ユーザのクエリ変更意図に基づく先読み検索

藤田　遼治† 　太田　学† 　徳永　徹郎††

† 岡山大学大学院自然科学研究科　〒 700–8530 岡山県岡山市北区津島中 3-1-1
†† NTT レゾナント株式会社　〒 108–0023 東京都港区芝浦 3-4-1 グランパークタワー 8F
E-mail: †{fujita,ohta}@de.cs.okayama-u.ac.jp, ††t-tokunaga@nttr.co.jp

あらまし　ユーザの連続した検索からクエリ変更意図を推測し，ユーザの検索の先を読み，自動で検索する先読み検索を提案する．本稿ではまず，クエリログを分析して，ユーザのクエリ変更意図を分析する．その分析結果を用い，SVM によるクエリ変更意図の自動分類を行い，クエリ変更意図毎に先読み検索を行うことで，クエリ変更意図を考慮した先読み検索を実現する．実装したプロトタイプシステムにより先読み検索の適合率を定量的に評価する．

キーワード　情報検索，先読み検索，クエリ変更意図，検索意図，SVM，情報推薦

A Prediction Search Based on Query Reformulation Intent

Ryoji FUJITA†, Manabu OHTA†, and Tetsuro TOKUNAGA††

† Graduate School of Natural Science and Technology, Okayama University
3-1-1 Tsushima-naka, Kita-ku, Okayama-shi, Okayama, 700–8530 Japan
†† NTT Resonant Inc.
4-1-8F Granpark tower Shibaura 3-chome, Minato-ku, Tokyo, 108-0023 Japan
E-mail: †{fujita,ohta}@de.cs.okayama-u.ac.jp, ††t-tokunaga@nttr.co.jp

Key words　Information retrieval, Prediction search, Query reformulation intent, Search intent, SVM, Information recommendation

1. はじめに

Google(注1)などのサーチエンジンは，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，一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クエリの修正と再検索を繰り返す．つまり，ユーザは検索結果に満足しなければ何回か連続して検索する．

そこで本研究では，ユーザの連続した検索に着目し，そのユーザの検索意図を推測し，検索支援に利用する．具体的には，推測した検索意図を用い，ユーザの検索の先を読んで検索クエリを自動生成して検索する先読み検索を実現する．本研究では，まず SVM を用いてユーザのクエリ変更意図を推測し，その検索意図に沿う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 節で関連研究について述べ，3 節で検索結果からの特徴語の抽出方法を説明する．4 節で SVM を用いたクエリ変更意図の推測，5 節でクエリ変更意図に基づく先読み検索について説明し，6 節でプロトタイプシステムの実装について述べる．7 節で評価実験を行い，8 節でまとめと今後の課題について述べる．

(注1)：http://www.google.co.jp/

2. 関連研究

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，クエリを拡張することができる．

また，単語の出現する文同士の距離から関連単語を抽出し，クエリ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クエリに追加することでクエリ拡張を行う．本研究でも検索結果中に出現する語を用いて先読み検索のクエリを生成するが，その際に検索意図を考慮する点が異なる．

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，検索の過程をグラフで表すことで視覚的に検索を進める試み [3] も行われている．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．そして，検索語と取得した検索語候補をノードとしたグラフを提示する．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，自動生成して検索を行う点が異なる．

長畑ら [4] は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を行った．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を検索結果から抽出した．さらに，抽出した語をバネモデルで可視化することで，検索語の想起を支援している．本研究でも同様に出現頻度の変化を興味の変化とみなして特徴語を抽出する．しかし，本研究ではクエリ変更意図を考慮する．また，検索クエリの変化を利用する研究には，クエリ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研究 [5] もある．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 クエリ変更意図

Rieh ら [6] はユーザのクエリ変更意図を，意味内容を変更している “Content”，表記レベルの変更をしている “Format”，検索対象のデータに合わせた変更をする “Resource” に大別した．さらに詳細に，“Content” は「絞込」,「汎化」,「同義語への変更」,「関連語への変更」の 4 タイプ，“Format” は「ステミング」,「検索記号（AND，OR など）の追加」,「誤字修正」の 3 タイプ，“Resource” は「専門検索（ニュース，画像，動画など）から一般検索への変更とその逆の変更」,「URL への変更」の 3 タイプに分類している．

