情報処理学会論文誌　データベース　Vol. 3　No. 3　33–45 (Sep. 2010)

概念ベースに基づくWeb検索のクエリタイプ判定手法とその評価

廣嶋 伸章[†1]　戸田 浩之[†1,*1]
松浦 由美子[†1,*2]　片岡 良治[†1]

Web検索において，あるクエリが入力された際に，そのクエリの種別を知ることができれば，それに応じてシステムの応答を変化させることが可能となり，適切な検索結果を提示することができる．たとえば，あるクエリの種別が「グルメ」であることが分かれば，レシピ検索とブログ検索の結果を提示することができる．このようなシステムの応答を変化させるための条件であるクエリの種別をクエリタイプと呼ぶことにする．クエリの属するクエリタイプを知ることで，上で述べたような利便性の高い検索サービスが実現できる．そこで本論文では，様々なクエリに対してクエリタイプを判定する手法を提案する．提案手法では，単語に対してその単語の分野を表す概念ベクトルが付与された概念ベースを参照して，クエリに関する文書から得られたクエリ分野ベクトルと各クエリタイプ分野ベクトルとのコサイン距離に基づきクエリタイプを判定する．実験では，27のクエリタイプに対し，提案手法単独で64.6%，Wikipediaなどの情報を利用した手法を組み合わせることにより77.1%の精度で判定を行うことができた．

A Query Type Inference Method Using Concept Base and Its Evaluation

Nobuaki Hiroshima,[†1] Hiroyuki Toda,[†1,*1]
Yumiko Matsuura[†1,*2] and Ryoji Kataoka[†1]

In web search, identifying categories of the given query is useful in enhancing the effectiveness, because the system can display appropriate results according to the query's category. For example, if the system knows the given query can be classified as a category "Food", it can display the result of the recipe search at the higher position. We call these categories "query types". We propose a method for inferring the query type or types for each query. The method employs the concept vectors obtained from a concept base. A query vector and query type vectors are generated from concept vectors and the similarity between the query vector and each query type vector are computed. Experimental results show that the proposed method achieves 64.6% accuracy and the combined method including the proposed method achieves high accuracy, 77.1%, in inferring 27 query types.

1. はじめに

情報検索サービスは，ある物事についての詳細な情報を調べたり，行動を起こしたりするうえで必要不可欠なものとなってきている．情報検索では，クエリを入力し，クエリに関するテキストや画像などの情報を検索する．情報検索を行うためのクエリとしては自然言語・音声・画像など様々なものが利用できるが，現在最も広く用いられているクエリは自然言語で表現されたものであり，その中でも特に単語や複合語を用いるのが一般的である．そこで本論文では，単語や複合語をクエリとして取り扱う．

情報検索サービスをより人々にとって利便性の高いものとするためには，人々が入力するクエリなどから検索意図を汲み取り，その検索意図に応じて適切な検索結果を提示することが考えられる．そのため，検索意図に応じた検索を行うための様々な研究が行われている[1)–3)]．検索意図に応じた検索を行うための方法の1つとして，クエリがある特定の種別に該当していればシステムの応答を変化させるということが考えられる．このようなシステムの応答を変化させるための条件であるクエリの種別をクエリタイプと呼ぶことにする．クエリの属するクエリタイプを知ることで，たとえばクエリタイプが「地名」であればその地理情報を考慮した検索に切り替えるというように，クエリタイプに応じて検索方法を切り替えることが可能となる．また，たとえばクエリタイプが「グルメ」であればレシピ検索とブログ検索の結果を提示するというように，クエリタイプに応じて様々な専門検索の中から適切な専門検索を選択してその結果のみを提示することが可能である．さらに，たとえばクエリタイプが「企業名」であれば被リンク数を重視したランキングを行うというように，クエリタイプに応じてランキング関数の変更を行うことも可能である．このように，クエリの属するクエリタイプが分かれば検索意図に応じた質の高い情報検索サービスが実現でき

†1 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
NTT Cyber Solutions Laboratories, NTT Corporation
*1 現在，NTT コミュニケーションズ株式会社
Presently with NTT Communications Corporation
*2 現在，NTT サービスインテグレーション基盤研究所
Presently with NTT Service Integration Laboratories

ると考えられる．特に，絞り込みのためのクエリを追加して再検索を行ったりするのが比較的容易な PC での検索に比べ，クエリの入力に労力がかかり操作が煩雑になりがちな携帯端末での検索においては，大幅な質の向上が期待できる．そこで本論文では，携帯端末における質の高い情報検索サービスの実現に向けて，様々なクエリがどのクエリタイプに属するかを自動的に判定することを目的とする．なお，本論文では，人によらずほぼ共通の集合を連想させるものをクエリタイプとして扱う．たとえば，「芸能人」は，人によらずテレビや映画などに出演する人物の集合を連想させるものであるため，クエリタイプとして扱う．一方，「友人」は，人によって連想する集合が異なり，個人に関する情報が必要となるため，本論文の範囲外とする．

クエリタイプを判定するにあたり，大きく分けて以下の 3 点が課題として考えられる．

(1)　クエリに対する網羅性
まず，クエリの分布はロングテールであるため，頻度の少ないクエリであっても正しくクエリタイプが判定できる必要があり，クエリに対する網羅性の高さが要求される．

(2)　様々な粒度のクエリタイプの定義
クエリタイプは「人名」「地名」「組織名」といった従来の固有表現抽出において付与されてきたレベルの細かさでは検索意図を十分に反映できない．たとえば，クエリがある人物を表している場合でも，それが芸能人であるかスポーツ選手であるかによって提示する検索結果を変えたいということが想定される．芸能人はさらに細かく俳優やミュージシャンなどに分類したいという場合も想定される．よって，クエリタイプに応じて検索システムの動作を制御できるようにするために，様々な粒度のクエリタイプを定義して分類できる必要がある．

(3)　クエリタイプスコアの付与
クエリによっては複数のクエリタイプに判定されることが適切である場合が存在する．たとえば，テレビで放映されて人気が出たため映画化されたアニメや，映画で上映されて人気が出たため続編がテレビドラマ化された物語は，クエリタイプとして「テレビ番組」「映画」の両方に判定されることが望ましい．しかし，前者は「映画」よりも「テレビ番組」としての性格が強く，後者は「テレビ番組」よりも「映画」としての性格が強いと考えるのが自然であり，両者は区別されるべきである．そのため，クエリタイプ判定のタスクにおいては，それぞれのクエリタイプに属するかどうかとともに，そのクエリタイプらしさを表すスコアが付与されることが求められる．クエリタイプにスコアが付与されることで，たとえばクエリタイプ「映画」のスコアよりもクエリタイプ「テレビ番組」のスコアが高ければ，検索結果の最上部にテレビ番組に関する検索結果を提示し，その下部に映画に関する検索結果を提示するというような利用が可能となる．

