなのみの語
- (3) 数字のみの語
- (4) アルファベット 1 字の語
- (5) カタカナ 1 字の語
- (6) 漢字 1 字の語
- (7) 日時などを表す語
- (8) 「ん」「ー」で始まる語

最終的に得られた特徴語をその単語頻度や文書頻度とともに保存する．

3.2 検索結果中の興味の推移を表す語

本研究では，一連の検索行動による特徴語の出現文書率の増減からユーザの興味の推移を推測する．本研究では，出現文書率の増減を単純に特徴語に対するユーザの興味の増減と見なす．さらに，新たに興味を持った語を「推測新規語」，興味を失った語を「推測否定語」，興味が維持されている語を「推測維持語」と定義し，検索結果から抽出する⁵⁾．なお，この抽出は推測新規語，推測維持語，推測否定語の順で行い，1 度抽出された語は後から抽出される推測語の候補から除く．よって 1 語が 2 種類以上の推測語に選ばれることはない．

3.2.1 推測新規語

語の出現文書率が前回の検索結果に比べ増大している場合，その語に対する興味が強まったと考える．そこで， i 回目の検索における特徴語 t_j の出現文書増加率を式 (1) で定義し，出現文書増加率の高い特徴語を推測新規語とする．

$$RDF_INCREASE_{i,j} = RDF_{i,j} - RDF_{i-1,j} \quad (1)$$

$$RDF_{i,j} = \frac{df_{i,j}}{n_i} \quad (2)$$

ここで $df_{i,j}$ は i 回目の検索における特徴語 t_j の文書頻度， n_i は i 回目の検索における全検索結果文書数である．よって，式 (2) は i 回目の検索における特徴語 t_j の出現文書率を表す．

3.2.2 推測維持語

一連の検索において，出現文書率が高い値で変わらずに推移している場合，その特徴語は検索の話題を表していると考えられる．そこで， i 回目の検索における特徴語 t_j の出現文書維持率を式 (3) で定義し，出現文書維持率の高い特徴語を推測維持語とする．

$$RDF_KEEP_{i,j} = \frac{\min(RDF_{i,j}, RDF_{i-1,j})}{\sqrt{|RDF_INCREASE_{i,j}|} + 1} \quad (3)$$

3.2.3 推測否定語

語の出現文書率が前回の検索結果に比べ減少している場合、その語に対する興味が弱まったと考える。そこで、 i 回目の検索における特徴語 t_j の出現文書減少率を式 (4) で定義し、出現文書減少率の高い特徴語を推測否定語とする。

$$RDF_DECREASE_{i,j} = RDF_{i-1,j} - RDF_{i,j} \quad (4)$$

3.2.4 先読み検索質問の生成

これらの推測語を用いて先読み検索用の検索質問 Q_{inf} を生成する。推測新規語 INT 、推測維持語 IKT 、推測否定語 IDT をそれぞれ n 語ずつ用いて、以下の式により生成する。ここで、 Q_i は i 回目のクエリで、1 語以上の検索語の並びとする。

$$\begin{aligned} Q_{inf} = Q_i \quad & \text{AND} (INT_1 \text{ OR } INT_2 \text{ OR } \dots \text{ OR } INT_n) \\ & \text{AND} (IKT_1 \text{ OR } IKT_2 \text{ OR } \dots \text{ OR } IKT_n) \\ & \text{NOT } IDT_1 \text{ NOT } IDT_2 \dots \text{ NOT } IDT_n \end{aligned} \quad (5)$$

4. 検索質問の変化に基づく先読み検索

鈴木ら¹²⁾ は検索質問の変化からユーザの検索意図を推定するための調査を行った。その中で、検索質問の変化パターンを表 3 の 9 種類に分類し、その検索意図を推測している。我々は、検索質問の変化を削除変化、追加変化、変更変化の 3 種類に分類し (表 4)、それぞれの変化パターンについて検索意図を推測した。推測した検索意図に基づき、式 (1)、(3)、(4) を利用した新たなスコアを定義し、それにより検索質問の変化を考慮した推測検索語を獲得する。本研究では以下に述べるように、それぞれの検索質問変化パターンごとに先読み

表 3 鈴木ら¹²⁾ による検索質問の変化パターンの分類
Table 3 Classification of query change patterns by Suzuki, et al.¹²⁾.

