DEIM Forum 2011 B2-3

検索クエリログを用いたクエリ変更意図の自動推定

関口裕一郎† 杉崎 正之† 内山 匡† 藤村 滋†† 望月 崇由††
鈴木 智也††

† 日本電信電話株式会社 NTT サイバーソリューション研究所 〒 239–0847 神奈川県横須賀市光の丘 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: †{sekiguchi.yuichiro,sugizaki.masayuki,uchiyama.tadasu}@lab.ntt.co.jp,
††{fujimura,mochizuki,to.suzuki}@nttr.co.jp

あらまし ユーザは検索エンジンに試行錯誤をしながらクエリを入力する．これらのクエリ変更のログは，ユーザのクエリに対する意図や検索結果に対する満足度といったを表すものとして，検索精度の向上やクエリ推薦に用いる情報源として有効であると考えられる．本論文では，クエリの変更意図を絞込，汎化，関連，修正，新規の 5 タイプに分類し，それらを自動分類する手法について取り扱う．文字・語句単位の差違による素性と検索結果スニペットから抽出した素性を用いて，先行クエリと後続クエリからなる複数クエリの組がどのクエリ変更タイプに該当するかを出力する分類器を SVM を用いて構築した．また，ローマ字や英語表記の読みモデルを用いた表記ゆれ照合を行うことにより，日本語特有の英語・平仮名・片仮名間の表記ゆれを考慮したクエリ分類手法を構築した．上記の手法を商用検索エンジンのログに対して適用し，精度 85.4%での分類が可能であることを確認した．
キーワード 検索意図，クエリ推薦，検索エンジン

Classification of Query Reformulation Intent using Search Engine Logs

Yuichiro SEKIGUCHI†, Masayuki SUGIZAKI†, Tadasu UCHIYAMA†, Shigeru FUJIMURA††,
Takayoshi MOCHIZUKI††, and Tomoya SUZUKI††

† NTT Cyber Solutions Laboratories, NTT Corporation Hikarinooka 1–1, Yokosuka-shi, Kanagawa, 239–0847 Japan
†† NTT Resonant Inc. Granpark Tower 8F, Shibaura 3–4–1, Minato-ku, Tokyo, 108–0023 Japan
E-mail: †{sekiguchi.yuichiro,sugizaki.masayuki,uchiyama.tadasu}@lab.ntt.co.jp,
††{fujimura,mochizuki,to.suzuki}@nttr.co.jp

1. はじめに

ウェブ閲覧中に，検索エンジンを用いてページを探す行為は極めて一般的になってきている．その一方で，自分の知りたいことを数語からなる検索ワードとして表現することは一般ユーザにとって容易でなく，検索ワードを試行錯誤しながら目的の情報を探し出そうとする行為がよく見られる．また，旅行の計画を立てている場合など，調べたい事柄が複数の要素から構成されている場合は，それぞれの要素に関連する言葉をそれぞれ検索し，それらを総合して解釈することにより必要とする情報を得ているケースも多く見られる．例えば，伊豆周辺へ旅行する際に寄る場所を調べたい場合において,「伊豆 名所」「伊豆 旅館」「伊豆 温泉」「熱海 温泉」といった関連する複数のクエリを連続して入力し，それらの検索結果を総合して最終的な行き先を定めること等がある．このように，ユーザの検索エンジンとのインタラクションは複数のクエリにまたがって行われ，検索エンジンのログにはその記録が蓄積されている．

上記に述べたような検索クエリ変更の内容は，ユーザが現在どういった事柄を求めているのかを知るのに有用な情報である．直前にユーザがクエリを変更した意図を把握することができれば，検索結果の精度やクエリ推薦等の技術の向上に有用な知見を得ることができると考えられる．例えば,「伊豆 観光」の後に「伊豆 旅館」といれたユーザは，伊豆に関する情報をいろいろと集めていると判断できるので,「伊豆 旅館 修善寺」といったクエリ以外にも,「伊豆 交通」や「伊豆 地図」といった伊豆に関する未検索のワードを推薦することが考えられる．