クエリタイプ判定の関連研究においては，これらの課題をすべて解決できる手法は提案されてこなかった．本論文では，単語に対してその単語の分野を表す概念ベクトルが付与された概念ベースを利用した手法を提案する．詳細については 4.1 節に示すが，概念ベースを利用することで，たとえば「首相」と「総理」のように類似した概念を持つ単語に類似したベクトルを割り当てることが可能となる．この概念ベースを参照して，あらかじめ各クエリタイプの分野を表現するクエリタイプ分野ベクトルを生成しておき，入力となるクエリに関する文書から算出したクエリ分野ベクトルと各クエリタイプ分野ベクトルとのコサイン距離に基づきクエリタイプを判定する．これにより，出現頻度が少ないクエリを含めた幅広いクエリに対して網羅的に適用可能であり，任意のクエリタイプを定義可能であり，かつクエリタイプとともにクエリタイプらしさを表すスコアを付与することが可能となる．

以下，2 章ではクエリタイプ判定に関連する研究について述べる．3 章では学習データを利用することにより実現できるクエリタイプ判定手法およびそれらの手法を用いた評価について述べる．4 章では提案手法において利用する概念ベースおよび概念ベースを用いたクエリタイプ判定手法について述べる．5 章では提案手法の有効性を検証するための実験について述べる．

2. 関 連 研 究

クエリタイプ判定に関する研究は，扱うタイプにより，検索意図に基づくクエリタイプ判定と意味に基づくクエリタイプ判定に大別することができる．検索意図によるクエリタイプ判定では，クエリに対し informational，navigational などのタイプを判定する研究[4)–6)]や，クエリに対しどのような結果を求めているか（ニュースを読みたい，画像を検索したいなど）のタイプを判定する研究[7)] が行われている．特に，informational か navigational かのタイプを判定する研究では，単語の分布や相互情報量などに基づく手法[8)] や，アンカテキストごとのリンク先の分布に基づく手法を任意のクエリが扱えるように拡張した手法[9)] などが提案されており，タイプに応じて検索のモデルを切り替えることにより検索精度の向上に成功している．一方，本論文では，クエリに対し「芸能人」「映画」「グルメ」などのタイプを付与するような意味に基づくクエリタイプ判定を扱う．

意味に基づくクエリタイプ判定の研究は，その手法により，コーパスから語彙とクラスの

関係を抽出する手法，検索エンジンのクエリのタイプを直接推定する手法に大別することができる．以下では，それぞれの手法ごとの関連研究とその問題点について述べる．

2.1　コーパスから語彙とクラスの関係を抽出する手法

コーパスから語彙とクラスの関係を抽出する手法は，語彙とクラスのペアを抽出する方法と，特定のクラスに対して語彙を抽出する方法の 2 つに分けることができる．

2.1.1　語彙とクラスのペアを抽出する手法

隅田らは，Wikipedia から上位下位関係にある語彙のペアを取得している[11]．Wikipedia のテキストの構造を利用して上位下位関係にある可能性のある候補を抽出し，SVM を用いて上位語および下位語に関する様々な素性をもとに正しい上位下位関係にあるかどうかを判定している．また，Paşca は，膨大な Web 文書の中から “X such as N” や “X including N” のようなパターンにマッチする名詞句 N と X をそれぞれ語彙とクラスとする手法を提案している[12]．取得した語彙とクラスをもとにパターンの拡張も行われている．Shinzato らは，HTML 文書の構造や df などの統計量などを用いて HTML 文書から単語間の上位下位関係を獲得している[13]．しかし，これらの手法では，抽出されるクラスがテキストの記述に依存するため，ほとんどのクラスは定義したクエリタイプと一致しない．たとえば，「ブロッコリー」という語彙に関して「緑黄色野菜」というクラスが抽出されたとしても，定義した「グルメ」というクエリタイプと一致しないことが考えられる．語彙をクエリタイプに結び付けるためには，たとえば「緑黄色野菜」→「野菜」→「食べ物」→「グルメ」といったような複数の関係がうまく抽出されていることが必要となる．しかし，このような関係がすべて抽出できる例はあまり多くないことが予想される．よって，クエリタイプ判定に適用した場合に網羅性の点で問題が生じると考えられる．

2.1.2　特定クラスの語彙のみを抽出する手法

Shinzato らは，上位下位関係の獲得とは別に，HTML 文書中に現れる箇条書きとその表題を用いて，あらかじめ指定された上位語に対する下位語の獲得も行っている[14]．上位語としてクエリタイプを指定して下位語を獲得することで，そのクエリタイプに属するクエリの集合を得ることが可能である．しかし，獲得できる語は直接的な下位語に限定されるため，本来クエリタイプに属するクエリとしたい語を獲得できない可能性がある．たとえば，「浅田真央」を「スポーツ」というクエリタイプに分類したいと考えていても，この手法で「スポーツ」から獲得できる下位語は「野球」「サッカー」などであるため，「浅田真央」は獲得できない．新たな上位語として「スポーツ選手」を設けることで獲得できる範囲が広がるが，クエリタイプごとにこのような上位語を漏れなく用意するのは難しいと考えられる．よって，クエリタイプ判定に適用した場合に網羅性の点で問題が生じると考えられる．

小町らは，用語間の 2 項関係を抽出するための Espresso アルゴリズム[15] をもとに同一クラスに属する語彙の抽出に適用するための改良を行った Tchai アルゴリズムを提案している[16]．検索ログを用いて語彙とパターンの抽出をブートストラップにより取得することで，ある意味カテゴリに属する少量の語彙の集合から同じ意味カテゴリに属する語彙の集合を獲得している．しかし，この手法により様々なクラスに対して語彙を取得した場合，複数のクラスに同一の語彙が出現する場合があると想定されるが，そのような場合にどちらのクラスがそのクエリに対する代表的なクラスであるのかを判断するのは難しい．よって，クエリタイプ判定に適用した場合にクエリタイプらしさを表すスコアを付与できないという問題が生じると考えられる．

2.2　検索エンジンのクエリのタイプを推定する手法

Vallet らは，入力のクエリを含む Wikipedia 記事を検索して検索結果上位の記事に含まれる固有表現をもとに固有表現タイプを決定している[10]．固有表現に対してアノテーションが行われたコーパスを用いて統計的な固有表現抽出器を作成し，これを Wikipedia の各記事に適用して固有表現を抽出して記事に付与している．しかし，この方法では，取得できる固有表現タイプは固有表現抽出器が出力可能な固有表現タイプに限られるため，現状の固有表現抽出器では MUC[17] や IREX[18] で定義された「人名」「地名」「組織名」というようなレベルでの判定しか行えない．大規模な固有表現タイプとして拡張固有表現階層が提案され，これに基づく固有表現タガーも作成されている[19)–22]．しかし，この拡張固有表現タイプをクエリタイプとして用いた場合，必要とされるクエリタイプに対応する拡張固有表現タイプが存在しないという状況が考えられる．たとえば，「芸能人」というクエリタイプを考えた場合，それに対応する拡張固有表現タイプは存在しない．よって，クエリタイプ判定に適用した場合に様々な粒度のクエリタイプを定義して分類することができないという問題が生じると考えられる．

