

情報要求の言語化を支援する クエリ拡張型 Web 検索システムに関する一検討

大塚 淳史^{†1} 関 洋平^{†1}
神門 典子^{†2} 佐藤 哲司^{†1}

コミュニティQA サイトでは、ユーザの疑問や知りたいことが自然言語で記述された質問記事として投稿されている。本論文では、Web 検索で必須な言語化された検索クエリの想起を、質問記事を提示することで支援する、クエリ拡張型 Web 検索システムを提案する。提案システムでは、検索者から入力されたキーワードと関連する質問記事を複数のカテゴリから抽出し、拡張するキーワードとともに提示するクエリ拡張を実現する。検索者は、提示された質問記事を閲覧することで、自身の情報要求を拡張の根拠として言語化された検索クエリとして確認することができる。潜在的意味インデキシングを用いてシステムを実装し、実運用されたコミュニティQA サイトの質問記事を用いて評価を行い、提案システムによって多様なクエリ拡張と、根拠となる質問記事を提示できることを確認した。

Diversified-query Generating System Using Community QA Resources to Verbalize Latent Information Needs

ATSUSHI OTSUKA,^{†1} YOHEI SEKI,^{†1} NORIKO KANDO^{†2}
and TETSUJI SATOH^{†1}

In community QA websites, huge number of questions written in natural language are posted by users. We propose a new query expansion system to support Web search users by using community QA resources. In our system, users can get diversified query candidates from questions in multiple Yahoo! chiebukuro (Yahoo! Answers) categories relevant to the search query, and will find queries relevant to their information needs by browsing question articles. We also report experimental results to extract diversified question articles with latent semantic indexing.

1. はじめに

膨大な Web ページから所望のページを探し出す Web 検索エンジンでは、1 語以上のキーワード組からなるクエリを入力する。そのため、検索ユーザは、自らの知りたいことである情報要求を、いくつかのキーワードの組として言語化しなければならない。ここで情報要求とは、ユーザの思考の中で、知りたいことが的確に言語化されている状態とする。しかし、ユーザの情報要求が言語化できないほど曖昧な場合、思いついた 1 語のクエリで検索を開始することも少なくない。たとえば、マスメディアで紹介されたキーワードについて検索を行う場合などは、検索のゴールが定まっておらず、情報要求の言語化が困難になる。また、検索エンジンが提供する急上昇ワードランキングも、ユーザにとって未知なキーワードが多く、現在の Web 検索では、情報要求が言語化されないままに検索を行わざるをえない状況が存在している。

膨大な Web ページ集合に対して、1 語程度のキーワードで検索を行っても、大量の類似したページが検索され、ユーザの情報要求を満足させる検索結果を得ることは難しい。このため、キーワードの追加や変更などの操作を行い、クエリを修正する必要があるが、この修正の際にも情報要求を言語化しなければならない。検索結果を閲覧することで次の検索のヒントを得ることができるが、ユーザにとって未知のキーワードで検索した結果から、適切なキーワードを追加することは難しいといえる。

このような情報要求の言語化にかかる、ユーザの負担を軽減する手法として、入力されたキーワードからクエリを補完するクエリ拡張がさかに行われている。商用の検索エンジンでも積極的に利用されている手法であるが、提示されたキーワードを知らないユーザにとっては、適切なキーワードを選択することは難しい。提示された拡張クエリがユーザの情報要求を満足させるものであったとしても、ユーザがそのことに気付かなければ、有効な拡張クエリとはならない。

本論文では、拡張の根拠を理解できる形式にして提示するクエリ拡張法を提案する。ユーザは、提示された拡張の根拠と自身の中にある情報要求を照らし合わせ、追加・修正されたキーワードが適切な拡張であるかの判断を支援する。提案システムでは、コミュニティQA

^{†1} 筑波大学大学院図書館情報メディア研究科

Graduate School of Library, Information and Media Studies, University of Tsukuba

^{†2} 国立情報学研究所

National Institute of Informatics

の質問記事を言語化された情報要求と見なし、検索エンジン用のキーワード組と対で提示する。質問記事は、コミュニティQAの質問者が知りたいことや疑問を自然言語で記述したものであり、質問を投稿する際に質問者が行った言語化の過程を、検索エンジンへのクエリ入力に再現する。クエリ拡張の根拠を質問記事として提示することで、検索ユーザは自身の情報要求を的確に表現している拡張クエリを選択できると考える。