本論文ではユーザがクエリを変更して再検索する意図を自動推定する手法について扱う．連続して入力された 2 つのクエリ

表 1　文字編集によるクエリ変更のタイプ分類
Table 1　Text editing taxonomies of query reformulation

種別	Huang09 [1]	Teevan07 [3]	Bruza97 [2]
語順変更	-	word order	word reorder
語句追加	ADD	add word	add words
		duplicate words	
		add noun phrases or location	
		add stop word	
		add words	
語句削除	DEL	remove stop words	remove word
		remove noun phrases or location	
		remove word	
語句置換	SUB	word swap	word substitution
		reformulation	substring
		synonyms	superstring
	ABR	abbreviation	abrreviation
			acronym
ステミング	DER	stemming and pluralization	stemming
スペル修正	SPL	misspellings	spelling correction
スペース編集	SPE	extra whitespace	white space and punctuation
		word merge	
記号編集	PUN	non-alphanumeric	
URL 編集	-	domain	url stripping
繰り返し	REP	exact	-

表 2　変更意図によるクエリ変更のタイプ分類
Table 2　Intent based taxonomies of query reformulation

種別	Jansen07 [7]	Rieh01 [4]	Lau99 [5]	He02 [6]
絞込	specialization	specification	specialization	specialization
		special resource		
汎化	generalization	generalization	generalization	generalization
		special resource		
関連	reformulation	replace with synonym	reformulation	reformulation
語句修正	-	term variations	-	
		error correction		
		site url		
新規	new	-	new	new
繰返	content change	-	-	
推薦	assistance	-	-	relevance feedback
その他	-	operator usage	blank queries	others

を入力とし，そのクエリ変更意図をあらかじめ定義された 5 つのクエリ変更意図クラスへ自動分類することを目的とする．日本語の検索クエリログを解析対象として用い，サポートベクターマシン (SVM) を用いた機械学習手法で解くこととする．

2. 関 連 研 究

検索クエリの変更内容の分類については，主にユーザの検索行動のモデル化やクエリ推薦技術の知識源として使用すること目的として，いくつかの試みがなされている．これらの先行研究における分類は，文字編集内容を基準とした分類 [1] [2] [3] と，ユーザの変更意図を基準とした分類 [4] [5] [6] [7] の大きく 2 種類の分類体系が存在する．

2.1　文字編集内容を基準とした分類

文字編集内容を基準とした分類は，先行クエリと後続クエリの文字列における文字の追加，削除，置き換えといった書き換え作業に注目した分類方法である．このようなクエリ変更の分類は，ユーザの検索行動解析のための手法の一環として行われてきた．

例えば Huang ら [1] は，語句順変更，空白句読点変更といったクエリ変更の分類を定義し，それぞれのクエリ変更タイプとクリックスルーレート等の検索行動との関係性を分析している．また Teevan らはクエリの変更内容の分類を用いることにより，長期間にわたって繰り返し検索される内容の解析を行っている．

表 1 に，文字編集内容を用いたクエリ変更分類の代表的な例を挙げる．手法によって多少の差違はあるが，語句の追加・削除・置換といった語句レベルでの編集と，ステミング・スペル修正・スペースや記号の追加削除といった文字レベルでの編集とに 2 分されている．

これらの手法は各パターンを語句や文字の編集内容として厳密に定義できるため，分類ルールを整備することにより高い精度で自動分類が可能になる特徴がある．

2.2　文字編集内容を基準とした分類

もう一つのクエリ変更の分類定義として，ユーザがクエリを変更することによって，どのような検索結果を得ようとしていたかによる分類が存在する．これらを文字編集内容を基準とした分類と呼ぶこととする．これらの分類手法は，berry-picking-model に代表されるような比較的長期間における検索興味の変遷を分析するために使われてきた．