Guo らは，クエリを固有表現とその固有表現の出現する文脈と固有表現タイプの 3 つ組であると見なし，クエリから考えられる様々な 3 つ組の中から同時確率が最大となるものを求めることにより，クエリに含まれる固有表現のタイプの推定を行っている[23]．同時確率を計算するために必要な各種確率は，少量の固有表現と固有表現タイプの組および検索ログを用いて推定している．しかし，この方法では，固有表現に関連する確率を求める際に，既知の文脈を含むクエリから文脈を取り除いた部分のみを固有表現として扱うため，検索ログ中にあまり出現しない固有表現に対して正しくパラメータを推定できず，固有表現タイプ

を求めることができない．よって，クエリタイプ判定に適用した場合に網羅性の点で問題が生じると考えられる．

Beitzel らは，ラベルつきの訓練データとラベルなしの訓練データをもとに，クエリ中の単語との一致およびクラスの個数分のパーセプトロンを組み合わせた手法を提案している[24]．しかし，人名などのクエリに含まれる単語のようにそのクエリ中でしか用いられない場合があるため，分類が不可能なクエリが存在すると考えられる．よって，クエリタイプ判定に適用した場合に網羅性の点で問題が生じると考えられる．

Broder らは，クエリに関する文書を取得し，文書分類器を用いて各文書に対して文書がそれぞれのクラスに分類される確率を求め，それらの確率から重みつき多数決や確率を素性とした 2 値分類手法によりクラスを求める手法を提案している[25]．しかし，文書分類器の学習を行う際にクラスラベルが付与された文書を大量に用意する必要があるだけでなく，クエリタイプに変更が行われた場合にはそれに応じて文書のクラスラベルを更新する必要があり，膨大なコストがかかる．よって，様々な粒度のクエリタイプを定義して分類することが難しいという問題が生じると考えられる．

3. 学習データを用いたクエリタイプ判定手法

クエリに対して人手によりクエリタイプを付与することで，ある程度の量であれば正解データを作成することができる．この正解データの一部を学習データとしてクエリタイプの判定に必要な情報を獲得し，これらの情報および既存の情報源を用いることによりクエリタイプを判定する単純な手法がいくつか考えられる．以下では，学習データを用いたクエリタイプ判定手法の評価を行い，問題点を示す．

3.1 クエリ中の単語の利用

まず，学習データを用いて，特定のクエリタイプによく出現するクエリ中の単語を知ることが可能である．そこで，クエリタイプごとにまとめたクエリ中の単語の頻度を調べて頻度がある閾値を超えた単語を頻出語として抽出しておき，実際に入力されたクエリ中に頻出語が出現した場合はそれに対応するクエリタイプを割り当てるという手法が考えられる．これにより，たとえば「動物占い」などの「占い」という頻出語を含むクエリに対して精度良くクエリタイプ「占い」を割り当てることが可能となる．

3.2 Wikipedia の情報の利用

Wikipedia においては，記事のタイトルは 1 つの用語であり，本文はその用語に関する説明であるが，その説明の中でも**図 1** の上部にある例のようなリード文は用語を端的に説明

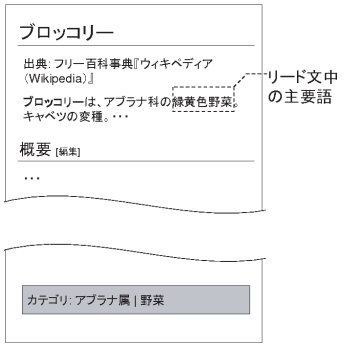

図 1　Wikipedia 記事の例
Fig. 1　Example of Wikipedia document.

した文となっている．リード文の多くは，「A は，… B である．」という形式で書かれており，B は A がどのようなタイプに属するかを強く指し示すものとなっている．この B を主要語と呼ぶことにすると学習データを用いて主要語とクエリタイプの組を収集し，クエリタイプ判定に利用することができると考えられる．具体的には，クエリに対してクエリタイプが付与された学習データを用意し，学習データ中の各クエリをタイトルに持つ Wikipedia 記事を取得して主要語を取得する．これをすべての学習データに対して行い，主要語が与えられた場合に各クエリタイプに属する確率を計算しておく．入力としてクエリが与えられた場合には，クエリをタイトルに持つ Wikipedia 記事を取得し，記事のリード文に含まれる主要語を抽出する．そして，その主要語が与えられた場合に各クエリタイプに属する確率が最大となるようなクエリタイプを選択することでクエリタイプの判定が行える．

3.3 学習データを用いたクエリタイプ判定手法の評価

上記で述べた学習データを用いたクエリタイプ判定手法により，クエリに対してどの程度正しくクエリタイプが付与できるかを確認するための実験を行った．

3.3.1 実 験 環 境

学習データを用いたクエリタイプ判定手法における学習や評価を行うために，正解データを作成した．国内のモバイル向け商用検索エンジンで 2008 年の 1 年間に高い頻度で検索された検索クエリ上位 30,000 件に対して，1 人の評価者が人手により必要とされるクエリタイプの定義を行い，定義されたクエリタイプをもとに 5 人の評価者が人手によりクエリタイプを付与した．クエリタイプの定義および付与にあたって評価者に提示した情報および評

表 1　用意したクエリタイプ
Table 1　Manually prepared query types.

クエリタイプ	クエリの例	クエリタイプ	クエリの例	クエリタイプ	クエリの例	クエリタイプ	クエリの例
芸能人	SMAP	スポーツ	浅田真央	作家	村上春樹	アナウンサー	筑紫哲也
政治家	東国原英夫	企業家	孫正義	キャラクター	キティちゃん	ゲーム	ドラクエ
書籍	たまごクラブ	車名	ヴィッツ	PC 関連	Windows	家電	デジカメ
地名	横浜	企業名	NTT 東日本	サイト名	mixi	映画	タイタニック
番組名	笑っていいとも	曲名	グリーングリーン	イベント名	クリスマスイブ	俗語	イケメン
健康/医療	花粉症	グルメ	しゃぶしゃぶ	ギャンブル	宝くじ	着メロ	発車メロディー
画像	風景画	占い	タロット	アダルト	（省略）		

価者に指示した内容は以下のとおりである．

(1)　クエリタイプ定義時

(a)　評価者に提示した情報

クエリの集合

(b)　評価者に指示した内容

どのような検索結果を提示したいかを考慮してクエリタイプを定義する．たとえば，俳優や芸人やミュージシャンの場合は画像検索や出演情報などの結果を提示できればよいと考え，個別のクエリタイプとはせずにまとめて「芸能人」というクエリタイプを定義する．一方，作家の場合は画像検索の結果は必要ではなくかわりにサイン会などの情報を提示できればよいと考え，「芸能人」とは別に「作家」というクエリタイプを定義する．