検索質問変化パターン*	説明
A → AB	最初の検索語から検索語の数を増やす場合
AB → AC	先頭の検索語は変えずに後ろの検索語を他の検索語に変える場合
AB → A	1 回目に複数個の検索語で検索を行ったが、2 回目で検索語を減らす場合
A → AB → AC	2 回目の検索で検索語を増やし、3 回目の検索で後ろの検索語を変えている場合
A → AB → A	2 回目の検索で検索語の数を増やし、3 回目の検索で検索語を減らし元に戻す場合
A → ABC → AB	2 回目の検索で 2 語増やしたが、3 回目の検索で 1 語減らす場合
AB → A → AC	2 回目の検索で検索語を減らし、3 回目で検索語を増やす場合
AB → A → AB	検索語を減らした後、再び同じ検索語を増やす場合
AB → AC → AB	検索語を変えた後、再び同じ検索語に戻す場合

*A, B, C はそれぞれ異なる検索語を表す。

検索の検索質問を生成する。ただし、 i 回目のクエリ Q_i は、曖昧でない限り Q_i 中の全検索語の集合も表すものとする。

4.1 削除変化

$Q_{i-1} \supset Q_i$ のとき、この検索質問の変化を削除変化と定義する。このとき、削除された語は検索に何らかの理由で不適当であったが、検索意図と無関係というわけではないと考える。

4.1.1 削除時新規語

削除変化の際に興味が増加した語を削除時新規語と定義する。具体的には、削除変化の際に式 (6) で定義するスコアが高い語を削除時新規語とする。

$$SCORE_{i,j} = RDF_INCREASE_{i,j} + COO_{i,j}(D) \quad (6)$$

$$COO_{i,j}(D) = \frac{cooccurrence_{i,j}(D)}{n_i} \times \frac{1}{2} \quad (7)$$

ここで、 D は削除変化によって削除された検索語の集合、 $cooccurrence_{i,j}(D)$ は i 回目の検索結果において、 D 中のいずれかの語と特徴語 t_j が共起する文書の数である。また、 n_i は i 回目の検索における全検索結果文書数である。

4.1.2 削除時維持語

削除変化の際に興味が維持されている語を削除時維持語と定義する。興味が維持されている語は検索質問によらず高い出現文書率が維持されている語である。よって、推測維持語と同様に式 (3) の値が高い語を削除時維持語とする。

4.1.3 先読み検索質問

削除時新規語 DNT 、削除時維持語 DKT をそれぞれ n 語ずつ用いて、先読み検索用の検索質問 Q_{del} を生成する。

$$\begin{aligned} Q_{del} = Q_i \quad & \text{AND} (DNT_1 \text{ OR } DNT_2 \text{ OR } \dots \text{ OR } DNT_n) \\ & \text{AND} (DKT_1 \text{ OR } DKT_2 \text{ OR } \dots \text{ OR } DKT_n) \end{aligned} \quad (8)$$

表 4 検索質問の変化パターンの例
Table 4 Examples of query change patterns.

検索質問変化パターン	例
削除変化	パワーポイント 図 tex 取り込む → パワーポイント 図 tex
追加変化	内閣閣僚 → 内閣閣僚 鳩山
変更変化	新型インフルエンザ 対策 → 新型インフルエンザ 予防

4.2 追加変化

$Q_{i-1} \subset Q_i$ のとき、この検索質問の変化を追加変化と定義する。この場合、追加された語により絞り込み検索を行っていると考える。

4.2.1 追加時新規語

追加変化の際に興味が増加した語を追加時新規語と定義する。すなわち、追加変化の際に式 (9) で定義するスコアが高い語を追加時新規語とする。

$$SCORE_{i,j} = RDF_INCREASE_{i,j} \times W_{i,j} \quad (9)$$

$$W_{i,j} = \sqrt{TF_{i,j} \times IDF_{i,j}} \quad (10)$$

$$TF_{i,j} = \max_d \{tf_{i,j,d}\} \quad (11)$$

$$IDF_{i,j} = \log \frac{n_i}{df_{i,j}} \quad (12)$$

ここで、 $tf_{i,j,d}$ は i 回目の検索結果の文書 d における特徴語 t_j の出現頻度である。

4.2.2 追加時否定語

追加変化の際に興味が減少した語を追加時否定語と定義する。検索結果中の多くの文書に出現する語は絞り込みの目的にあわないと考えられる。そこで、式 (13) の値が高い語を追加時否定語とする。