また，関口ら [7] は，SVM を用いてクエリの変更意図を「絞込」「汎化」「関連」「語句修正」「新規」の 5 種類に分類した．まず，日英音訳アルゴリズムを用いて，語句修正かどうかを判定し，語句修正でない場合は SVM によって残りの 4 種類に分類する．SVM に与える素性にはクエリの編集距離や共通語数などの表記上の関係を元にした素性，スニペットベクトルのコサイン類似度，および Skew divergence などの非対称な素性を用いている．本稿では，この関口ら [7] の手法を参考にしてクエリ変更意図を分類する．

3. 特徴語の抽出

我々も，酒井ら [1] や大石ら [2] と同様に検索結果から検索語候補を抽出する．長畑ら [4] の特徴語抽出法を参考に特徴語を抽出し，先読み検索の検索語に利用する．まず，検索結果のタイトルおよびサマリから形態素解析器 Sen(注2)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，別に定めた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．例えば，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） 別に定めた除外語リストに含まれる語
（2） ひらがなのみの語
（3） 数字のみの語
（4） アルファベット一字の語
（5） カタカナ一字の語
（6） 漢字一字の語
（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．

表 1 連 結 の 例

形態素	連結後
道，の，駅	道の駅
晴れ，の，国，岡山	晴れの国岡山
スティーブ，・，ジョブズ	スティーブ・ジョブズ

表 2 分 割 の 例

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	

4. クエリ変更意図の自動分類

我々は，ユーザのクエリ変更意図は，そのユーザの検索意図を推測する上で重要な手がかりとなると考えた．そこで本稿では，そのクエリ変更意図を先読み検索に利用するため，SVM を用いてユーザのクエリ変更意図を自動分類する．そのために，4.1 節で実際の商用検索エンジンのクエリログを調査し，クエリ変更意図を人手で分類する．次に 4.2 節で，その手動分類の結果に基づいて，SVM を用いたクエリ変更意図の自動分類を試みる．さらに，その自動分類の精度を 4.3 節で確認する．

(注2)：https://sen.dev.java.net/

表 3　クエリ変更意図の分類

種類	説明	例	クエリ対の数
絞込	絞り込まれた詳細な情報を得るための変更	子供の反抗期→中学生の反抗期	71
比較	複数の対象を比較するための検索	古本市場 買取→ブックオフ 買取	39
修正	同じ内容を表す別の語への変更	TIFF 結合 フリー→ TIFF ファイル結合 フリー	20
汎化	より一般的な情報を得るための変更	彼氏へ誕生日 プレゼント→プレゼント	17
語句	表記のゆれなどの修正	透析看護→透析 看護	13
道程	観光地などの情報収集のための検索	大沼トンネル→函館新道 道央道	10
固有名詞	固有名詞を調べるための変更	台湾女優→リンチーリン	2
調査	ある目的についての調査	ゴミ ゴミ問題→飲食 ゴミ	1

4.1　クエリログの分析

実際の商用検索エンジンに，2011 年 11 月中の連続した 3 日間に入力されたクエリログを解析した．まず，それぞれの日において複数回の検索を行っているユーザを 100 名ずつ無作為に抽出した．そして，合計 300 名のユーザの入力クエリを人手で分析し，クエリ変更意図を推測した．具体的には，この 300 名の入力した 1,434 個のクエリのうち検索意図が一貫していると判断でき，10 分以内の間隔で入力された連続した 2 回のクエリをクエリ対として抽出した．そして，抽出した 173 個のクエリ対を人手で確認し，クエリ変更意図を分類した．その結果を表 3 に示す．

なお，クエリ対のうち先に入力されたクエリを先行クエリ，後に入力されたクエリを後続クエリと定義する．例えば，「子供の反抗期→中学生の反抗期」というクエリ対では，先行クエリは「子供の反抗期」であり，後続クエリは「中学生の反抗期」である．また，表 3 に示すように，この例のクエリ変更意図は「絞込」であると推測している．

表 3 に示すように，8 種類のクエリ変更意図に分類できた．「絞込」の変更意図に最も多くのクエリ対が分類されている一方，「調査」の変更意図に分類したクエリ対は最も少なく 1 個だけであった．

4.2　SVM を用いたクエリ変更意図の自動分類

ユーザのクエリ変更意図を自動分類する方法を説明する．我々は，同一検索意図に基づく連続した検索を対象としているため，表 3 に示す 8 種類のクエリ変更意図への自動分類を試みる．具体的には，2 回の連続した検索から特徴量となる素性を抽出し，SVM を用いて表 3 に示した 8 種類のクエリ変更意図に分類する．なお，クエリ変更意図の自動分類は関口ら [7] の手法を参考とし，彼らの実験で最も高い分類精度を示した素性の組み合わせを利用した．学習に用いた素性を表 4 に示す．