(2)　クエリタイプ付与時

(a)　評価者に提示した情報

クエリタイプおよび各クエリ

(b)　評価者に指示した内容

クエリタイプの名称のみをもとにクエリタイプを付与する．クエリタイプの属するクエリを例示することは行わない．

クエリを見ただけではクエリタイプが判断できない場合は，深く調べることはせず，不明とする．

クエリタイプの定義を行った結果，**表 1** に示す 27 のクエリタイプが定義された．このクエリタイプをもとにクエリタイプの付与を行い，5 人の評価者のうち 3 人以上が不明を除く同一のクエリタイプを付与したものを正解としたところ，件数は約半数の 14,858 件となった．27 のクエリタイプのうち，クエリ数が最も多いクエリタイプは「芸能人」となり，その次にクエリ数が多いクエリタイプは「企業名」となった．定義したクエリタイプがどの程度のクエリを網羅しているかを調べるのは，該当するクエリタイプが存在せずに不明としている場合と単純に判断できずに不明としている場合の区別ができないため難しいが，5 人の評価者が不明としたクエリの数を調べることである程度推定できると考えられる．そこでそのようなクエリの数を調査したところ，その数は少なく，663 件であった．よって，これらのクエリタイプでおおよそのクエリをカバーできており，よく検索されるクエリに対してシステムの応答を変化させるという用途においては実用的なクエリタイプであると考えられる．14,858 件のうち 9 割を学習データとして利用し，残りの 1 割を評価データとして利用した．

3.3.2　実 験 方 法

評価データに対して，クエリタイプの判定を行った．

評価指標として，以下の 3 つを用いた．

(1)　カバー率

クエリのうち正解かどうかを問わず何らかのクエリタイプが付与されたクエリの割合を表す．クエリに関する文書が取得できない場合などにはクエリタイプを付与することができないため，どの程度のクエリに対して手法が適用できるかを確認するためにこの指標を利用した．

(2)　適合率

何らかのクエリタイプが付与されたクエリのうち，正解のクエリタイプが付与されたクエリの割合を表す．

(3)　正解率

カバー率と適合率の積であり，全体としてクエリのうち正解のクエリタイプが付与さ

表 2　学習データを用いたクエリタイプ判定結果
Table 2　Results of example-based method.

手法	カバー率	適合率	正解率
クエリ中の単語利用	0.231	0.926	0.214
Wikipedia 主要語利用	0.602	0.604	0.364
学習データ利用	0.771	0.723	0.557

れたクエリの割合を表す．

手法として，上記で述べた学習データを用いた手法を組み合わせた手法を用意した．具体的には，まず学習データ中のクエリに対して形態素解析を行い，名詞および連続する未知語を単語として抽出した．「動物占い」のように名詞が連続する場合は，まとめて 1 つの単語とはせず，「動物」「占い」をそれぞれ別の単語として扱った．クエリタイプごとに単語の頻度をカウントし，頻度が閾値以上の単語を頻出語として抽出した．評価データ中のクエリに対して形態素解析を行って名詞および連続する未知語を単語として抽出し，単語の中に頻出語が含まれるかどうかを調べ，含まれていればそれに対応するクエリタイプを付与した．閾値の決定にあたっては，様々な値を閾値として頻出語を抽出し，学習データを用いて評価を行い，適合率が最も高くなるような値を選択した．正解率を用いた場合，学習データを用いて頻出語を抽出し同じ学習データを用いて評価を行うと，閾値を低くするほどカバーできる範囲が増えて正解率が高くなるということになり，閾値の決定のための指標とはならない．一方，適合率を用いた場合，閾値を低くしていくと，カバーできるクエリ数は増加するが，一般的な語を頻出語として抽出するためクエリタイプの判定を誤るクエリ数も増加し，適合率は下がっていく．また，閾値を高くしていくと，クエリタイプの推定を誤りにくい頻出語が残るため適合率は上がる傾向にあるが，カバーできるクエリ数が減少して適合率が下がる場合もある．そのため，適合率が高くなるような値が閾値として適切であると考えた．このようにして閾値を決定した結果，閾値は 20 となった．なお，クエリタイプの異なる頻出語が複数個含まれている場合は，後方に出現する頻出語を優先した．特定の頻出語が含まれていなかった場合には，クエリをタイトルに持つ Wikipedia 記事のリード文に含まれる主要語を抽出し，その主要語が与えられた場合に各クエリタイプに属する確率が最大となるようなクエリタイプを選択した．

3.3.3　実 験 結 果

実験結果を**表 2** に示す．結果より，半数強のクエリに対しては正しくクエリタイプを付与できたが，残りの半数弱のクエリに対しては正しくクエリタイプを付与できなかったことが分かる．ここで問題となるのは，カバー率の低さであると考えられる．結果より，2 割以上のクエリに対して何らかのクエリタイプを付与することに失敗していることが分かる．

表 3　クエリタイプ判定に失敗したクエリの例
Table 3　Failure examples.

チェジュウ	ブラットピット	ほっかほか亭	脳ないメーカー	ミクシィー

クエリタイプ判定に失敗したクエリの例を**表 3** に示す．クエリ中の単語を用いる手法では，クエリタイプを付与できるクエリは頻出語を含むものに限られるため，多くのクエリに対してクエリタイプを付与することができないという問題がある．また，Wikipedia の情報を利用した手法では，クエリが Wikipedia 記事のタイトルとなっている場合のみしかクエリタイプを判定できないという問題がある．現在，日本語版 Wikipedia には 65 万を超える豊富な記事が存在するが，Web 検索のクエリとしてよく出現するニックネームや表記の誤りを含むクエリは Wikipedia のタイトルとして存在しないため，この手法では対応できない．

4. 提 案 手 法

提案手法では，概念ベースおよび各クエリタイプを特徴付ける数個の特徴語のみを利用して，学習データを必要とすることなくクエリタイプを判定する．以下では，概念ベースについて述べた後，概念ベースを利用したクエリタイプ判定手法について説明する．

4.1　概念ベース

概念ベースは，様々な単語に対してその単語の概念を表す概念ベクトルが付与されたデータベースである[26]．以下では，概念ベースの生成方法および性質について述べる．

4.1.1　概念ベースの生成方法

概念ベースの生成方法はいくつか存在するが，基本となるのは単語間共起に基づく手法である．単語間共起に基づく手法では，まずコーパス中の各文を形態素解析し，得られた形態素列の中から内容語を取り出す．内容語のうちの 1 つを取り出して概念語とし，残りの内容語を共起語として，概念語と共起語が文中で共起する頻度を算出することにより，概念語を行，共起語を列とする共起行列を生成する．共起行列の各行の行ベクトルを共起ベクトルと呼ぶ．共起ベクトルが類似している概念語どうしは共起のパターンが類似しているため，類似した概念を持つと考えられるため，単語間の類似性を測るのに利用できる．しかしながら，共起ベクトルは次元数が大きく，データスパースネスの問題も存在するため，概念語お

単語	1	2	3	···	D
学校	-0.052	0.045	0.040	···	0.074
総理	-0.011	0.010	0.017	···	0.013
首相	-0.015	0.012	0.022	···	0.015
···	···	···	···	···	···

図 2　概念ベースの例
Fig. 2　Example of concept base.