提案手法によるクエリ拡張の有効性を評価することを目的として、作成された拡張クエリにより検索された Web ページからキーフレーズ抽出を行い、抽出できたキーフレーズ数を比較する評価実験を行った。いくつかの商用の検索エンジンで実装されている拡張クエリと比較して、提案手法によって拡張されるクエリの多様さをキーフレーズ数で比較した。

本論文の構成は以下のとおりである、まず2章で、先行する関連研究について述べる。3章で、本論文が提案する拡張クエリについて説明し、4章でその実装について述べる。5章では、提案手法の評価として、出力結果の例とキーフレーズ抽出実験について説明する。6章で考察を行い、7章でまとめと今後の課題について述べる。

2. 関連研究

本研究は、クエリ拡張に際してコミュニティQAの情報を利用することに特徴があり、クエリを拡張する際の根拠とした質問記事を合わせて提示することで、検索ユーザによる情報要求の言語化を支援する。情報検索でのクエリ拡張は、適合性フィードバックやシソーラスの応用技術とされる¹⁾。Yin ら²⁾は、Web 検索エンジンで検索した Web ページのスニペットから拡張クエリを作成することが効果的であることを示している。村田ら³⁾は、多くの検索ユーザからアクセスされる Web ページのスニペットから拡張クエリを作成することで、追加する語が少ない場合の検索精度を向上させている。Web 検索結果の一覧から必要な Web ページを探す際、ユーザは検索結果のスニペットを見て判断していることが多く、文書量の少ないスニペットからでも十分に高精度なクエリ拡張が行えるとしている。

Web 上の特徴的な情報源を利用する研究も行われている。堀ら⁴⁾は、Web 百科事典 Wikipedia から作成した拡張クエリは、Web 検索結果の疑似適合フィードバックから作成した拡張クエリよりも、ユーザ満足度が高くなることをユーザ実験によって示している。また、Web 上ではユーザ自身が積極的に情報を発信していることから、水野ら⁵⁾は、ユーザが記述した blog やブックマークから作成したユーザプロフィールを情報源とすることで、ユーザの趣向にあった拡張クエリを作成できるとしている。

多義的なクエリをクエリを推薦する手法としては今井ら⁶⁾の研究がある。今井らは、ク

エリと URL からなる 2 部グラフを用いたクラスタリングを行い、意味が偏らないクエリ推薦を行うことが可能であることを明らかにしている。

Yoon ら⁷⁾は、ユーザの要求はコミュニティQAのカテゴリと質問記事に出現する主要な名詞に対応するとして、Web 検索結果の分類をカテゴリを用いて分類する手法を提案している。高田ら⁸⁾は、コミュニティQAの質問記事と回答記事に着目し、回答記事の別解情報を含む Web ページを収集することで、質問記事に関連のある Web ページを効率的に閲覧できる手法を実現している。

コミュニティQAを用いて検索意図の候補を提示する研究には山本ら⁹⁾がある。質問記事の中に出現する形容詞と名詞で構成される語“修飾語付き観点”は、検索ユーザの検索意図であるとして、修飾語付き観点をユーザに提示することで、通常では思いつきにくい観点から検索が可能になるとしている。

以上で述べたように、従来のクエリ拡張では、検索精度を向上させることを目的に、キーワードを追加していた。情報源の違いは、抽出されるキーワードの違いに反映され、その評価は、精度や再現率などの検索精度の向上を狙いとしていた。本論文では、コミュニティQAの質問記事を“ユーザの情報要求を直接的に表現したもの”であるとして、検索ユーザに提示することを主たる目的とする。ユーザは“情報要求の候補”を自然言語で書かれたテキストとして閲覧することで、曖昧な情報要求を具体化することができると考える。本研究で生成される拡張クエリは、検索精度の向上だけでなく、ユーザの情報要求を適切に反映するものになることを目指している点で、従来研究と大きく異なる。

3. コミュニティQAを用いた拡張クエリの提案

本論文では、情報要求が曖昧なユーザに対し、拡張の根拠となるコミュニティQAの質問記事を対にした拡張クエリを提示する。提案手法を用いて拡張した質問記事付きの拡張クエリの提示システムを図1に示す。図1の上部は、コミュニティQAのカテゴリとタグクラウドを用いた多様性を持つクエリを展開している。図1の下部は、選択された拡張クエリの根拠となる質問記事を示し、より具体的な拡張クエリを提示する部分である。以下、それぞれについて詳述する。