例えば Jansen ら [7] は，絞込 (specification) のクエリ変更タイプを「直前のクエリと同じ話題のクエリで，検索ユーザがより詳細な情報を求めている場合」と，検索ユーザの意図を用いて定義を行っている．また Rieh らによる分類 [4] は，まずクエリ変更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 の 3 つに大別し，Content には絞り込み，汎化，同義語，関連語の 4 タイプを，Format にはステミング，検索記号 (OR,NOT など) 追加，誤字修正の 3 タイプを，Resource には専門検索 (ニュース，画像，動画等) から一般検索への変更，一般検索から専門検索への変更，URL への変更の 3 タイプを定義している．

表 2 に，変更意図によるクエリ変更分類の代表的な例を挙げる．手法によって様々に差違があるが，絞込検索，汎化検索，関連検索，といった項目については共通している．また Rieh ら以外は，関連検索を同じ話題であるが絞込・汎化のどちらにも当てはまらないクエリ変更として定義している点に共通性がある．

これらの手法は編集内容に基づいた分類と比べ定義が難しく，

表 3　クエリ変更意図の 5 タイプ
Table 3　Taxonomies of query reformulation

種類	説明	例
Specification: 絞込	前のクエリより詳細で絞り込まれた情報を得るための変更	りんご→青りんご
Generalization: 汎化	前のクエリより幅広い情報を得るための変更	青りんご→りんご
Parallel: 関連	前のクエリと関連する情報を得るための変更	りんご→バナナ
Format: 語句修正	表記の揺らぎや，誤字脱字，入力変換ミスの修正	リンギ→リンゴ
New: 新規	新たな情報を得るための全く異なるクエリへの変更	りんご→花火大会

機械的に判別することも難しいが，連続した検索行動中において検索意図が変化した時点の抽出 [8] 等のユーザの意図に着目した行動解析に利用可能であるという利点がある．

2.3　本論文で用いるクエリ変更の分類体系

本論文ではユーザの意図に基づいたクエリ変更の分類体系を採用することとする．Rieh らによる分類を元とし，日本語検索クエリで一定割合以上見られる，絞込，汎化，関連，語句修正，新規の 5 タイプによる分類を採用する．表 3 にそれぞれのクラスの定義と例を表記する．

関連度が薄いクエリの組を，関連とするか新規とするかの基準としては，同一もしくは関連するトピックを示しているか否かで行うこととする．多義語については，その示す意味の 1 つがもう一方のクエリのトピックと関連している場合は，関連にクラス分けすることとする．また複数語から構成されるクエリでそれぞれの語句が異なるトピックである場合には，そのうち片方がもう一方のクエリのトピックと関連していれば，関連と判別することとする．トピックの関連性については，Open Directory プロジェクトの階層構造を参考にすることとし，兄弟関係や親子関係にある場合に関連していると判断することとする．

同一の検索クエリが繰り返し入力される場合については，クエリの一致を見ることで簡単に判別できるため，自動分類の項目外とした．Jansen ら，He らの分類に見られる推薦クエリからの選択については，推薦クエリの選択が絞込や関連といった何らかの意図に基づいて行われていると考え，別個の項目としては扱わないこととした．

3. 提 案 手 法

図 1 に提案手法の処理の概要を示す．提案手法はあるユーザによって連続して入力された 1 組の検索クエリを入力として，その検索クエリの組が表 3 に定義されたどのクエリ変更タイプに該当するか自動分類し出力する．入力として扱う連続したクエリのうち，先に入力されたクエリを先行クエリ，後に入力したクエリを後続クエリと呼ぶこととする．

行われる処理は大きく 2 段階に分かれており，1 段階目では入力されたクエリの組に対して表記ゆれの有無の判定を行い，表記ゆれ関係だった場合「語句修正」と判定する．また，クエリの一部分に表記ゆれ関係があった場合には，表記ゆれ部分を揃えるように置換を行い，後段の処理に回す．2 段階目では表記揺れが解消されたクエリから素性を抽出し，その内容を SVM に入力し，どのクエリ変更タイプかを分類する．