よび共起語を高頻度の語に限定し，さらに特異値分解を行って次元を圧縮する．この次元を圧縮して得られた行列の各行の行ベクトルが概念ベクトルとなる．

単語間共起に基づく手法では，計算量削減のために，特異値分解を行う際に共起語を高頻度の語に限定するため，低頻度の語が考慮されないという問題点がある．また，共起語の中には「りんご」と「みかん」のように同じカテゴリに属する共起語との頻度を別々に算出しているため，ある概念語はコーパス中で「りんご」とよく共起し，別の概念語はコーパス中で「みかん」とよく共起するような場合にこれらの概念ベクトルが類似したものにならない場合があるという問題がある．これらの問題を解消するために，単語・意味属性間共起に基づく手法では，概念語と単語の共起頻度ではなく，概念語と単語の意味属性との共起頻度を算出する．これにより，単語間共起に基づく手法において別々に算出されていた「りんご」および「みかん」との共起頻度は，「果物」という意味属性との共起頻度としてまとめられるため，上記の問題を解消することができる．

汎用的で質の高い概念ベースを生成するためには，様々な分野について書かれた質の高い文書を利用して生成を行う必要がある．ブログ記事には様々な分野の記事が存在するが，記述のミスなどを多く含むため，質の高い文書とはいえない．また，新聞記事は質の高い文書であるが，分野に偏りがあると考えられる．一方，Q&A 文書は様々な分野について記述されており，また質問者や回答者に読んでもらって理解してもらう必要があるため，比較的質の高い文書であると考えられる．そのため，Q&A 文書を利用して概念ベースを生成するのが望ましいと考えられる．

4.1.2　概念ベースの性質

生成された概念ベースの例を**図 2** に示す．この例における「総理」と「首相」のように，類似する概念を持つ単語の概念ベクトルは類似するため，ベクトル間の距離が近くなるという性質を持つ．概念ベクトルの各次元は何らかの分野を表しており，概念ベクトルはそれぞれの分野に対して相関を持つかどうかを表していると考えられるため，概念ベクトルはその単語がどのような分野において用いられるかを表現するものであるととらえることができる．提案手法では，クエリや文書などの分野を概念ベクトルを用いて表現し，クエリと各クエリタイプの分野の近さをもとにクエリタイプを判定する．

4.2　概念ベースを用いたクエリタイプ判定

提案手法では，単語の概念ベクトルをもとに，クエリに関する分野を表すクエリ分野ベクトルとクエリタイプごとのクエリタイプ分野ベクトルの距離を求めることによってクエリタイプスコアを求める．概念ベースは単語に対して概念ベクトルが付与されたデータベースであるため，クエリ中の単語から直接概念ベクトルを取得してクエリ分野ベクトルを生成することも場合によっては可能である．しかし，クエリが新語などであった場合にはクエリ中の単語から概念ベクトルを取得できない場合もある．クエリ中の単語が概念ベクトルを持つもので構成されているかどうかを調査したところ，全体の 61.6%のクエリについてはクエリ中の単語がすべて概念ベクトルを持っていたが，残りのクエリについては概念ベクトルを持たない単語を含んでいた．調査に用いた概念ベースは 4.1.1 項で述べた理由により Q&A 文書から生成されているが，Q&A 文書では質問や回答の内容を理解してもらうために誤字などは極力修正されて投稿されると想定される．一方，Web 検索のクエリでは，たとえば人名の漢字が分からず平仮名で入力されるというようなことは頻繁に起こりうる．これが，クエリ中に概念ベクトルを持たない単語が多く存在した原因であると考えられる．そこで，クエリ中の単語から直接概念ベクトルを取得することはせず，以下の手順によりクエリタイプを判定する．

(1)　各クエリタイプの分野を表すクエリタイプ分野ベクトルを生成（4.2.1 項）
(2)　クエリの分野を表すクエリ分野ベクトルを生成
　(a)　クエリタイプ判定の対象となるクエリに関連する文書を取得（4.2.2 項）
　(b)　文書中の単語からクエリ分野ベクトルを生成（4.2.3 項）
(3)　クエリ分野ベクトルと各クエリタイプ分野ベクトルとの類似度を算出し，類似度をもとにクエリタイプを判定（4.2.4 項）

このようにして関連する文書を取得して文脈を用いることにより，出現頻度の低いクエリであっても，文脈からクエリの分野を正しく推定することが可能となる．

4.2.1　クエリタイプ分野ベクトルの生成

前処理として，各クエリタイプを特徴付ける単語（以下では特徴語と呼ぶ）を数個用意し，それらの単語をもとにクエリタイプ分野ベクトルを生成する．特徴語は人手で用意することを想定している．たとえば，「芸能人」というクエリタイプに対する特徴語として「俳

優」「芸人」「ミュージシャン」を用意し,「グルメ」というクエリタイプに対する特徴語として「レストラン」「レシピ」を用意する．特徴語の選択基準としては，上記の例のようにクエリタイプの表す分野がいくつかの詳細な分野に分かれるような場合は主にそれらの詳細な分野を表す語を特徴語とし，そうでない場合は主にクエリタイプのラベルとして用いられている 1 語のみを特徴語として用いた．用意した特徴語の概念ベクトルの重心をクエリタイプ分野ベクトルとする．

$$\boldsymbol{t}_k = \frac{1}{|T_k|} \sum_{w \in T_k} \boldsymbol{c}_w$$

ここで，$\boldsymbol{t}_k$ は k 番目のクエリタイプのクエリタイプ分野ベクトル，T_k は k 番目のクエリタイプを特徴付ける単語の集合，$\boldsymbol{c}_w$ は単語 w の概念ベクトルを表す．

4.2.2　クエリに関連する文書の取得

入力されたクエリに対し，クエリの分野を表すために必要な文書を取得する．ここで取得されるべき文書は，単にクエリを含んでいる文書ではなく，クエリの詳細な説明などが記述されたクエリに関連する文書である．このようなクエリに関連する文書を取得するための手法として，文書検索の分野では BM25 などの様々な適合性スコアが提案され，このスコアに基づいて文書の検索を行う文書検索エンジンが利用可能である．よって，ここでは文書検索エンジンを利用し，クエリを入力として上位の検索結果の概要文を文書として取得する．Web 上には様々な情報源が存在し，それぞれ取得できる文書が異なるため，クエリタイプ判定の結果も異なると考えられる．そこで，様々な情報源からそれぞれの文書を取得する．情報源としては 3 種類のものを利用する．クエリが一般にどのように利用されているかを判断するための第 1 の情報源として，ブログを利用する．クエリが語義としてどのように利用されているかを判断するための第 2 の情報源として，Wikipedia を利用する．クエリタイプを判定しようとするクエリが入力された検索エンジンから得られる第 3 の情報源として，携帯サイトを利用する．