3.1 多様性を展開する拡張クエリ

ユーザの情報要求は、入力されたクエリ、すなわちキーワードが同じであったとしても異なる場合がある。たとえば、“北海道”というクエリを入力した場合、そのユーザが“北海道の観光地”を知りたいのか、“北海道の名産”について知りたいのか、“北海道の方言”に

3 情報要求の言語化を支援するクエリ拡張型 Web 検索システムに関する一検討

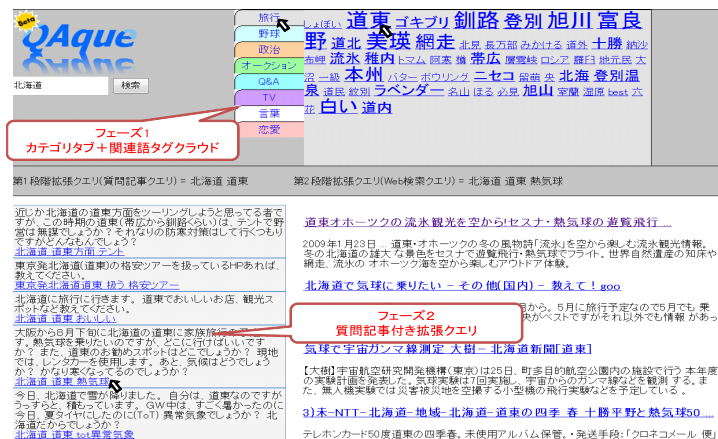


図 1 システムの実行画面

Fig. 1 Screenshot of our query expansion system.

ついて知りたいのか、入力された“北海道”のみから判断することは難しい。

このような多様な情報要求に対応するには、できる限り広い話題から多くの選択肢を提供することが必要と考える。提案手法では、幅広い話題をコミュニティQAのカテゴリ分類を用いて展開する。コミュニティQAでは質問者は質問記事を投稿する際、その質問に最もふさわしいと思うカテゴリを選択するのが一般的である。そのため、カテゴリは質問の内容をある程度反映していると考えることができる。本論文では、カテゴリ別に拡張クエリを作成することで、より多様な話題を持つ拡張クエリの候補を抽出する。

拡張クエリの提示方法として、情報検索の可視化技術として広く使用されているタグクラウドを用いる¹⁰⁾。ここでは、入力されたクエリに対する関連語が表示される。コミュニティQAの各カテゴリごとにタグクラウドを作成し、タブによってカテゴリを切り替える手法を提案する。カテゴリ別にタグクラウドを作成することで、タグクラウドには、カテゴリの特徴を反映した関連語が提示され、検索結果をカテゴリによって分類する手法⁷⁾の拡張と位置づけられる。また、タブに付与されるカテゴリ名は、ユーザの興味の方向性を示すものとなるため、ユーザは、タブを切り替えること同時に、自身の興味の方向性が変化することを確認できると考えられる。

3.2 拡張の根拠を示す質問記事付き拡張クエリ

提示される関連語の中には、ユーザにとって未知な語や、入力語との関連を把握できない

語が提示される場合がある。それらの語は、ユーザにとって仮に有用な語であったとしても、ユーザがそれに気付かなければ、有効に活用することができない。そこで、提案システムでは、選択された関連語に対して、入力語と関連語が含まれるコミュニティQAの質問記事を提示することで、拡張クエリの根拠を提示する。質問記事は自然言語によって記述されているため、ユーザは質問記事を読むことで、提示された関連語がどのような疑問に対して使用されているのか容易に把握することができる。拡張クエリの根拠が示されるため、ユーザは未知の単語であっても的確に選択することができるようになると考える。

コミュニティQAでは一問一答によって質問のやりとりが行われる。そのため、質問の内容は具体的であり、同じ関連語を含む質問記事でも、質問の内容が異なる場合も多い。そこで本システムでは、関連語を含む質問記事を複数提示することとした。その際に、質問記事の内容を反映したより具体的なキーワード組による拡張クエリを付与する。質問記事を参照しながらそこから発生した拡張クエリを使用することで、より具体的な検索を行うことが可能になる。