Levenshtein 編集距離は 2 つの文字列の類似度を表す指標で，文字の挿入・削除・置換で一方の文字列を他方に変形するために必要な最小手順を数えたものである．たとえば，先行クエリ「子供の反抗期」と後続クエリ「中学生の反抗期」の Levenshtein 編集距離は「子」「供」の「中」「学」への置換と，「生」の挿入の 3 となる．Jaccard 係数は 2 つの集合の共起度合いを表す指標であり，先行クエリの構成語の集合 S_1 と後続クエリの構成語の集合 S_2 について式 (1) で定義される．

$$Jaccard(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \tag{1}$$

表 4　素 性 一 覧

素性	内容
levenshtein	クエリの Levenshtein 編集距離
common_char_l	クエリの左側からの共通文字数
common_char_r	クエリの右側からの共通文字数
common_word	クエリ中の共通語数
jaccard	クエリの構成語の Jaccard 係数
cosine	スニペットベクトルのコサイン類似度
char_add	後続クエリに追加された文字数
word_add	後続クエリに追加された語数
skew1	先行クエリから後続クエリへの Skew divergence
skew2	後続クエリから先行クエリへの Skew divergence

素性 cosine に用いるスニペットベクトルは，先行クエリ，後続クエリのそれぞれの検索結果上位 30 件のタイトルとサマリから抽出した特徴語を構成要素とした bag-of-words のベクトルとし，その重みは検索結果上位 30 件における出現文書率とした．i 回目の検索における検索結果上位 n 件中の特徴語 t の出現文書率を式 (2) に示す．なお，$df_{i,n}(t)$ は i 回目の検索における検索結果上位 n 件での特徴語 t の文書頻度である．

$$RDF_{i,n}(t) = \frac{df_{i,n}(t)}{n} \tag{2}$$

なお，検索結果数が 30 に満たない場合は n の値を得られた検索結果数とする．

先行クエリと後続クエリのどちらが意図的に絞り込まれているかを文字数から判断するため char_add と word_add を素性として導入した．char_add，word_add 共に後続クエリで文字（検索語）が追加された場合は正の値を，削除された場合は負の値を示す．例えば，先行クエリが「シーチキン　玉ねぎ　お弁当」，後続クエリが「ツナ　お弁当」なら，word_add は「シーチキン」と「玉ねぎ」の削除による −2 と，「ツナ」の追加による +1 の合計 −1 となる．

同様に検索結果が絞り込まれているかを素性として扱うためにスニペットベクトルの Skew divergence を導入した．スニペットベクトル $\boldsymbol{s_1}$ と $\boldsymbol{s_2}$ の Skew divergence を式 (3) に示す．

$$skew(\boldsymbol{s_1}, \boldsymbol{s_2}) = D(\boldsymbol{s_1}||\alpha\boldsymbol{s_2} + (1-\alpha)\boldsymbol{s_1}) \tag{3}$$

$$D(\boldsymbol{s}||\boldsymbol{s'}) = \sum_k \boldsymbol{s}(k) log \frac{\boldsymbol{s}(k)}{\boldsymbol{s'}(k)} \tag{4}$$

Skew divergence は式 (3) に示すように，Kullback-Leibler 距離 $D(\boldsymbol{s}||\boldsymbol{s'})$ をスムージングした指標である．$\boldsymbol{s}(k)$ はベクトル $\boldsymbol{s}$