4.2.3　クエリ分野ベクトルの生成

得られた複数の文書をもとに，クエリに関する分野を表すクエリ分野ベクトルを生成する．ここでは，クエリに関連する各文書に含まれる単語の概念ベクトルの重心を文書分野ベクトルとし，複数の文書分野ベクトルの重心をクエリ分野ベクトルとする．

$$\boldsymbol{q} = \frac{1}{M} \sum_{i}^{M} \boldsymbol{d}_i$$

$$\boldsymbol{d}_i = \frac{1}{|D_i|} \sum_{w \in D_i} \boldsymbol{c}_w$$

ここで，$\boldsymbol{q}$ はクエリ分野ベクトル，M は取得した文書数，$\boldsymbol{d}_i$ は i 番目の文書の文書分野ベクトル，D_i は i 番目の文書に含まれる単語の集合を表す．

4.2.4　クエリタイプの判定

得られたクエリ分野ベクトルと各クエリタイプごとのクエリタイプ分野ベクトルとの距離をクエリタイプスコアとして算出する．ベクトル間の距離にはコサイン距離を用いる．

$$S_k = \frac{\boldsymbol{q} \cdot \boldsymbol{t}_k}{|\boldsymbol{q}||\boldsymbol{t}_k|}$$

ここで，S_k は k 番目のクエリタイプのクエリタイプスコアを表す．クエリタイプスコアの最も高いクエリタイプをクエリに最もマッチするクエリタイプと判定する．

4.2.5　複数の判定結果の統合

これまでに述べてきたクエリタイプ判定方法により，3 種類の情報源を用いて 3 種類の判定結果が得られる．これらの判定結果を統合して最終的な判定結果を得るため，それぞれの手法に重みを設定し，クエリタイプごとのクエリタイプスコアの線形和を最終的なクエリタイプスコアとして算出する．

5. 実　　験

提案手法の有効性を検証するために，実験を行った．実験環境については 3.3.1 項，実験方法については 3.3.2 項と同様である．クエリに関連する文書の取得では，多くの検索サービスでは検索結果が 10 件ずつ提示され，検索エンジンは 1 回の検索で少なくとも 10 件の検索結果を取得できると考えられるため，上位 10 件の検索結果を用いた．適合性スコアとしては，情報検索において広く用いられている TFIDF を用いた．検索結果の概要文の生成方法としては，100 文字のウインドウ幅を持つウインドウを文書の先頭からスライドさせていき，最も多くのクエリを含む箇所を概要文とする手法を採用した．ブログ記事には 2009 年 4 月時点で収集可能な数千万件の文書を利用した．携帯サイト記事には 2007 年 11 月時点で収集可能な数億件の文書を利用した．概念ベースの作成には，Web ポータルサイトで提供されている Q&A サービスから取得可能な約 300 万件の Q&A 文書を利用した．複数の判定結果の統合では，ブログに対する重みを 0.5，Wikipedia に対する重みを 0.25，携帯サイトに対する重みを 0.25 とした．3 章で述べた手法をベースライン手法として提案手法と比較を行った．

表 4　ベースライン手法との比較結果
Table 4　Comparison with baseline method.

手法	カバー率	適合率	正解率
ベースライン手法	0.771	0.723	0.557
提案手法	0.991	0.651	0.646
提案手法＋ベースライン手法	0.995	0.775	0.771

表 5　個々の情報源によるクエリタイプ判定結果
Table 5　Quality of each information resource.

情報源	カバー率	適合率	正解率
ブログ記事	0.989	0.529	0.524
携帯サイト	0.776	0.495	0.384
Wikipedia	0.772	0.696	0.538
すべて	0.991	0.651	0.646

5.1 実験結果

以下では，ベースライン手法との比較結果，クエリタイプ分野ベクトルを正解データから生成した場合との比較結果，個々の手法によるクエリタイプ判定結果について述べる．

5.1.1 ベースライン手法との比較結果

はじめに，ベースライン手法との比較を行った．その結果を**表 4** に示す．

実験結果より，提案手法はベースライン手法よりも正しくクエリタイプを判定できることが示された．提案手法では全体的にカバー率が高くなっていることが分かる．クエリは基本的に検索システムの利用者が入力するものであるため，クエリの中には誤字を含むものや愛称などで入力されるものも少なくない．しかし，ブログ記事などはクエリと同様に校正されることなく投稿されることが多いため，同様の誤字などの表現を含んでいる．これがカバー率の高くなった原因であると考えられる．一方，ベースライン手法では提案手法に比べてカバー率が低く，適合率が高くなっていることが分かる．カバー率が低くなった原因は，誤字を含むような場合に Wikipedia の記事を検索できないためであると考えられる．適合率が高くなった原因は，より確実性の高い情報を用いているためであると考えられる．クエリ中の単語を利用した手法は，たとえば「占い」を含むクエリはクエリタイプも「占い」であるというあまり例外のない情報を利用している．また，Wikipedia の主要語を利用した手法でも，たとえば主要語として「アナウンサー」が取得できた場合にはほぼ確実にクエリタイプも「アナウンサー」であるといえる．

カバー率の高い提案手法と適合率の高いベースライン手法は，組み合わせることにより精度が高まると考えられる．そこで，2 つの手法の結合を行った．各クエリタイプ判定方法により得られた判定結果を統合して最終的な判定結果を得るため，それぞれの手法に重みを設定し，クエリタイプごとのクエリタイプスコアの線形和を最終的なクエリタイプスコアとして算出した．それぞれの手法に対する重みを設定するため，ここでは学習データに含まれるクエリに対してできるだけ正しくクエリタイプが判定できるように遺伝的アルゴリズムを用いて重みを変更しながら最適値を求めた．その結果，提案手法を単独で用いた場合に比べ，10 ポイント以上高い正解率 77.1%を実現できた．

5.1.2 個々の情報源によるクエリタイプ判定結果

提案手法によってクエリタイプ判定を行い，結果を統合する前の個々の情報源による手法がどの程度有効に働いているかの実験を行った．その結果を**表 5** に示す．

実験結果より，個々の手法では良いものでも 5 割強のクエリに対してのみしか正しくクエリタイプが判定できないが，個々の判定結果を統合することにより多くのクエリに対して正しくクエリタイプを判定できることが示された．

情報源ごとに比較を行うと，携帯サイトを情報源として利用した場合の正解率が低くなった．これは，利用した携帯サイト記事が少し古いものであったため，最近登場した芸能人などのクエリに対応できなかったのではないかと考えられる．

5.2 考　　察

ここでは，判定結果を実際に目視により確認して得られた知見について述べる．まず，提案手法によりクエリタイプ判定を行った結果の一例を**表 6** に示す．

この例ではシステムが 1 位に出力したクエリタイプが正解と一致した例をあげているが，システムが 2 位以下に出力したクエリタイプもある程度クエリから連想されるクエリタイプであるものが多いことが分かる．提案手法が有効に働いた結果であると考えられる．この結果をシステムの応答を変化させるために利用することを考える．一例として，1 位のクエリタイプスコアの 95%以上のスコアを持つクエリタイプを選択し，クエリタイプスコアの順にクエリタイプに対応する検索結果を提示することができる．クエリ「オードリー・ヘプバーン」の場合には，クエリタイプとして「芸能人」が選択され，「芸能人」に対応する検索結果として画像検索の結果を提示することができる．また，クエリ「ガンダム」の場合には，クエリタイプとして「番組名」「ゲーム」が選択され，検索結果として最上部に番組に関する情報を提示し，その下部にゲームの攻略に関する情報を提示することができる．このように，システムの応答を適切に変化させることができるため，得られたクエリタイプ判定

表 6　クエリタイプ判定結果の例
Table 6　Examples of query type assignment.