4. 拡張クエリの実装

提案システムでは、カテゴリ別に作成された関連語のタグクラウドとカテゴリタブ、そして質問記事を順次参照していくことにより、ユーザが自然に興味のある内容について情報要求を具体化させながら検索を行うことを可能にする。本システムでは、タグクラウドから関連語を選択したときに作成される“入力語+関連語”のクエリを第1段階拡張クエリ、質問記事に付与される拡張クエリを第2段階拡張クエリと定義する。提案手法を実装したシステムの構成を図2に示す。拡張クエリ作成のためのデータセットは、Yahoo!知恵袋コーパス^{*1}（2004年4月～2005年9月）から、2005年9月時点で投稿質問数が多かった上位10カテゴリを利用する。各カテゴリからランダムに20,000文書を抽出し、形態素解析器MeCab^{*2}により形態素解析し、以下の条件を満たすキーワードを抽出する。

- 動詞，形容詞，名詞（非自立，接尾，代名詞を除く）。
- 2文字以上で構成されている（漢字は1文字でも可）。
- MySQLのストップワードリスト^{*3}に含まれない。
- 3文書以上に含まれる。

*1 <http://research.nii.ac.jp/tdc/chiebukuro.html/>

*2 <http://mecab.sourceforge.net/>

*3 <http://dev.mysql.com/doc/refman/5.1/ja/fulltext-stopwords.html>

4 情報要求の言語化を支援するクエリ拡張型 Web 検索システムに関する一検討

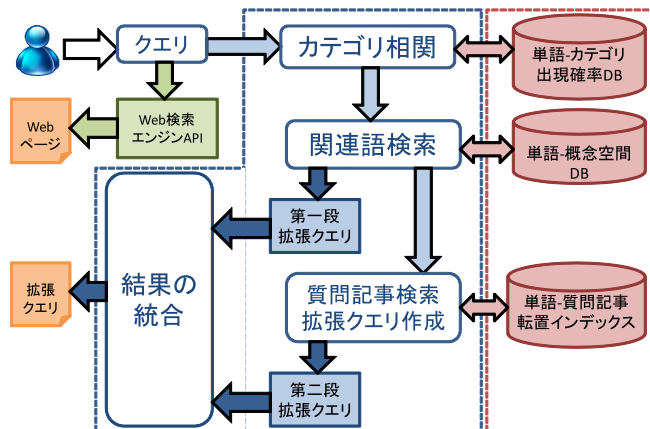


図 2 提案手法の構成

Fig. 2 Configuration diagram of our system.

表 1 上位 10 カテゴリと抽出キーワード数

Table 1 Number of keywords in top 10 categories.

カテゴリ	キーワード数	カテゴリ	キーワード数
政治・社会問題 (seiji)	9,781	健康・症状・ヘルスケア (health)	7,650
恋愛相談・人間関係の悩み (renai)	7,626	国内 (travel)	7,263
パソコン・周辺機器 (pc)	6,504	テレビ・ラジオ (tv)	8,560
Yahoo!オークション (auction)	6,521	野球 (baseball)	7,033
Yahoo!知恵袋 (bukuro)	7,350	言葉・語学 (kotoba)	8,257

使用したカテゴリと抽出したキーワード数を表 1 に示す．以後カテゴリ名は括弧内のローマ字表記を利用する．

4.1 第 1 段階拡張クエリの実装

第 1 段階拡張クエリでは，関連のあるカテゴリ順にタブを実装し，それぞれでタグクラウドを作成する．以下カテゴリタブの実装法とタグクラウドの実装法について詳述する．

4.1.1 カテゴリタブの実装

入力されたクエリに対して，関連のあるカテゴリ順に動的にタブを生成する．入力クエリとカテゴリとの関連度の指標として出現確率を用いる．ここで出現確率は，カテゴリ内の 20,000 質問記事に対して，入力クエリを含む質問が何件あるかを示す割合である．カテゴリ C ，入力語 t としたとき，カテゴリ内の全文書数を N_C ，入力語 t を含む質問記事数を

$N_{C,t}$ とすると出現確率 $P_{C,t}$ は， $P_{C,t} = \frac{N_{C,t}}{N_C}$ で表される．データセットの全カテゴリで出現確率を計算し，確率の高い順にソートすることでタブの表示ランキングを作成する．出現率が 0，すなわち，カテゴリ内のすべての質問記事に対して入力クエリの語が 1 度も出現しなかったカテゴリは表示しない．