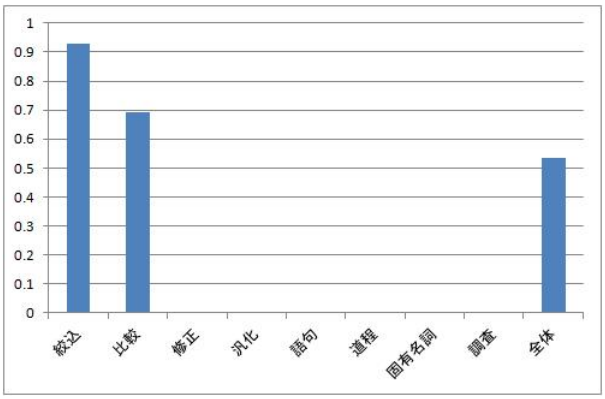

図 1　分 類 精 度

の k 成分である．表 4 に示すように skwe1 では $\boldsymbol{s_1}$ が先行クエリによる検索結果のスニペットベクトルであり，$\boldsymbol{s_2}$ が後続クエリによる検索結果のスニペットベクトルである．また，skew2 では $\boldsymbol{s_1}$ と $\boldsymbol{s_2}$ が逆となる．式 (4) から分かるように，$\boldsymbol{s_1}$ の各成分が大きいとき，$skew(\boldsymbol{s_1}, \boldsymbol{s_2})$ も大きくなる．例えば，様々な話題を含む検索結果から特定の話題の検索結果に絞り込むような検索の場合，先行クエリによる検索では特徴語の種類が分散しているため，各特徴語の出現文書率が低くなるが，後続クエリによる検索では特徴語の種類が絞られるため，各特徴語の出現文書率が高くなる．そのため，skew1 は小さな値をとり，skew2 は大きな値をとる．このとき，先行クエリによる検索結果に比べ後続クエリによる検索結果が絞り込まれていると判断できる．

4.3　クエリ変更意図の自動分類精度

4.1 節でクエリ変更意図の定義のために分析した 173 個のクエリ対を用いて，4.2 節のクエリ変更意図の自動分類実験を行った．SVM には LIBSVM(注3)，カーネル関数は RBF，その他も標準の設定を用いた．

5 分割交差検定で分類精度を求めた結果を図 1 に示す．なお，式 (3) の α の値は 0.1〜0.9 までを 0.1 刻みで実験し，全体の分類精度が最も高かった 0.3 としてある．図 1 を見ると，全体の分類精度は 0.53 となっている．また,「絞込」の分類精度は 0.93,「比較」は 0.69 であり，その他のクエリ変更意図ではすべて 0 であった．表 3 を見ると,「絞込」や「比較」はクエリ対が比較的多く，SVM が学習できたが，その他のクエリ変更意図については学習に十分な量のクエリ対がなかったためこのような結果になったと考えられる．

5.　クエリ変更意図に基づく先読み検索

ユーザが入力したクエリが，どのクエリ変更意図であるかを自動分類した後，先読み検索を行う．具体的には，2 回以上の連続した検索が行われた際に，現在入力された検索クエリによる検索と前回の検索から 4.2 節の自動分類法で，クエリ変更意図を自動分類し，クエリ変更意図毎に用意した先読みクエリ生成法により，先読みクエリを生成する．その後，生成した先読みクエリで自動検索を行う．本稿では，表 3 で頻度の高かった「絞込」と「比較」について先読みクエリの生成法を説明する．まず，5.1 節で先読みクエリの生成に用いる特徴語の選択法について説明する．その後,「絞込」の先読みクエリ生成法を 5.2 節,「比較」の先読みクエリ生成法を 5.3 節でそれぞれ説明する．

5.1　特徴語の選択

本稿では，連続した検索における特徴語の出現頻度の推移を用いて，その特徴語へのユーザの興味の推移を推測する．具体的には，連続した 2 回の検索において，出現頻度の増加した特徴語を注目語 (NT：Notable Term)，出現頻度が高く維持されている語を主題語 (TT：Theme Term)，出現頻度が減少した語を廃棄語 (DT：Disuse Term) として抽出する．これらの抽出した語を用いて先読みクエリを生成する．

5.1.1　注　目　語

特徴語の出現頻度が前回の検索に比べ増大している場合，その特徴語に対する注目度合が増していると考える．そこで，i 回目の検索における特徴語 t の出現文書増加率 $RDF_INCREASE_i(t)$ を式 (5) のように定義し，$RDF_INCREASE_i(t)$ の値が高い特徴語を注目語として抽出する．なお，ここで n_i は i 回目の検索における全検索結果文書数である．

$$RDF_INCREASE_i(t) = RDF_{i,n_i}(t) - RDF_{i-1,n_{i-1}}(t) \quad (5)$$

5.1.2　主　題　語

2 回の検索で出現頻度が高く維持されている特徴語はその検索の主題を表すと予想できる．そこで，i 回目の検索における特徴語 t の出現文書維持率 $RDF_KEEP_i(t)$ を式 (6) で定義し，$RDF_KEEP_i(t)$ の高い特徴語を主題語として抽出する．

$$RDF_KEEP_i(t) = \frac{min(RDF_{i,n_i}(t), RDF_{i-1,n_{i-1}}(t))}{\sqrt{|RDF_INCREASE_{i,n_i}(t)|} + 1} \quad (6)$$