クエリ	正解クエリタイプ	システム出力結果（括弧内はスコア）		
		1 位	2 位	3 位
オードリー・ヘプバーン	芸能人	芸能人（0.512）	映画（0.392）	番組名（0.389）
ガンダム	番組名	番組名（0.434）	ゲーム（0.431）	映画（0.416）
花札	ゲーム	ゲーム（0.517）	キャラクター（0.408）	書籍（0.398）
スクリーンセーバー	PC 関連	PC 関連（0.424）	ゲーム（0.351）	サイト名（0.342）
デジタルカメラ	家電	家電（0.424）	PC 関連（0.334）	画像（0.316）
ランドマークタワー	地名	地名（0.473）	企業名（0.431）	イベント名（0.373）
ホームページ無料	サイト名	サイト名（0.528）	PC 関連（0.488）	曲名（0.437）
卒業式	イベント名	イベント名（0.374）	曲名（0.345）	書籍（0.317）
ダイエット方法	健康/医療	健康/医療（0.464）	グルメ（0.453）	イベント名（0.373）
うなぎ	グルメ	グルメ（0.481）	地名（0.409）	健康/医療（0.398）

表 7「番組名」および「映画」のスコアが高くなったクエリの例
Table 7　Examples of queries whose scores of “TV Program” and “Movie” got high.

$SCORE_{TV} > SCORE_{MOVIE}$	$SCORE_{MOVIE} > SCORE_{TV}$
ルパン	リロ&スティッチ
ヤッターマン	スーパーマン
ボウケンジャー	
踊る大捜査線	

表 8　正解数と正解出力数
Table 8　Number of appropriate query type and correct query type.

		正解出力数		
		1	2	3
正	1	72	—	—
解	2	5	20	—
数	3	0	1	2

結果が有効であるということが確認できた．

得られたクエリタイプのスコアが人間の直観と一致しているかどうかについて検証する．「ガンダム」の例では，1 位と 2 位のクエリタイプのスコアの値の差は 2 位と 3 位のクエリタイプのスコアの差に比べて小さいことから，1 位の「番組名」と 2 位の「ゲーム」が「ガンダム」に対するクエリタイプであると見なせる．ここで得られたクエリタイプおよびその順位は人間の直観とほぼ一致していると考えられる．もう少し詳細な検証を行うため，クエリタイプ「番組名」および「映画」のスコアが高くなったクエリを抽出した．その結果得られたクエリの例を**表 7** に示す．

表 7 において，左側のクエリはクエリタイプ「映画」のスコアよりも「番組名」のスコアのほうが高かったものであり，右側のクエリはクエリタイプ「番組名」のスコアよりも「映画」のスコアのほうが高かったものである．これを見ると，左側のクエリは一般的にテレビ番組として認知されており，テレビ番組として人気が出たため後に映画化されたものが集まっている．逆に右側のクエリは一般的に映画として認知されており，映画として人気が出たために後にテレビ番組化されたものが集まっている．このことから，得られたクエリタイプのスコアが人間の直観とおおよそ一致することが確認できた．

次に，提案手法により適切に複数のクエリタイプを判定できるかどうかを検証するための分析を行った．ここでは正解のクエリタイプが複数存在すると考え，システムが 1 位に出力したクエリタイプが正解と一致した例の中からランダムに 100 件を抽出し，正解とすべきクエリタイプの数および正解のクエリタイプが提案手法により上位に連続して出力される数について調査を行った．その結果を**表 8** に示す．

結果より，クエリの 1/4 以上は複数のクエリタイプが付与されるべきであることが分かる．また，複数のクエリタイプを付与すべきクエリの約 8 割に対して正しく複数のクエリタイプを付与できることが分かる．複数のクエリタイプが考えられる場合でも提案手法により複数のクエリタイプを正しく付与できることが確認できた．

表 9　誤りの原因
Table 9　Cause of error.

原因	件数
「番組名」を「書籍」と判定	11
「企業名」を「車名」と判定	9
「アダルト」を判定不可	8
「サイト名」を「書籍」と判定	7
「企業名」を「地名」と判定	6
「サイト名」を「俗語」と判定	6
「サイト名」を「PC 関連」と判定	6
「健康/医療」を「俗語」と判定	5
「企業名」を「スポーツ」と判定	5

次に，システムが 1 位に出力したクエリタイプが正解と一致しなかった 340 件の例を集計し，誤りの原因を分析した．その結果のうち，件数が 5 件以上となった原因を**表 9** に示す．

この結果から，以下の 4 つが主な誤りの原因として考えられる．

複数のクエリタイプを持つクエリの存在

「番組名」を「書籍」に判定した誤りは最も多かったが，実際のクエリを見てみると漫画がテレビアニメ化されたものやテレビ番組が書籍化されたもので占められていた．このようなクエリタイプを正しく判定するのは難しい問題であるが，2 つのクエリタイプのうちどちらが代表的なものであるかを判定する手法について検討する必要がある．

一部に異なるクエリタイプを持つ語句が存在するクエリの存在

「企業名」を「地名」に判定した誤りでは，企業名の文字列の一部に地名が含まれているものが多かった．また，「企業名」を「スポーツ」に判定した誤りにおいても，「日本サッカー協会」のように文字列の一部にスポーツを表す語が含まれていた．このようなクエリを正しく判定するためには，クエリに含まれる長単位の文字列を優先するなどの手法について検討していく必要がある．

重複した概念を持つクエリタイプの存在

「サイト名」を「PC 関連」に判定した誤りでは，サイトが提供されているのはインターネット上であるため，PC に関連すると考えられる．そのため，この 2 つのクエリタイプは重複した概念を持つと考えられる．また，「企業名」を「車名」に判定する誤りでは，自動車を製造する企業を表すクエリが含まれていた．このようにクエリによっては概念が重複してしまうケースも存在する．このケースはベースラインにおける誤りとしても多く含まれており，提案手法を導入してもうまく解決できなかったものであることを示している．このようなケースに対しては，固有表現抽出の結果得られる固有表現タイプを用いるなどの手法について検討していく必要がある．