$$SCORE_{i,j} = RDF_DECREASE_{i,j} \times RDF_{i,j} \quad (13)$$

4.2.3 先読み検索質問

追加時新規語 ANT 、追加時否定語 ADT をそれぞれ n 語ずつ用いて、先読み検索用の検索質問 Q_{add} を生成する。

$$Q_{add} = Q_i \text{ AND } (ANT_1 \text{ OR } ANT_2 \text{ OR } \dots \text{ OR } ANT_n) \\ \text{NOT } ADT_1 \text{ NOT } ADT_2 \dots \text{ NOT } ADT_n \quad (14)$$

4.3 変更変化

$Q_{i-1} \not\subset Q_i$ かつ $Q_{i-1} \not\supset Q_i$ かつ $Q_{i-1} \neq Q_i$ のとき、この検索質問の変化を変更変化と定義する。このとき、変更変化によって Q_{i-1} から削除された語は何らかの理由で検索に不要になったと考える。

4.3.1 変更時新規語

変更変化の際に興味が増加した語を変更時新規語と定義する。すなわち変更変化の際、式 (15) で定義するスコアが高い語を変更時新規語とする。

$$SCORE_{i,j} = RDF_INCREASE_{i,j} - COO_{i,j}(M) \quad (15)$$

ここで、 M は変更変化によって Q_{i-1} から削除された検索語の集合、すなわち $Q_{i-1} -$

$(Q_{i-1} \cap Q_i)$ である。

4.3.2 変更時否定語

変更変化の際に興味が減少した語を変更時否定語と定義する。すなわち変更変化の際、式 (16) で定義するスコアが高い語を変更時否定語とする。

$$SCORE_{i,j} = RDF_DECREASE_{i,j} + COO_{i,j}(M) \quad (16)$$

4.3.3 先読み検索質問

変更時新規語 MNT 、変更時否定語 MDT をそれぞれ n 語ずつ用いて、先読み検索用の検索質問 Q_{mod} を生成する。

$$Q_{mod} = Q_i \text{ AND } (MNT_1 \text{ OR } MNT_2 \text{ OR } \dots \text{ OR } MNT_n) \\ \text{NOT } MDT_1 \text{ NOT } MDT_2 \dots \text{ NOT } MDT_n \quad (17)$$

5. 実 装

図 1、図 2 は鳩山由紀夫内閣の組閣前の 2009 年 9 月 9 日に行った検索の実行例であり、



図 1 通常検索の実行画面

Fig. 1 Screenshot of the ordinary search.



図 2 先読み検索の実行画面

Fig. 2 Screenshot of the prediction search.

実装したプロトタイプシステムの実行画面である。図 1 はユーザが「内閣閣僚」で検索した後、「内閣閣僚 鳩山」で検索したときの画面である。このとき、プロトタイプシステムは 4 章で説明した方法で先読み検索用の検索質問を生成し、先読み検索を行っており、ユーザは「先読み検索」のタブを選択することで図 2 のように先読み検索の結果を閲覧できる。通常検索の画面である図 1 では「麻生内閣」や「鳩山一郎内閣」についてのページが検索結果リストに混じっているが、先読み検索の画面である図 2 では「鳩山由紀夫次期内閣」の関連文書が多数提示されている。

6. 実験

まず、4 章で提案した先読み検索の検索質問生成法を評価するため、式 (8), (14), (17) において $n = 1, 2, 3$ として比較を行った。次に、 $n = 3$ の場合について、検索質問の変化を考慮することの効果をも、3 章の方法と比較して評価した。さらに、擬似適合性フィードバックおよび Google サジェストとも比較を行い提案手法の有効性を評価した。