3.1　複数文字種間の表記ゆれ解消

日本語は平仮名・片仮名・漢字・英数字といった複数の文字種が混在するため，検索クエリにおいても多様な表記方法が存在する．例えば「DEIM2011」について調べる場合において,「デイム 2011」という片仮名による表記もあり得るし，片仮名表記を変換する前の「deimu2011」という表記が検索エンジンに入力されることも考えられる．

これらの表記揺れは，既存の表記揺れ解消アルゴリズムを用いることにより解決が可能であるため，SVM を用いた変更タイプ分類の前にクエリ間の表記揺れ関係を判定し，表記揺れの関係にあるものは「語句修正」と分類することとした．

今回表記揺れの判定には，英日音訳アルゴリズム [10] に用いている統計的音訳モデルを，ローマ字や平仮名・片仮名間の対応付けも可能なように拡張して用いた．このアルルゴリズムは統計的に文字種間の読みを対応付けさせるもので，例えば「chance」と「チャンス」いう 2 つの単語について，それぞれに対して統計的に音節の区切りを加え，その対応関係から読みの同一性を判定する．上の例であれば,「チャンス」は「チャン-ス」「チャ-ン-ス」等のいくつかの読みの区切りがあり得るが,コーパスから学習した英語と日本語の各音節間の対応確率を元に,「cha-n-ce」と「チャ-ン-ス」の対応付けをおこない，表記揺れの関係にあると抽出する．

片仮名・平仮名の音素は,「ヴ」と「う゛」のような一部の文字を除いて一対一で対応するため，コーパスを用いた確率モデルの学習は行わず，対応する対は全て等しく現れるという仮定で確率モデルを構築した．またローマ字と仮名の対応については，JIS X 4063 を元にいくつかのローマ字かな変換エンジンで用いられている変換規則を追加する形で構築した．これらについてもコーパスからの学習ではなく，等しく出現するという仮定に基づいて確率モデルを構築した．

また表記揺れのパターンには，組みとなるクエリを構成する一語のみが表記ゆれ関係になっている場合も存在する．例えば,「デイム 2011 会場」と「DEIM2011 プログラム」においては,「デイム 2011」と「DEIM2011」は表記揺れの関係にあるが,「会場」もしくは「プログラム」という追加語句が異なるので,検索クエリ全体としては表記ゆれ関係とはならない．このような場合に対応するために，クエリの一部の語句が表記ゆれ関係にあった場合は，対応する語句を片方に寄せるように置換した上で SVM による分類器に渡すように処理を行う．これにより，時節で述べる SVM に入力する素性が正しく取れるようになる．

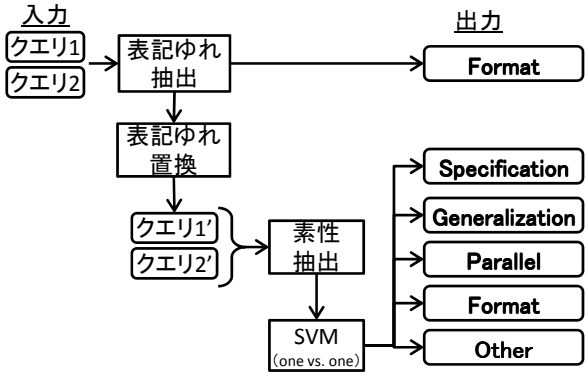

図 1　処理の流れ

Fig. 1　Process image

表 4　Jones ら [8] における素性一覧

Table 4　Features proposed by Jones et al.