5.1.3　廃　棄　語

特徴語の出現頻度が前回の検索に比べ減少している場合，その特徴語は検索に不要な語と考えられる．よって，これらを含む検索結果を取り除くことでユーザを支援できると予想される．そこで，i 回目の検索における特徴語 t の出現文書減少率 $RDF_DECREASE_i(t)$ を式 (7) で定義し，$RDF_DECREASE_i(t)$ の値が高い特徴語を廃棄語として抽出する．

$$RDF_DECREASE_i(t) = RDF_{i-1,n_{i-1}}(t) - RDF_{i,n_i}(t) \quad (7)$$

5.2　絞　　込

「絞込」のクエリ変更意図における先読みクエリ生成法を述べる．絞込み検索では，ユーザは様々な話題を含む検索結果から，少数の話題を含む検索結果に絞り込みたいと考える．そのため，検索結果中に出現する特徴語も，先行クエリでの検索に比べ，後続クエリでの検索では種類が減っていると予想できる．

(注3)：http://www.csie.ntu.edu.tw/~cjlin/libsvm/

つまり，特定の特徴語の出現する割合が多くなり，注目されていると予想できる．そこで，$RDF_INCREASE_i(t)$ の値の高い注目語 NT を n 語，入力されているクエリ Q_i に付加して先読みクエリ Q_{narrow} を生成する．ただし，既にクエリ内に存在する語は付加しないように n 語選択する．

$$Q_{narrow} = Q_i \text{ AND } (NT_1 \text{ OR } NT_2 \text{ OR } ... \text{ OR } NT_n) \quad (8)$$

5.3 比　　較

「比較」のクエリ変更意図における先読みクエリ生成法を説明する．比較の検索の場合，本稿の先読み検索では，入力されたクエリと同じ主題に属する別の対象を検索する．例えば，ユーザが「古本市場→ブックマーケット」と検索した場合，「ブックオフ」のような別の古本屋を検索するクエリを生成することを目標とする．そのため，入力されているクエリは用いずに，主題語 TT と廃棄語 DT を n 語ずつ用いて，先読みクエリ $Q_{compare}$ を生成する．なお，まず主題語を n 語定め，それらと重複しないように廃棄語 n 語を NOT 検索として付加して先読みクエリを生成する．

$$Q_{compare} = (TT_1 \text{ OR } TT_2 \text{ OR } ... \text{ OR } TT_n) \text{ NOT } DT_1 \text{ NOT } DT_2 \text{ ... NOT } DT_n \quad (9)$$

6. プロトタイプシステム

図 2 は「子供の反抗期」で検索した後,「中学生の反抗期」で再検索したときの画面である．このとき，先読み検索システムは SVM を用いたクエリ変更意図の自動分類により，このクエリ対を「絞込」と分類している．よって，5.2 節で説明した，「絞込」のための先読みクエリ Q_{narrow} を生成し，バックグラウンドで先読み検索を実行している．

ユーザは図 3 のように「Prediction search」のタブを選択することで，その先読み検索の結果を閲覧することができる．なお,「Prediction search」タブを選択すると，クエリ入力部分には先読みクエリが表示される．また，画面の左側には抽出した注目語，主題語，廃棄語が全て表示されており，タブを切り替えることで閲覧できる．

7. 評 価 実 験

まず，7.1 節で，先読みクエリについて考察する．そして 7.2 節では，5 節で説明した先読み検索に基づいて，プロトタイプシステムが出力する先読み検索結果の上位 10 件の適合率を求め，TOD47 で著者らが提案した先読み検索 [8] と比較する．

7.1 先読みクエリの考察

表 5 に示したクエリをプロトタイプシステムに入力し，注目語，主題語および廃棄語を抽出した．抽出した上位 10 語ずつを表 6〜表 9 に示す．なお，N1 と N2 は「絞込」のクエリ対であり，C1 と C2 は「比較」のクエリ対である．N1 はお祝いの言葉を検索しており，N2 は誕生日プレゼントを探している検索である．また，C1 はカー用品店の比較をしており，C2 は北海道のテレビ局の比較をしている．これらの表の中の，青い特徴語は後続クエリ中に存在する語もしくは，その語に検索に影響しない記号を付加した語である．