分野ベクトルの作成が難しいクエリタイプの存在

「俗語」というクエリタイプは，クエリタイプ分野ベクトルの生成に必要な単語の与え方が難しく，クエリタイプ分野ベクトルが正しく分野を表していなかったことが考えられる．このようなクエリタイプに対して適用可能な手法について検討する必要がある．

6. ま と め

本論文では，概念ベースを参照してクエリに関する文書から得られたクエリ分野ベクトルと各クエリタイプ分野ベクトルとのコサイン距離に基づきクエリタイプを判定する手法を提案した．実験では，27 のクエリタイプに対し，提案手法単独では 64.6%，Wikipedia の情報などを利用した手法と組み合わせることにより 77.1%の精度で判定を行うことができた．

実験の結果，クエリが複数のクエリタイプを持っていたり，クエリタイプの持つ概念に重複があるような場合に誤りが起きやすいことが分かった．今後は，このような区別が難しいクエリタイプに対して別途 2 値分類の手法を適用するなどの手法などについて検討していきたい．

参 考 文 献

1) Murata, M., Toda, H., Matsuura, Y. and Kataoka, R.: Query-Page Intention Matching using Clicked Titles and Snippets to Boost Search Rankings, *Proc. 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp.105–114 (2009).
2) Baeza-Yates, R.: Applications of Web Query Mining, *Proc. 27th ECIR*, pp.7–22 (2005).
3) Cui, H., Wen, J., Nie, J. and Ma, W.: Probabilistic Query Expansion using Query Logs, *Proc. 11th International Conference on World Wide Web*, pp.325–332 (2002).
4) Broder, A.: A taxonomy of web search, *SIGIR Forum*, Vol.3, pp.3–10 (2002).
5) Rose, D.E. and Levinson, D.: Understanding user goals in web search, *Proc. 13th International Conference on World Wide Web*, pp.13–19 (2004).
6) Lee, U., Liu, Z. and Cho, J.: Automatic identification of user goals in web search, *Proc. 14th International Conference on World Wide Web*, pp.391–400 (2005).
7) König, A.C., Gamon, M. and Wu, Q.: Click-Through Prediction for News Queries, *Proc. 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.347–354 (2009).
8) Kang, I. and Kim, G.: Query Type Classification for Web Document Retrieval,

Proc. 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp.64–71 (2003).
9) Fujii, A.: Modeling Anchor Text and Classifying Queries to Enhance Web Document Retrieval, *Proc. 17th International Conference on World Wide Web*, pp.337–346 (2008).
10) Vallet, D. and Zaragoza, H.: Inferring the Most Important Types of a Query: A Semantic Approach, *Proc. 31th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.857–858 (2008).
11) 隅田飛鳥，吉永直樹，鳥澤健太郎：Wikipedia の記事構造からの上位下位関係抽出，自然言語処理，Vol.16, No.3, pp.3–24 (2009).
12) Paşca, M.: Acquisition of Categorized Named Entities for Web Search, *Proc. 13th ACM International Conference on Information and Knowledge Management*, pp.137–145 (2004).
13) Shinzato, K. and Torisawa, K.: Acquiring Hyponymy Relations from Web Documents, *Proc. HLT-NAACL 2004*, pp.73–80 (2004).
14) Shinzato, K. and Torisawa, K.: Extracting Hyponyms of Prespecified Hypernyms from Itemizations and Headings in Web Documents, *Proc. COLING 2004* (2004).
15) Pantel, P. and Pennacchiotti, M.: Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations, *Proc. 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL*, pp.113–120 (2006).
16) 小町　守，鈴木久美：検索ログからの半教師あり意味知識獲得の改善，人工知能学会論文誌，Vol.23, No.3, pp.217–225 (2008).
17) Grishman, R. and Sundheim, B.: Message understanding conference - 6: A brief history, *Proc. COLING-96* (1996).
18) Sekine, S. and Isahara, H.: IREX: IR and IE evaluation project in Japanese, *Proc. 2nd International Conference on Language Resources and Evaluation* (2000).
19) Sekine, S., Sudo, K. and Nobata, C.: Extended named entity hierarchy, *Proc. LREC2002* (2002).
20) 関根　聡，鈴木久美：検索ログによる拡張固有表現辞書の整備，言語処理学会第 13 回年次大会 (2007).
21) 新納浩幸，関根　聡：拡張固有表現タガーの作成とその問題点の考察，言語処理学会第 12 回年次大会発表論文集 (2006).
22) 橋本泰一，乾　孝司，村上浩司：拡張固有表現タグ付きコーパスの構築，情報処理学会研究報告，Vol.2008-NL, No.113, pp.113–120 (2008).
23) Guo, J., Xu, G., Cheng, X. and Li, H.: Named Entity Recognition in Query, *Proc. 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.267–274 (2009).
24) Beitzel, S.M., Jensen, E.C., Frieder, O., Grossman, D., Lewis, D.D., Chowdhury, A. and Kolcz, A.: Automatic web query classification using labeled and unlabeled training data, *Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.581–582 (2005).
25) Broder, A., Fontoura, M., Gabrilovich, E., Joshi, A., Josifovski, V. and Zhang, T.: Robust Classification of Rare Queries Using Web Knowledge, *Proc. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.231–238 (2007).
26) 別所克人，内山俊郎，内山　匡，片岡良治，奥　雅博：単語・意味属性間共起に基づくコーパス概念ベースの生成方式，情報処理学会論文誌，Vol.49, No.12, pp.3997–4006 (2008).

(平成 22 年 3 月 19 日受付)
(平成 22 年 7 月 6 日採録)

(担当編集委員　藤井 敦)

廣嶋　伸章（正会員）

2000 年慶應義塾大学大学院理工学研究科修士課程修了．同年日本電信電話（株）入社．現在，NTT サイバーソリューション研究所勤務．入社以来，文書要約，評判情報抽出，情報検索の研究に従事．言語処理学会会員．

戸田　浩之（正会員）

1999 年名古屋大学大学院工学研究科博士課程前期課程修了．同年日本電信電話（株）入社．2007 年筑波大学大学院システム情報工学研究科博士後期課程修了．現在，NTT コミュニケーションズ（株）勤務．入社以来，情報検索，情報抽出，Web マイニングの研究に従事．博士（工学）．ACM，電子情報通信学会，日本データベース学会，人工知能学会各会員．

松浦由美子（正会員）

1993 年慶應義塾大学大学院理工学研究科修士課程修了．同年日本電信電話（株）入社．以来，音楽データの特徴部分抽出の研究，電子透かしシステムの研究開発，コンテンツ流通システムの研究開発，ポータルサービスシステムの研究開発に従事．現在，NTT サービスインテグレーション基盤研究所所属．

片岡　良治（正会員）

1987 年千葉大学大学院電子工学専攻修士課程修了．同年日本電信電話（株）入社．以来，トランザクションの並行処理制御方式の研究，マルチメディア情報システムの研究，ポータルサービスシステムの研究に従事．現在，NTT サイバーソリューション研究所所属．