4.1.2 タグクラウドの実装

タグクラウドに表示する関連語の検索には，潜在的意味インデキシング (LSI) を用いる．LSI を行うために，質問記事と出現単語の関係を表す文書・単語行列を作成する．行列の各要素を $tf.idf$ 値とし， tf 値は質問記事の語彙数で正規化する．質問記事 Q_j ，単語 t_i の要素 $q_{i,j}$ は以下のとおりである．

$$q_{i,j} = tf \times idf = \frac{\log_2(f_{i,j} + 1)}{\log_2(K_j)} \times \left(\log_2 \frac{N_C}{N_{C,i}} + 1 \right) \quad (1)$$

ここで $f_{i,j}$ は，質問記事 Q_j の単語 t_i の出現回数． K_j は，質問記事 Q_j での出現単語の種類数である．

得られた 20,000 次元の質問記事ベクトルを LSI によって 100 次元のクラスに次元圧縮することとした．次元圧縮には特異値分解 (SVD: Singular Value Decomposition) を用いる． $m \times n$ の質問記事行列 Q は以下のように分解できる．

$$Q_{m \times n} = U_{m \times r} \Sigma_{r \times r} V_{n \times r}^T \quad (2)$$

次元圧縮した行列から，単語間のコサイン距離による類似度を求め，類似度上位 50 件の単語をタグクラウドに表示することとした．

タグクラウドでは，表示の際の要素として，表示の順番と表示の大きさがある．一般に使用されるタグクラウドでは，表示の順番は単語のアルファベットや 50 音順，重要度や文書のヒット数を文字の大きさに割り当てる．また，意味の近い単語を同じ行に割り当てる手法も知られている．タグクラウドに表示される関連語は LSI による類似度，そして質問記事のヒット数という 2 つの評価尺度を有するそこで，順番は類似度順位の昇順，各単語の大きさは，その単語を選ぶことで検索される質問記事数によって決定する．コサイン距離，検索質問記事数はそれぞれ別の観点によって関連語を評価していることから，これらを組み合わせて提案する．検索される質問記事が 0 の関連語はタグクラウドには表示しない．従来のタグクラウドと比較し，より多くの観点から関連語を推薦することができる．

4.2 第 2 段階拡張クエリの実装

第 2 段階の拡張クエリの質問記事検索には，転置インデックスを用いる．入力クエリと第

5 情報要求の言語化を支援するクエリ拡張型 Web 検索システムに関する一検討

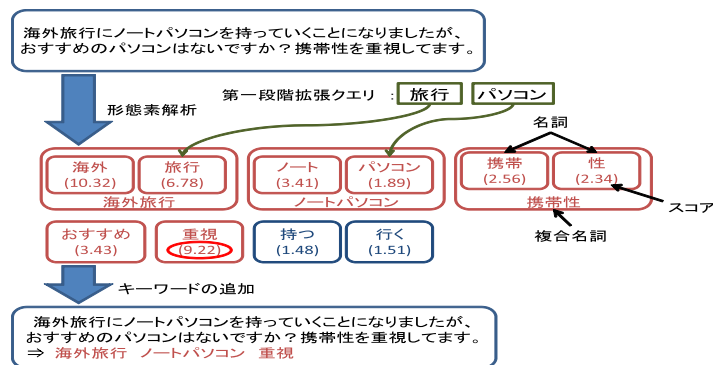


図 3 2 段階目キーワード作成

Fig.3 Flow of second stage query expansion.

1 段階の拡張クエリで選択された関連語により AND 検索を行い、AND 検索でヒットした質問記事のスコアを計算する。スコアは、質問記事中の各単語の *tf.idf* 値の総和とする。

第 2 段階目の拡張クエリの作成手順を図 3 に示す。まず、質問記事を形態素解析し、各形態素に分解する。形態素の中で名詞が連続している箇所を接続し、複合名詞を作成する。また、質問記事では、“きれいな景色”などの表現が多く使用される⁹⁾。そのため、形容詞と名詞を複合名詞と同様に連結する。次に、第 1 段階の拡張クエリのキーワードと比較し、第 1 段階の拡張クエリのキーワードが複合名詞の一部に含まれる場合、クエリのキーワードを複合名詞に置き換える。第 2 段階拡張クエリでは、まず無条件に第 1 段階の拡張クエリを使用する。図 3 の例では、第 1 段階拡張クエリとして“旅行”と“パソコン”が選択されるが、複合名詞化されているので“海外旅行”と“ノートパソコン”が第 2 段階拡張クエリとなる。最後に、転置インデックスに付与されている各単語の *tf.idf* 値から最もスコアの低い単語を、複合名詞化の処理を施してキーワード組に追加する。合計のキーワード数が 3 語になるまで、キーワードを追加する。