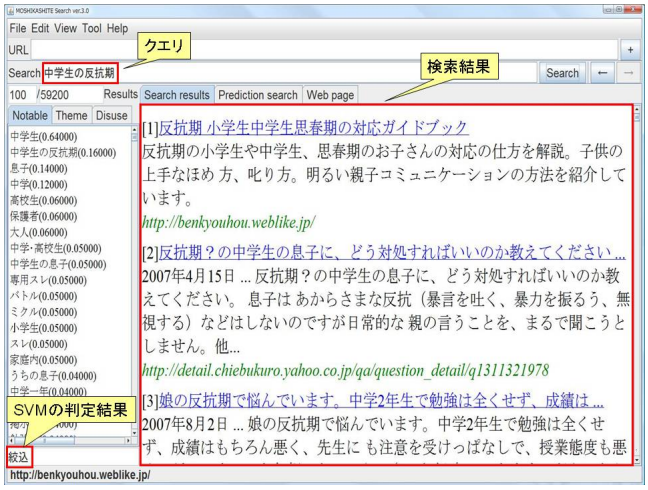

図 2　通常検索の実行画面

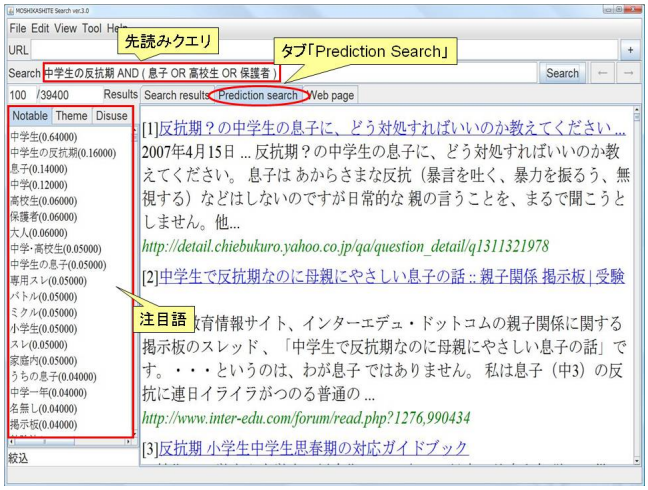

図 3　先読み検索の実行画面

表 5　先読みクエリの評価用クエリ対

	先行クエリ	後続クエリ
N1	一言	一言お祝いの言葉を
N2	プレゼント	プレゼント 誕生日
C1	オートバックス	イエローハット
C2	ＨＴＢ	ＴＶＨ

N1，N2 は絞込みをしているので，青の特徴語は用いずに先読みクエリ Q_{narrow} を生成している．表 6，表 7 より，青の特徴語以外の「注目語」は「結婚式」や「彼氏」など，検索を絞り込むのに適した語が多いことが分かる．また,「主題語」は「自分」や「贈り物」など絞込みには不十分な語が多かった．さらに,「廃棄語」では「ブログ」や「プレゼント情報」など検索意図と本当に無関係か判断しかねる語や，NOT 検索に用いると，必要な情報まで取り除いてしまう恐れのある語がみられた．そのため,「絞込」の場合，式 (8) の先読みクエリ生成式に一定の妥当性があると判断できる．

また，表 8，表 9 より,「カー用品」や「北海道」「テレビ」などの「主題語」は，比較対象と同じ主題に属するオブジェクトの検索に有効であることが分かる．さらに,「オートバックス」

表 6　N1 の注目語・主題語・廃棄語

注目語	主題語	廃棄語
言葉	一言	一言
お祝い	紹介	一言!
一言お祝い	自分	日前
お祝いの言葉	開催	日本
一言お祝いの言葉	受賞	ブログ
結婚式	参加	お送り
代表	仕事	東日本大震災
スピーチ	皆様	一言主神社
挨拶	言葉	blog
出席	年賀状の一言	ご案内

表 7　N2 の注目語・主題語・廃棄語

注目語	主題語	廃棄語
誕生日	プレゼント	プレゼント情報
彼氏	贈り物	キャンペーン
ギフト	誕生日	抽選
プレゼント	紹介	懸賞
友達	ご紹介	プレゼント!
フラワーギフト	アイテム	ご応募
贈り物	彼氏	応募
男性	記念日	present
花束	お祝い	お答え
お誕生日	プレゼント!	無料

表 8　C1 の注目語・主題語・廃棄語

注目語	主題語	廃棄語
イエローハット	車検	オートバックス
タイヤ	カー用品	オートバックスセブン
紹介	店舗	スーパーオートバックス
イエローハットグループ	運営	ご案内
ナビ	店舗情報	ドットコム
東京都	展開	クルマ
発表	販売	カー用品
国内	タイヤ	カーライフ
本社	com	サポート
企業	お店	車検