比較には検索結果上位 10 件の適合率 $P@10$ を用い、適合判定は岡山大学工学部情報工学科 4 年生の男性 5 名が行った^{*1}。5 名にはそれぞれ「高適合 (5)」から「不適合 (1)」までの 5 段階評価を行ってもらい、5 名の平均が 3 より高いものを適合文書と判定した。

また、実験は 2009 年 9 月 24 日、2010 年 2 月 23 日および 2010 年 5 月 31 日に行った。実験に用いた検索質問および検索意図を表 5 にまとめる。D1～D5 は削除変化、A1～A5 は追加変化、M1～M5 は変更変化であり、これらの検索質問は著者らの検索履歴などから収集した。さらに、これらの検索質問で検索を行った際、先読み検索が使用した検索語を表 6 にまとめる。たとえば D1 の場合、表 5 に示すように 1 回目に「パワーポイント 図 tex 取り込む」、2 回目に「パワーポイント 図 tex」という検索質問で通常検索を行う。このとき表 6 から、追加する検索語が 3 語、つまり $n = 3$ であれば「パワーポイント 図 tex AND (作成 OR 変換 OR powerpoint) AND (方法 OR pdf OR eps)」という先読み検索質問が自動生成されることが分かる。なお、表 6 に示した新規語、維持語、否定語ではそれぞれ左の語ほどスコアが高い。たとえば D1 の削除時新規語であれば、「作成」のスコアが高く、「変換」「powerpoint」となるに従ってスコアが低くなる。

6.1 先読み検索質問の評価

表 5 に示した検索質問の検索結果上位 10 件の適合率 $P@10$ を比較する。1 回目のクエリ、2 回目のクエリでの通常検索および、先読み検索の $P@10$ を表 7 に示す。先読み検索では、式 (8), (14), (17) において $n = 1, 2, 3$ として生成した先読み検索質問を用いて行った検索の結果をそれぞれ示す。

6.1.1 通常検索の適合率が高い場合

表 7 の D2, D4, A4, M1, M3 は 2 回目の通常検索で 0.8 以上と高い適合率を示している。これらと先読み検索の適合率を比較すると、D4, M1 のように n の値によっては先読み検索が適合率を改善している場合もあるが、A4, M3 のように先読み検索が逆効果となることもある。たとえば、「口蹄疫が人に感染するのか」を検索意図とする A4 では、追加時新規語の「宮崎県」の影響により「宮崎県の口蹄疫のニュース」が検索結果中に多くなった。 $n = 3$ で初めて「ウイルス」が追加されたことにより「人に感染するかどうか」の情報が現れ、適合率が改善されているがそれでも通常検索より適合率が低くなっている。この例のように、通常検索の適合率が高い場合は、追加される語が逆効果になり、先読み検索がうまく働かない場合がある。しかし、先読み検索は通常検索のバックグラウンドで行われ、ユーザ

*1 実験を複数回実施したため、延べ人数は 10 名である。

表 5 検索質問と検索意図
Table 5 Queries and their search intention.

	1 回目のクエリ	2 回目のクエリ	検索意図
D1	パワーポイント 図 tex 取り込む	パワーポイント 図 tex	パワーポイントで作成した図を $\text{\LaTeX}$ へ取り込む方法を知りたい
D2	tex 図 左右 並べる	tex 図 並べる	$\text{\LaTeX}$ で図を左右に並べる方法を知りたい
D3	バンクーバー 男子 フィギュア 結果	バンクーバー 男子 フィギュア	バンクーバーオリンピックの男子フィギュアスケートの結果を知りたい
D4	年金 学生 書類 提出	年金 学生 書類	国民年金の学生納付特例を申請する際に必要な書類を知りたい
D5	オリジナルジーンズ 製作 体験	オリジナルジーンズ 製作	オリジナルジーンズの製作体験のできる店やツアーなどを知りたい
A1	内閣閣僚	内閣閣僚 鳩山	鳩山由紀夫内閣の閣僚人事を知りたい
A2	BALLAD	BALLAD 原案	映画『BALLAD 名もなき恋のうた』の原案が何か知りたい
A3	GDP	GDP 略語	略称である GDP の正式名称を知りたい
A4	口蹄疫 感染	口蹄疫 感染 人	口蹄疫が人に感染するのかどうか知りたい
A5	Sleipnir ブラグイン	Sleipnir ブラグイン おすすめ	ブラウザ Sleipnir の便利なブラグインを知りたい
M1	新型インフルエンザ 対策	新型インフルエンザ 予防	新型インフルエンザに対する対策法を知りたい
M2	Twitter なう 解説	Twitter なう 意味	Twitter で使われる『なう』という語の解説を知りたい
M3	岡山 土産	岡山 名産	岡山の土産として適当な特産品や名産品を知りたい
M4	覚える 憶える 辞書	覚える 憶える 違い	「覚える」と「憶える」の 2 つの漢字の意味と、その使い方の違いを知りたい
M5	就職超氷河期 内定率	2010 卒 内定率	2010 年前後の就職超氷河期世代の内定率を知りたい