素性	内容
Levenshtein	Levenshtein 編集距離
comm_char_l	左端からの共通字数
comm_char_r	右端からの共通字数
comm_word_l	左端からの共通語数
comm_word_r	右端からの共通語数
num_comm_word	クエリ全体の共通語数
jaccard	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

3.2　SVM を用いたクエリ変更意図の分類

表記揺れの抽出の結果「語句修正」タイプではないと判別されたクエリ対から，特徴量となる素性を複数抽出し，SVM を用いて 5 タイプの分類を行う．

学習に用いる素性は，Jones らによる検索セッションを分割する手法 [8] において提案されている語句変更内容と検索スニペットベクトルから素性群に，絞込・汎化の区別を行うための先行クエリ・後続クエリ順序関係に対して方向性のある素性を追加して用いた．

表 4 に Jones ら [8] による素性の一覧を示す．この素性群はクエリ対の表記上の関係性を表す文字内容を元にした素性 (levenshtein から jaccard まで) と，各クエリの意味的な関係性を表す検索結果のスニペットを元にした素性 (cosine) の大きく 2 種類が存在する．

素性 cosine に用いるスニペットベクトルは，先行クエリ・後続クエリそれぞれでウェブ検索を行った結果の上位 N 件のスニペットから構築する．スニペットを形態素解析し，得られた名詞を構成要素として bag-of-words ベクトルを構築する．ベクトルの重みは，その要素を含むスニペット数を tf とみなし，全検索クエリ数 N 中でのその要素を含むクエリ数を df と見なして計算した tf-idf 値を用いる．よってクエリ q のスニペットベクトルの語句 t に対する要素の値 $v_q(t)$ は，式 1 のようになる．

$$v_q(t) = tf_q(t) \times \log \frac{\mathrm{N}}{df_t} \tag{1}$$

絞込・汎化といった変更意図を判別するためには，先行クエリと後続クエリのどちらが意味的に絞り込まれているかという情報が重要である．例えば，一般的に多くの語句が含まれるクエリの方が，絞られた検索結果を返すと考えられる．その為，先行クエリに対して後続クエリの語数が増えたのか減ったのかという情報はクラス分類に置いて重要な手がかりである．上記の Jones らによる素性は，どれも先行クエリと後続クエリを入れ換えても値が同じになる対照的な素性であるため，このような情報が抜け落ちてしまっている．それに対応するため，語句要素，ベクトル要素共に先行クエリと後続クエリの組み合わせに対して非対称な素性を導入する．導入した非対称な素性の一覧を表 5 に示す．

表 5　非対称な素性の一覧

Table 5　Asymetric features

素性	内容
word_add	後続クエリに追加 (削除) された語句数
char_add	後続クエリに追加 (削除) された文字数
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

文字内容に関する非対称な素性として，語の増減，文字の増減を用いる．語の増減は先行クエリに対して後続クエリで語句が追加された場合に正の値を，削除された場合には負の値をとるものとする．文字の増減についても同様に正負の値を取るものとする．また，文字の増減については空白もカウントする事とする．例えば，先行クエリが「伊豆」で後続クエリが「伊豆 修善寺」ならば，word_add は+1，char_add は+4 となる．また，先行クエリが「伊豆 修善寺」で後続クエリが「DEIM2011」のように共通語句が存在しない場合には，word_add，char_add ともに 0 とする．

スニペットベクトルの類似関係に対する非対称な素性として，Skew divergence [9] を用いる．Skew divergence は式 2 に示すように，Kullback-Leibler 距離 $D(q\|r)$ をスムージングした距離指標であり，0 要素を含む bag-of-words ベクトルに対しても適用可能な特徴がある．

$$s_\alpha = D(r\|\alpha q + (1-\alpha)r) \tag{2}$$

$$\mathrm{skew}(q\|r) = \sum_y q(y)(\log q(y) \log r(y)) \tag{3}$$

Kullback-Leibler 距離を元としているため，ベクトル r とベクトル q が同一の際に値が 0 となる．またベクトル r の方がベクトル q に比べて確率分布が広がっている場合にベクトル r からベクトル q への距離 skew($q\|r$) が，ベクトル q からベクトル r への距離 skew($r\|q$) より大きくなり，非対称な値を取る．検索結果が絞り込まれていれば，スニペットに含まれる語句のバリエーションも狭まると考えられるので，先行クエリに対して後続クエリが絞込の意図で入力されている場合に skew2 が skew1 より大きくなり，先行クエリを汎化する意図で後続クエリが入力されている場合に skew1 が skew2 より大きくなる．また「伊豆」と「熱海」のような関連クエリに対しては skew1 と skew2