5. 拡張クエリの評価

本論文では、ユーザの情報要求の言語化支援を目的に、質問記事を拡張の根拠として提示する拡張クエリを生成する。そこで、実際に実装したシステムを用いてクエリを入力した際の出力結果の評価を 5.1 節で述べる。本手法の特徴である、多様性を持つ拡張クエリの評価を 5.2 節で述べる。

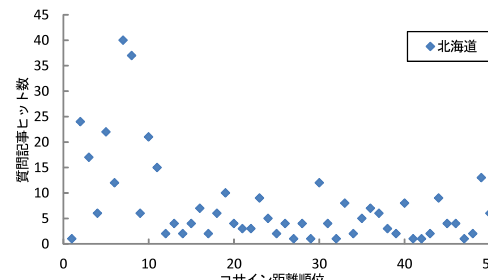


図 4 “北海道 + 関連語”の質問記事検索数

Fig.4 Number of question articles for query “Hokkaido and related terms”.

表 2 出現確率のランキング
Table 2 Category ranking from occurrence probability.

北海道		
順位	入力語	出現確率
1	travel	4.0×10^{-2}
2	baseball	3.9×10^{-3}
3	seiji	3.3×10^{-3}
4	auction	2.3×10^{-3}
5	bukuro	2.2×10^{-3}
6	tv	2.0×10^{-3}
7	kotoba	1.9×10^{-3}
8	renai	6.0×10^{-4}

5.1 システムの出力結果

提案システムにクエリが入力されると、入力クエリに関連のあるカテゴリ順にタグクラウドが作成される。入力語“北海道”について各カテゴリで出現確率を算出した結果を表 2 に示す。最も出現確率が高いカテゴリは travel カテゴリであり、2 位以下のカテゴリと比較して、約 10 倍の出現確率であることが分かる。一方で、kotoba カテゴリや auction カテゴリなど一見“北海道”と関係なさそうなカテゴリでも“北海道”が出現していることが分かる。タグクラウドは表示の順番を LSI のコサイン距離の順位、文字の大きさを検索される質問記事数としている。コサイン距離と質問記事検索数との関連を図 4 に示す。入力語“北海道”，travel カテゴリ，横軸はコサイン距離の順位，縦軸がクエリ“北海道+関連語”によって検索された質問記事数を表している。順位と検索質問記事数は対数的な分布になっていることが分かる。特に順位が下位の分布は順位に関係なく、質問記事が多数検索される関連語とほとんど検索されない関連語が混在している。15 位から 50 位までの順位と質問記事数の相関係数は 0.0256 であり、ほとんど相関がないことが分かる。

カテゴリごとに提示されるタグクラウドを図 5 に一覧する。travel カテゴリには“道東”，“富良野”，“ニセコ”などの観光地に関する関連語が提示されている。一方、野球カテゴリでは“苫小牧”，“優勝旗”などの高校野球関連，政治では“当選”，“比例”など選挙関連の単語が提示されていることが分かる。図 5 のタグクラウドから旅行カテゴリの“層雲峡”，野球カテゴリから“優勝旗”，政治カテゴリから“大地”を選択したときに表示される質問記事付き拡張クエリの例を図 6 に示す。“層雲峡”を含む質問記事を参照すると、層雲峡は千

6 情報要求の言語化を支援するクエリ拡張型 Web 検索システムに関する一検討

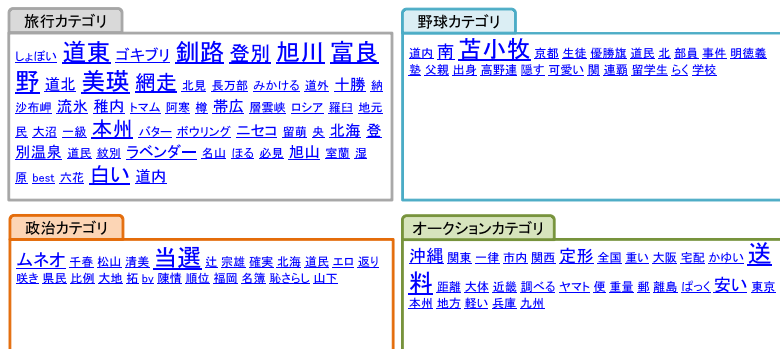


図 5 カテゴリごとのタグクラウド

Fig. 5 Tag-cloud in each category.