表 6 追加される検索語
Table 6 Added search terms.

	{ 削除 追加 変更 } 時新規語	{ 削除 追加 変更 } 時維持語	{ 削除 追加 変更 } 時否定語
D1	作成, 変換, powerpoint	方法, pdf, eps	-
D2	環境, figure, 方法	画像, 表示, 文字	-
D3	オリンピック, 銅メダル, フリー	高橋大輔, メダル, 高橋	-
D4	証明, 申請, 印鑑	在学証明書, 被保険者, 納付	-
D5	販売, 生地, リメイク	デニム, 岡山県, 児島	-
A1	副総理, 民主党, 新内閣	-	首相官邸, 内閣総理大臣, 氏名
A2	クレヨンしんちゃん, アニメ, 国大合戦	-	テレビ朝日, 製作委員会, 特集
A3	国内, gross domestic product, 英語	-	日本, 経済, 発表
A4	宮崎県, 疑い, ウイルス	-	処分, 確認, 発生
A5	firefox, ie, スレイブニル	-	フェンリル, 公開, 開発
M1	感染, 症状, ワクチン	-	対策, 平成, 設置
M2	産経新聞, 言葉, 謝罪	-	解説, アカウント, 当選
M3	マスカット, 特産品, 白桃	-	お土産, 定番, 和菓子
M4	コソ, 名前, 苦勞	-	辞書, 単語, 英語
M5	年卒, 調査, 発表	-	超氷河期, 就職氷河期, 就職

が検索結果に満足しない場合のみ閲覧されると想定している．よって，通常検索の適合率が高い場合は，先読み検索は使用されない可能性が高く，実用上は問題にならないと考える．

6.1.2 通常検索の適合率が低い場合

表 7 の D5, M2, M4 のように 2 回目の通常検索の適合率が 0.2 以下と低い場合は先読み検索は有効に機能しないことが多い．この中では，M2 で適合率が改善されているが，こ

表 7 提案手法の P@10
Table 7 P@10 of the proposed prediction search.

	通常検索		先読み検索		
	1 回目の検索	2 回目の検索	$n = 1$	$n = 2$	$n = 3$
D1	0.2	0.4	0.5	0.5	0.6
D2	0.5	0.8	0.6	0.6	0.8
D3	0.4	0.5	0.4	0.6	0.6
D4	0.6	0.8	0.4	0.8	0.9
D5	0.2	0.2	0.0	0.1	0.0
A1	0.4	0.5	0.4	0.2	0.4
A2	0.2	0.5	0.8	0.8	0.9
A3	0.5	0.7	0.8	0.9	0.8
A4	0.5	0.9	0.2	0.3	0.7
A5	0.5	0.7	0.0	0.1	0.1
M1	0.7	0.9	1.0	1.0	0.7
M2	0.0	0.1	0.1	0.4	0.4
M3	0.9	1.0	0.9	0.7	0.7
M4	0.3	0.2	0.0	0.0	0.0
M5	0.3	0.7	0.4	0.3	0.0

れは「言葉」が変更時新規語として追加されたため、「当選う」や「地震う」という実際のツイートに関する記事が検索結果から減り、知りたかった「～う」という言葉の解説が検索結果に多く含まれるようになったためと考えられる。検索ノイズを減らす効果はあったものの、それでも先読み検索の適合率は 0.4 と半分には達していない。また、D5, M4 では先読み検索が逆効果となっており、たとえば、D5 では「販売」といった、「ジーンズ」との関連はあるが、検索意図である「オリジナルジーンズの製作」とは異なる語が追加されてしまったため適合率が低くなっている。このように通常検索の適合率が低い場合は検索意図の推測が本質的に困難で、後述する他の検索支援手法でも適合率を改善できていない。