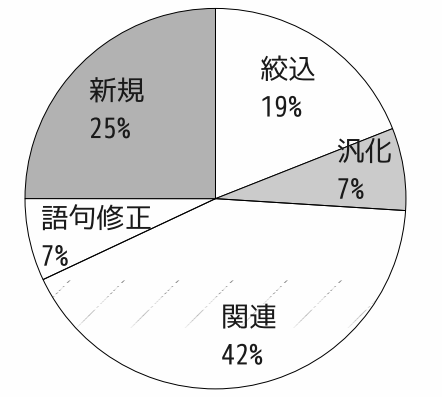

図 2　正解データ中の各パターン割合
Fig. 2　Proportion of reformulation classes

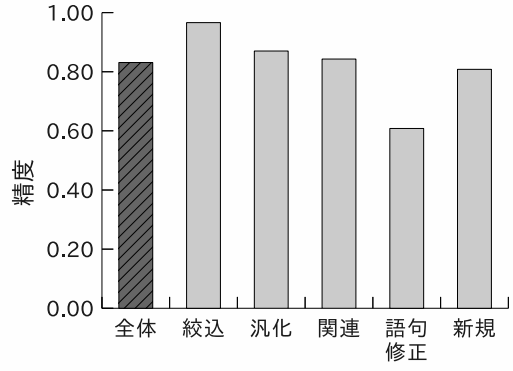

図 3　クラス分類の精度
Fig. 3　result of classification

がほぼ同じ値となると期待される．

4. 評 価 実 験

商用の検索エンジンに入力された検索クエリログデータを用いて，提案手法の評価実験を行った．

処理対象データとして商用検索エンジンに 2010 年 4 月〜6 月に入力された 10 名の検索クエリログを用いた．ログデータをユーザ毎に分割した上で同一のクエリが連続して入力されている場合を除いて，308 のクエリ遷移のペアを作成抽出し，それに人手で 5 タイプの正解付けを行った．正解付けは 2 名の作業者で行い，作業者間の正解付けに差異が出た部分に関しては，作業者同士で協議の上で一方を採用することとした．作成した正解におけるクエリ変更タイプの分布を図 2 に示す．

得られた 308 のクエリペアに対して各素性の値を求め，5 分割交差検定で精度評価を行った．分類器には LIBSVM を用いた SVM を構築し，カーネルは RBF を用いた．また，各クエリのスニペットベクトルの作成には，検索結果上位 30 件のスニペットを用いた．

実験結果の全体での精度と各クラスごとの精度を図 3 に示す．

全体での精度は 0.831 となり，クラス別では絞込が 0.966，汎化は 0.870，関連は 0.843，語句修正が 0.608，新規は 0.808 とそれぞれなった．最も分類ミスが多かった語句修正における精度低下の要因としては，平仮名・片仮名・漢字・英文字と複数の文字が混在する日本語特有の複雑な表記ゆれによるものが多かった．例えば，平仮名を英語に書き換えた場合には共通文字語数，共通文字数共に 0 になり，表記の変更により検索結果も大きく変わっているとスニペットベクトルの内容も異なってしまうので，各素性の値が新規の場合と似通った内容になるというパターンが見られた．関連と新規の間での分類ミスが次に多い精度低下要因だった．これは先行クエリと後続クエリが関連語句であっても得られるページの傾向が大きく異なる場合などに，2 つのスニペットベクトル中の共通語句が少なくコサイン類似度や Skew divergence で十分な類似度がでないことにより起きていた．