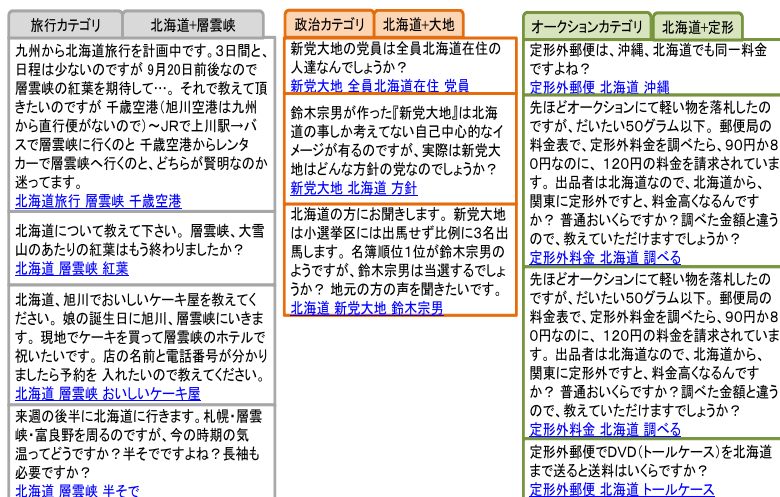


図 6 質問記事付き拡張クエリ

Fig. 6 Expansion queries with community question articles.

歳空港から電車やレンタカーで行くことのできる、紅葉の時期の観光地であることを知ることができる。そのほかにも店の情報や服装に関する質問記事を読覧することができる。政治カテゴリの“大地”は政治政党の名前であることを確認することができ、オークションカテ

ゴリでの“定形”では、北海道における郵便料金などに関する質問記事が提示されている。

5.2 Web ページからのキーフレーズ抽出実験

前節で多様な質問記事を提示できることを示したが、そのためには、タグクラウドに提示するキーワードに多様性が備わっていることが前提となる。そこで、本手法で提示させる拡張クエリの多様性を評価する。多様性の評価では、拡張クエリにおいて検索される Web ページが異なる内容について記述されたものかを調べる必要がある。たとえば、異なる URL の Web ページは異なる内容を扱うと仮定し、URL の重複を見ることで多様性を評価する手法¹¹⁾ などが知られている。

本論文では、拡張クエリから検索された Web ページからキーフレーズを抽出し、キーフレーズの種類数を比較する。理由は、同様の話題を扱う Web ページは、使用される単語も同様であり、単純に URL の異なりを調べただけでは多様なクエリ拡張が行えていると判断できない場合があると考えたためである。キーフレーズ抽出を行うことで、異なる話題の拡張クエリから検索される Web ページはより多くのキーフレーズを抽出できると考えられ、異なる URL で同様の話題を扱う Web ページが多様性に貢献しないと判断できる。

5.2.1 キーフレーズによる多様性の評価

提案するキーフレーズを用いた評価手法の妥当性を評価するために、商用の検索エンジンである bing^{*1}、Google^{*2}、Yahoo! JAPAN^{*3}の拡張クエリについて、評価実験を行う。

評価手法を、既存の検索エンジンのクエリ拡張を用いて実験評価する手順を以下に示す。

- (1) 作成した拡張クエリを Web 検索 API^{*4}に送信し、Web ページ検索結果を入手
- (2) 入手した Web ページ検索結果からスニペットを抽出
- (3) キーフレーズ抽出 API^{*5}を用いて、スニペットごとにキーフレーズを抽出
- (4) スコアが閾値以上キーフレーズを選択
- (5) Web ページ上位 10 件で 2 回以上出現したキーフレーズをカウント