6.1.3 通常検索の適合率が中程度の場合

D1, D3, A1~A3, A5, M5 は、通常検索において 1 回目比べて 2 回目の適合率が向上したが、2 回目の検索の適合率が 0.3 から 0.7 と中程度であり、まだ検索の支援が必要な事例である。提案手法はこのような事例についてうまく先読み検索を行うことが期待される。表 7 を見ると、削除変化の D1, D3, 追加変化の A2, A3 では先読み検索の適合率が最も高い値を示している。たとえば、D1 では「変換」や「eps」といった「図を $\LaTeX$ へ取り込む方法」という検索意図に沿った語が追加されている。また、A2 の通常検索では、「BALLAD」と同名だが検索意図と無関係な会社名や CD が検索結果に含まれていたが、先読み検索で

は、「クレヨンしんちゃん」といった語が追加時新規語として選ばれ、「BALLAD」の原案である「クレヨンしんちゃん 嵐を呼ぶ アッパレ！戦国大合戦」が検索結果に多く含まれるようになった。

しかし、A1, A5, M5 については期待に反して先読み検索がうまく機能していない。A1 は取得した検索結果中に多数のリンク切れが存在していたために低い適合率となっており、これは使用した検索 API の問題と考える。一方、「Sleipnir のおすすめプラグイン」を検索意図とする A5 では、“firefox” や “ie (IE)” などの “Sleipnir” と競合関係にあるブラウザ名が検索語として追加されたことにより検索意図と異なる検索結果になってしまった。また、M5 では「2010 年前後の就職超氷河期世代の内定率」を検索意図としているが、「就職超氷河期」や「就職」が変更時否定語として追加されたため就職関連の話題から遠ざかってしまい、適合率が低くなってしまった。

6.1.4 検索質問の変化パターンに基づく分析

検索質問の変化パターンごとに表 7 をみると、提案手法は削除変化に有効である。追加変化については、通常検索の適合率が中程度で前回の検索と比べて適合率の向上が見られる場合、つまり通常検索は必ずしも間違っていないが欲しい検索結果がそれほど多くない場合に、一定の効果を発揮することが確認できた。しかしその場合でも、A1, A5 では先読み検索が有効ではなかったため、検索意図の推測についてはさらに検討したい。変更変化については先読み検索がうまく機能する場合もあるが、検索意図の推測が困難な場合も多く、今後さらなる検討が必要である。また、表 7 から分かるように、 n の値、すなわち追加する検索語数によって適合率は変動する。そのため、追加する検索語数の決定方法も今後の検討課題である。

6.2 検索質問の変化を考慮しない先読み検索との比較

式 (5) において $n = 3$ として、3 章で説明した検索質問を考慮しない先読み検索を行った。その結果の P@10 を表 8 の左の列に示す。表 7 の先読み検索の $n = 3$ と表 8 を比べると、検索質問の変化を考慮する表 7 の方がおおむね高い適合率を示した。A3 と M3 では表 7 の方が低くなったが、A3 では、表 7 の検索質問の変化を考慮する先読み検索の $n = 2$ と比べると 0.1 の差しかなく、M3 では表 7 の $n = 1$ の結果が、検索質問の変化を考慮しない先読み検索と同じ適合率を示している。以上より、先読み検索において検索質問の変化を考慮することは、有効であると考えられる。

6.3 擬似適合性フィードバックとの比較

2.3 節で説明した関連研究は、ユーザからのフィードバックを利用するある種の「適合性

表 8 その他の検索支援手法の P@10
Table 8 P@10 of the other search support methods.