各素性の分類精度へのインパクトを確認するために，全素性から個別に素性を除いてみての精度測定を行った．その結果を表 7 に示す．両端からの共通語数 (comm_word_l と comm_word_r)，両端からの共通文字数 (comm_char_l と comm_char_r) および Skew divergence(skew1 と skew2) については，効果を分かり易く見るために 2 つ同時に除いた場合の精度を記載している．

文字数増減および Skew Divergence を除去した時に精度が大きく低下しており，非対称な素性が有効に働いていると考えられる．また対象な素性については，各端からの共通文字数以外については除去することにより精度が向上しており，有効に機能していないと考えられる．

最も精度が良くなった素性の組み合わせは，共通語数と語句増減数，Skew divergence を用いた場合であり，精度が 0.854 となった．全素性を用いた場合に比べ，0.023 の向上となっている．このことからも，非対称な素性がクエリ変更意図の分類に有効であると考えられる．

4.1 表記ゆれ対応による効果

SVM のみで分類した場合と，SVM に表記ゆれ抽出を加えた場合，SVM，表記ゆれ抽出およびクエリの一部に表記揺れがあった際の表記ゆれ置換の全てを行った場合についての精度の比較を実施した．その結果を表 6 に示す．また，表記ゆれ対応を行った場合の各クエリ変更タイプごとの精度の値を，図 4 に示す．

表記揺れ対応を行うことにより精度が 0.812 から 0.831 となり，若干の向上がおこなわれていることが分かる．また図 4 から，表記揺れ対応によって，語句修正の精度が 0.261 から 0.609 と向上している．これは，英語と日本語間の表記揺れ関係にあるクエリは，語句や文字の内容は全く異なっているため，何も処理を加えずに SVM 用の素性抽出を行うと，左右からの共通文字・語句数や，編集距離といった表記内容に依存する素性が極端に小さい値となり，学習及び分類に悪影響を与えているためと考えられる．

また，今回は素性抽出前に表記揺れを置き換える処理を行ったが，それによる効果は限定的であった．クラス別の精度を見ても，Specification では若干の精度向上がみられたが，全体的にはほぼ差が出なかった．これは，表記揺れの置換が行われるクエリ自体の件数がすくないことに合わせて，置換されていない部分の語句の情報がすでにあるため，表記揺れを補正することによるクエリ全体に対する影響が比較的小さいことによると

表 6　表記揺れ対応を行った際の精度

Table 6　Precision at abbreviation correction

	1.SVM のみ	2.SVM ＋表記ゆれ抽出	3.SVM ＋抽出＋置換
精度	0.812	0.825	0.831

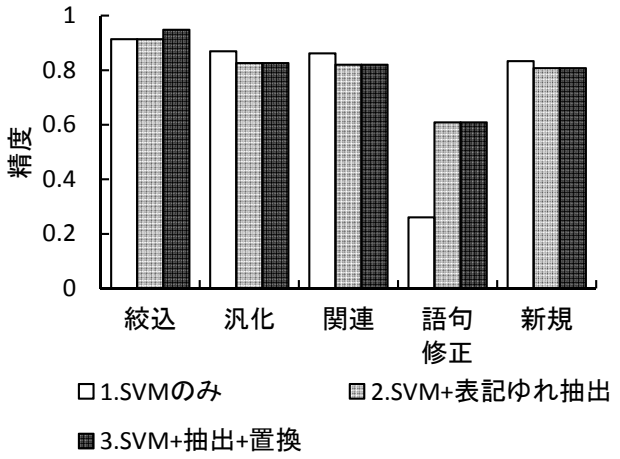

図 4　クエリ変更タイプ毎の精度

Fig. 4　Impact of abbreviation correction on each class

表 7　各素性を抜いた場合の精度

Table 7　precision for removing one feature

除去した素性	精度
none	0.831
Levenshtein	0.841
comm_word_l, comm_word_r	0.851
comm_char_l, comm_char_r	0.831
num_comm_word	0.851
jaccard	0.838
word_add	0.841
char_add	0.796
cosine	0.841
skew1, skew2	0.799