や「htb」などの「廃棄語」を NOT 検索に用いることで，先行クエリで検索されたオブジェクトを先読み検索から除去できると考えられる．一方,「テレビ北海道」や「タイヤ」などの「注目語」を用いると検索結果に余分な情報が含まれてしまったり，必要な情報が検索できない可能性がある．例えば,「テレビ北海道」を AND 検索に用いると，検索結果が「テレビ北海道」の情報のみに絞られてしまう恐れがあり，OR 検索に用いると「テレビ北海道 (TVH)」が検索結果に混ざり，“同じ主題に属する別の対象” を検索するという目的からはずれてしまう．また,「タイヤ」などを NOT 検索に用いると必要な情報まで除去してしまう恐れがある．そのため,「注目語」を用いない式 (9) は「比較」の先読みクエリとして一定の妥当性がある判断できる．

7.2　先読み検索の評価

表 10 のクエリ対を用いて，先読み検索を評価する．N3，N4

表 9　C2 の注目語・主題語・廃棄語

注目語	主題語	廃棄語
tvh	北海道	htb
テレビ北海道	放送	ハウステンボス
tvh 杯	テレビ	北海道テレビ放送
ジャンプ大会	札幌市	長崎
紹介	番組	htb クルーズ
葛西	イベント	北海道テレビ
ジャンプ	開催	on ちゃん
優勝	com	長崎県
札幌市	紹介	htb オンラインショップ
3 年ぶり	大倉山ジャンプ競技場	国内旅行

表 10　先読み検索の評価用クエリ対

	先行クエリ	後続クエリ
N3	時刻表	時刻表 jr
N4	子供の反抗期	中学生の反抗期
C3	古本市場 買取	ブックマーケット 買取
C4	大丸温泉	北温泉

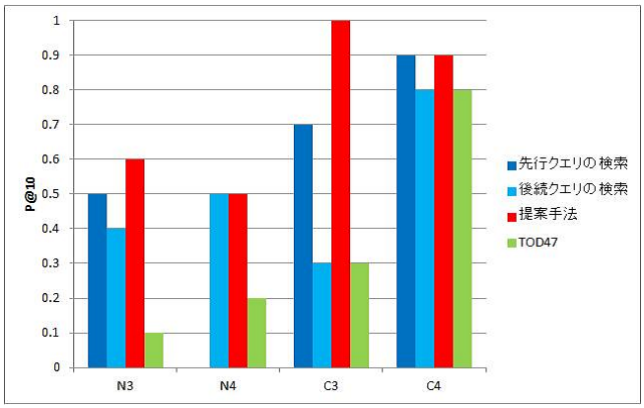

図 4　検索結果の適合率の比較

は「絞込」のクエリ変更意図であり，N3 は “JR の” 時刻表を検索しており，N4 は中学生の子供の反抗期の対処法について調べている．また，C3，C4 は「比較」のクエリ変更意図であり，C3 は古本屋の買取について比較しており，C4 は栃木県那須の温泉を比較している．これらのクエリを用いて，先行クエリによる検索，後続クエリによる検索および，本稿で提案した先読み検索の上位 10 件の適合率 P@10 を比較する．また，TOD47 で我々が提案した先読み検索 [8] の P@10 とも比較する．実験は 2012 年 2 月 24 日に行い，適合判定は第一著者が行った．それぞれの P@10 を図 4 に示す．なお，TOD47 の先読み検索はクエリの変化を「検索語の追加」,「検索語の削除」,「検索語の変更」の 3 パターン定義し，そのクエリ変化パターン毎に先読みクエリを生成する手法である．

TOD47 で報告した評価実験の結果，TOD47 の先読み検索は Google の提示する検索語候補と同等の適合率を示した．また，TOD47 で提案した先読みクエリは検索語を 3 語加えた場合が平均的に良い適合率を示したので，本稿では式 (8) および式 (9) の $n = 3$ として実験した．図 4 より，本稿の提案手法が最も高い適合率を示している．TOD47 の手法では「検索語の

追加」および「検索語の変更」の際に先読み検索があまり有効に機能しなかったが，本提案手法により改善できた．また，N4 は TOD47 では「検索語の変更」にあたり，絞込み検索という検索意図までは判断できなかったが，本手法ではそれを実現している．