	検索質問を考慮しない 先読み検索	擬似適合性フィードバック		Google サジェスト	
		1 回目の検索	2 回目の検索	1 回目の推薦	2 回目の推薦
D1	0.5	0.2	0.5*	-	-
D2	0.5	0.4	0.8	-	0.9
D3	0.3	0.2	0.1	-	-
D4	0.9	0.0	0.3	-	-
D5	0.0	0.1	0.0	-	-
A1	0.3	0.2	0.5	-	-
A2	0.9	0.0	0.4	0.8	-
A3	1.0	0.7	0.6	0.6	-
A4	0.2	0.0	0.2	0.1	-
A5	0.0	0.2	0.2	0.7	-
M1	0.7	0.0	0.9	0.8	0.8
M2	0.3	0.0	0.1	-	-
M3	0.9	0.5	0.6	0.9	0.9
M4	0.0	0.3	0.2	-	-
M5	0.0	0.1	0.6	-	-

*検索結果が 4 件しか取得できなかったためこだけ P@4 の値を示す。

フィードバック」といえる。一方、提案する先読み検索はユーザからの明示的なフィードバックは利用しないので、同じく明示的なフィードバックを利用しない「擬似適合性フィードバック」との比較を行った。

以下に提案する先読み検索の比較対象とした擬似適合性フィードバックについて説明する。通常検索の検索結果の上位 10 件を適合文書と見なし、その中に含まれる特徴語 t_j を式 (18) の値によって順位付けする。この上位 3 語の特徴語 ft_1, ft_2, ft_3 を i 回目のクエリに追加することで検索質問 Q_{prf} を生成する。

$$TFIDF_{i,j} = TF'_{i,j} \times IDF_{i,j} \quad (18)$$

$$TF'_{i,j} = \max_{d \in D_{10}} \{tf_{i,j,d}\} \quad (19)$$

$$Q_{prf} = Q_i \text{ AND } (ft_1 \text{ OR } ft_2 \text{ OR } ft_3) \quad (20)$$

ここで、 D_{10} は検索結果上位 10 件の文書集合である。1 回目および 2 回目の通常検索を利用して生成した Q_{prf} による検索結果の P@10 を表 8 の中央の列に示す。

表 7、表 8 に示すように、擬似適合性フィードバックの適合率は、その元になる通常検索の適合率と同程度の値を示すこともあるが、通常検索の適合率を改善するよりは悪化させる

ことの方が多かった。一方、擬似適合性フィードバック（2 回目の検索）と提案した先読み検索（ $n = 3$ ）と比較すると、A1、A5、M1、M4、M5 では擬似適合性フィードバックが良いが、そのほかは、先読み検索の方が良いものが 8 例、同じ適合率を示すものが 2 例となっている。なお先読み検索の方が悪かったこの 5 つの例はすべて、先読み検索（ $n = 3$ ）が通常検索を改善できなかった例に含まれている。また、D3、A2、M2 のように、通常検索の適合率が中程度以下で擬似適合性フィードバックが有効でない場合でも、提案する先読み検索では適合率を改善できる場合があることが分かる。

6.4 Google サジェストとの比較

Google が提供している検索キーワード候補推薦サービスとの比較を行う。1 回目のクエリ、2 回目のクエリを Google^{*1} に入力したときに推薦される検索質問候補の上位 3 件を我々のプロトタイプシステムで通常検索し P@10 を算出した。この 3 件の P@10 のうち最大値を表 8 の右の列に示す。なお、数値を書いていない欄は Google が検索キーワード候補を提示しなかったものである。

提案した先読み検索の $n = 1, 2, 3$ のうちの最良の結果と Google サジェストの結果を比較する。表 7 と表 8 をみると、D2、A5 については先読み検索の方が劣る結果となった。ここで、A5 では Google の 1 回目の推薦で「sleipnir プラグイン おすすめ」という検索質問が提示され、これは通常検索の 2 回目のクエリと同じ検索質問であるため 2 回目の通常検索と同じ適合率となっている。一方、A2～A4、M1 では提案した先読み検索の方が高い適合率を示している。さらに、Google では検索キーワード候補を提示できない場合があるが、提案手法では実験結果が示すように通常あらゆる場合に先読み検索を実行できる。

7. ま と め

本稿では、検索質問と検索結果の推移から、ユーザの検索意図を推測し、先読み検索を行う手法を提案した。評価実験では、実装したプロトタイプシステムを、擬似適合性フィードバック、Google の検索キーワード候補推薦サービスと比較した。さらに、検索質問の変化を考慮しない先読み検索と比較することで、先読み検索において検索質問の変化パターンを考慮することの有効性を評価した。