考えられる．

5. おわりに

本論文では，ユーザのクエリ変更意図に基づいた，絞込・汎化・関連・誤字修正・新規の 5 クラスからなるクエリ変更の分類を定義し，同一ユーザによって連続して入力された 2 つの検索クエリに対して，SVM を用いてクエリ変更のクラスを自動分類する手法について述べた．語句，文字の増減および Skew divergence による非対称な素性を提案し，商用の検索エンジンに入力された検索ログを用いた実験により精度 0.854 で分類可能であることを示した．また，表記揺れの補正を行うことにより，語句修正のクエリ変更クラスにおいて抽出精度が 0.348 向上することを確認した．

今後の課題としては，語句修正クラスについての分類精度の向上に取り組みたい．日本語は複数の文字種からなるため，複雑な表記揺れが存在する．特に検索エンジンに入力されるクエリは，インプットメソッドを介してローマ字から漢字への変換を通して作成されるため，最初のローマ字の部分で入力ミスした場合などでは完全に異なる語句が入力される．例えば「研究会 (kenkyuukai)」を「kenkyiukai」とタイプミスすると「剣きぃうかい」という表記となり，今回の表記揺れ抽出処理では対応が不可能な形となる．入力変換前のアルファベットの共通文字数を素性に入れるなどすることにより，これらに対する精度向上が考えられる．

また，今回構築したクエリ変更意図の自動分類結果を用いた，クエリ推薦技術の構築にも取り組みたい．直前のクエリ変更の意図が分かれば次に求める情報の傾向も絞れるため，絞込を行った後のユーザには関連のクエリ変更タイプになるクエリを推薦するといった手法を構築していきたい．

文　献

[1] Huang, J., Efthimiadis, E. N., ”Analyzing and evaluating query reformulation strategies in web search logs,” In Proceeding of the 18th ACM conference on Information and knowledge management, pp. 77–86, 2009.

[2] Bruza, P. D., Dennis, S., ”Query Reformulation on the Internet: Empirical Data and the Hyperindex Search Engine,” In Proceeding of 5th RIAO Conference, pp. 488–499, 1997.

[3] Teevan, J., Adar, E., Jones, R., Potts, M. A. S., “Information Re-Retrieval: Repeat Queries in Yahoo’s Log,” In Proceedings of the 30st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 151–158, 2007.

[4] Rieh, Soo Y. and Xie, Hong, ”Patterns and Sequences of Multiple Query Reformulations in Web Searching: A Preliminary Study,” In Proceedings of the ASIST Annual Meeting 2001, pp. 246–55, 2001.

[5] Lau, Tessa and Horvitz, Eric, ”Patterns of search: analyzing and modeling Web query refinement,” In Proceedings of the seventh international conference on User modeling 1999, pp. 119–128, 1999.

[6] He, D., Goker, A., Harper, D. J., ”Combining evidence for automatic Web session identification,” Information Processing & Management, vol. 38, no. 5, pp. 727–742, 2002.

[7] Jansen, B. J. and Spink, A. and Blakely, C. and Koshman, S., ”Defining a session on Web search engines,” Journal of the American Society for Information Science and Technology, vol. 58, no. 6, pp. 862–871, 2007.

[8] Jones, R., Klinkner, K. L., ”Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs,” Proceeding of the 17th ACM conference on Information and knowledge management, pp. 699–708, 2008.

[9] Lee, L., ”On the effectiveness of the skew divergence for statistical language analysis.,” In Proceedings of Artificial Intelligence and Statistics 2001, pp. 65–72, 2001.

[10] 齋藤邦子, 篠原章夫, 永田昌明, 小原永: “音声制御ブラウザ VCWeb の英日シームレス化,” 人工知能学会論文誌, vol.17, no.3, pp.343-347, 2002.