N3 で本稿の先読み検索が生成した先読みクエリは「時刻表 jr AND (列車 OR jr 西日本 OR jr 東日本)」であり，バスなどの時刻表が混ざった検索から JR の時刻表に絞り込むことができたため，適合率が改善された．なお，TOD47 の先読み検索が生成した先読みクエリは「時刻表 jr AND (料金 OR 特急 OR 常磐線) NOT 検索 NOT 路線図 NOT バス」であり,「特急」や「常磐線」が適合率を下げる原因となっていた．また，N4 で生成された先読みクエリは「中学生の反抗期 AND (息子 OR 対応ガイドブック OR 高校生)」であった.「対応ガイドブック」が原因でガイドブックの販売促進のページが多く検索されたためあまり高い適合率とならなかった．なお，TOD47 の方法で生成した先読みクエリは「中学生の反抗期 AND (息子 OR 対応ガイドブック OR 小学生中学生思春期) NOT 子供 NOT 子供の反抗期 NOT 第一次」であり，同じく「対応ガイドブック」が入っていた．また,「子供の反抗期」が NOT 検索されたことにより適合率が下がっていた．

C3 では「(買取 AND cd AND ゲーム) NOT 古本市場 NOT 新品 NOT フルイチオンライン」という先読みクエリを生成しており,「買取 AND cd AND ゲーム」により，狙い通り古本市場とブックマーケット以外の古本屋を検索することができた．また，TOD47 の先読みクエリは「ブックマーケット　買取　AND　(エーツー　OR　店舗　OR　ゲオ) NOT 古本市場 NOT 古本 NOT 新品」であり，ブックマーケットの店舗の地図などの店舗情報が検索結果に多く含まれたため適合率を下げていた．C4 で生成された先読みクエリは「(温泉 AND 栃木県 AND 那須) NOT 大丸温泉旅館 NOT 奥那須温泉 NOT 那須温泉郷」であり,「温泉 AND 栃木県 AND 那須」により栃木県那須の温泉を検索できている．さらに,「大丸温泉旅館」が NOT 検索されたことで，大丸温泉以外の温泉宿を検索できた．TOD47 の先読み検索では「北温泉 AND (建物 OR 日帰り温泉 OR 天狗) NOT 大丸温泉旅館 NOT 大丸温泉 NOT 大丸」という先読みクエリが生成されており，北温泉の情報しか検索できていなかった．

8. ま と め

本稿では，クエリ変更意図の推測とそれに基づく先読み検索を提案した．まず，クエリログを分析してクエリ変更意図のクラスを定義した．そして，ユーザの連続した検索におけるクエリ変更意図を SVM を用いてそのクラスに自動分類した．また，定義したクエリ変更意図のうち「絞込」「比較」について，それらのクエリ変更意図を考慮した先読み検索を実装した．評価実験により提案した先読み検索はユーザの検索以上の適合率を示し，TOD47 の先読み検索 [8] よりも高い適合率となった．また,「比較」の先読み検索では比較対象を増やすような検索に成功した．よって，クエリ変更意図を推測して，そのクエリ変更意図毎に先読みクエリ生成法を変えることは，ユーザの検索意図に合致する先読み検索を行う上で有効であることが確認できた．

しかし，クエリ変更意図の自動分類が上手くできない場合もあり，自動分類の精度を高め，今回対応していないクエリ変更意図にも今後対応していく必要がある．分類精度を高める方法として学習データの追加，素性の追加の他に 3 回以上の連続した検索からユーザの検索意図を推測することなどが考えられる．また，先読みクエリの生成法も改善の余地があり，注目語・主題語・廃棄語の選択や追加する語数などを動的に変更することで，適合率の向上が図れるのではないかと考える．

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] 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.

[4] 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 Forum2008）, 5A-3, 2008.

[5] Xuehua Shen, Bin Tan, and ChengXiang Zhai. Context-Sensitive Information Retrieval Using Implicit Feedback. In *Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 43–50, 2005.

[6] Soo Young Rieh and Hong (Iris) Xie. Patterns and Sequences of Multiple Query Reformulations in Web Searching: A Preliminary Study. In *Proc. the ASIST Annual Meeting 2001*, pp. 246–55, 2001.

[7] 関口裕一郎, 杉崎正之, 内山匡, 藤村滋, 望月崇由, 鈴木智也. 検索クエリログを用いたクエリ変更意図の自動推定. 第 3 回データ工学と情報マネジメントに関するフォーラム（DEIM2011）, B2-3, 2011.

[8] 藤田遼治, 太田学. 検索質問と検索結果の変化を利用した先読み検索. 情報処理学会論文誌：データベース, Vol. 3, No. 3(TOD47), pp. 78–87, 2010.