評価実験の結果、提案手法は通常検索の改善が可能であり、他の検索支援手法である擬似適合性フィードバックや Google の検索キーワード候補推薦サービスと比べても、検索結果

*1 <http://www.google.co.jp/>

の改善に効果的であることが分かった。また、先読み検索において検索質問の変化パターンを考慮することは先読み検索の精度向上に一定の効果があることを確認した。

一方、先読み検索で追加する検索語の数によっては適合率が下がる場合も確認された。今後は追加する推測検索語数の動的な決定方法の検討や、推測検索語の算出式のさらなる検討を行う予定である。また、追加する推測検索語の種類についても考慮する必要がある。さらに、連続する2回の検索の変化のみの考慮では検索意図の推測に限界があるため、将来的には3回以上連続した検索に基づく、より詳しい検索意図の推測についても検討したい。

参 考 文 献

- 1) 酒井浩之, 大竹清敬, 増山 繁: 絞り込語提示による一検索手法の提案, 言語処理学会第7回年次大会, pp.185-188 (2001).
- 2) 大石哲也, 倉元俊介, 峯 恒憲, 長谷川隆三, 藤田 博, 越村三幸: 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, 電子情報通信学会論文誌 D, Vol.J92-D, No.3, pp.281-292 (2009).
- 3) 飯崎智之, 志築文太郎, 三末和男, 田中二郎: 検索過程を確認しながら条件指定が行える検索インタフェースの提案, 第19回人工知能学会全国大会 (JSAI 2005), 3D1-02 (2005).
- 4) 中岡美華, 大島裕明, 田中克己: 同位語発見システムを用いた子どもの利用履歴からの思考推移の抽出および分析, 電子情報通信学会第19回データ工学ワークショップ (DEWS2008), B7-1 (2008).
- 5) 三浦聖史: ユーザの検索行動に基づく検索キーワードの予測, 岡山大学工学部情報工学科特別研究報告書, 岡山大学 (2008).
- 6) 長畑洋臣, 太田 学: 検索結果の推移の可視化による検索支援, Web とデータベースに関するフォーラム (WebDB Forum) 2008 論文集, 5A-3 (2008).
- 7) Shen, X., Tan, B. and Zhai, C.: Context-sensitive information retrieval using implicit feedback, *Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.43-50 (2005).
- 8) 柘植 寛, 獅子掘正幹, 北 研二: サポートベクターマシンによる適合性フィードバ

クを用いた情報検索, 情報処理学会研究報告, Vol.2001, No.9, pp.83-88 (2001).

- 9) 山本岳洋, 中村聡史, 田中克己: ContextRank: 語ベースフィードバックおよびそのコンテキストに基づく検索結果の再ランキング手法, 第1回データ工学と情報マネジメントに関するフォーラム (DEIM2009), C8-1 (2009).
- 10) Rerank.jp . <http://rerank.jp/>
- 11) Rerank . <http://search.yahoo-labs.jp/rerank/>
- 12) 鈴木俊輔, 山名早人: 時間間隔を用いた検索履歴のモデル化, 情報処理学会研究報告, Vol.2002, No.28, pp.103-110 (2002).

(平成 22 年 3 月 20 日受付)

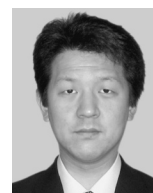
(平成 22 年 7 月 7 日採録)

(担当編集委員 中島 伸介)



藤田 遼治

2009 年岡山大学工学部情報工学科卒業。同年から岡山大学大学院自然科学研究科博士前期課程に在学。Web 情報検索に関する研究に従事。電子情報通信学会学生員。



太田 学 (正会員)

1994 年東京大学工学部電気工学科卒業。1999 年同大学大学院工学系研究科電気工学専攻博士課程修了。博士 (工学)。東京都立大学工学研究科助手を経て 2005 年岡山大学大学院自然科学研究科助教授。2007 年より同研究科准教授。Web 情報検索ならびに電子図書館の研究に従事。電子情報通信学会, 日本データベース学会, IEEE 各会員。