API のスコアは 0 ~ 100 に設定されている。今回は 50 を閾値として設定する^{*6}。1 回しか

*1 <http://www.bing.com/?cc=jp>

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

*3 <http://www.yahoo.co.jp/>

*4 <http://developer.yahoo.co.jp/webapi/search/>

*5 <http://developer.yahoo.co.jp/webapi/jlp/keyphrase/v1/extract.html>

*6 API のスコアは相対値であり、必ずスコア 100 のキーフレーズが存在する。そのため、どの Web ページでも最低 1 個はキーフレーズを抽出できる。一般語や日付など、内容を表すキーフレーズにふさわしくないキーフレーズのスコアを調査したところ 30 以下だったため、今回はより特徴的な語に絞ることを含めて 50 に設定した。

7 情報要求の言語化を支援するクエリ拡張型 Web 検索システムに関する一検討

表 3 入力語 “ウイルス” の拡張クエリ
Table 3 Expansion queries for “virus”.

bing	Google	Yahoo! JAPAN
ウィルスバスター	ウィルスバスター	ウィルスソフト 無料
ウィルスバスター 2011	ウィルス	ウィルスチェック
ウィルスソフト	ウィルスソフト	ウィルスソフト
ウィルスバスター 2010	ウィルスバスター 2011	ウィルスバスター
ウィルスバスター 無料	ウィルスミス	ウィルス性胃腸炎
ウィルス対策ソフト	ウィルスチェック	ベートーベンウィルス
ウィルスミス	ウィルス対策ソフト	ロタウィルス
ウィルス対策ソフト フリー	ウィルスセキュリティ	ノロウィルス

表 4 入力語 “mac” の拡張クエリ
Table 4 Expansion queries for “mac”.

bing	Google	Yahoo! JAPAN
マクドナルド	mac	mac book
マック	macbook pro	mac book air
マクドナルド クーポン	macbook air	mac Windows
マクドナルド メニュー	macbook	mac OS
マクロス F	mac アドレス	mac book pro
マクロミル	mac mini	mac DVD コピー
幕張メッセ	mac os	MAC アドレス
マカフィ	mac air	mac air

出現しないキーフレーズはその Web ページ固有のものである可能性が高いため、複数ページで使用されたものを話題を示すキーフレーズと設定した。スニペットのみを用いるのは 1 つの Web ページに複数の話題が記述されている可能性があるため、クエリ周囲の文章のみのスニペットを用いることで、クエリ周囲の話題のみを抽出するためである。

評価は、“ウイルス”と“mac”の 2 つの入力語で行っている。この語は、複数の意味を持つ単語であり、推薦される拡張クエリに違いが出やすいと判断したためである。bing で入手できる拡張クエリ数が最大 8 件なので、8 個の拡張クエリでのキーフレーズ抽出数を比較する。検索エンジンによって推薦された“ウイルス”の拡張クエリを表 3 に、“mac”の拡張クエリを表 4 に示す。また、各検索エンジンの拡張クエリによって検索された Web ページから抽出されたキーフレーズ数の一覧を表 5 に示す。各検索エンジンでキーフレーズの抽出数に違いがでていることが分かる。“ウイルス”において、bing ではコンピュータ関連のクエリが 9 個推薦されている。Google も同様に 9 個がウイルスセキュリティ関連のクエリである。Yahoo! JAPAN は 4 個がコンピュータに関連するクエリであるが、他の 3 個は病気関連、1

表 5 各検索エンジンのキーフレーズ抽出数
Table 5 Number of keyphrase in each Web search engine.

入力クエリ	bing	Google	Yahoo! JAPAN
ウイルス	66	82	91
mac	60	43	58

個はテレビドラマに関する語が提示されている。キーフレーズ数では、Yahoo! JAPAN が最も多くのキーフレーズを抽出でき、bing が最も抽出できたキーフレーズ数が少ない。“mac”では、bing は飲食店関係が 4 個、そのほか、アニメ、Web サービス、イベントホール、セキュリティソフトに関するクエリが推薦されている。Google と Yahoo! JAPAN はすべてコンピュータ関連のクエリが推薦されているが、Google はコンピュータの mac 製品に関するキーワードが大半なのに対し、Yahoo! JAPAN は“windows”や“dvd”など mac 製品以外に対するキーワードも含まれている点が異なる。抽出キーフレーズ数は、bing が最も多く、Yahoo! JAPAN、Google の順で抽出できたキーフレーズ数が減少している。

5.2.2 既存手法との比較による多様性の評価

提案手法の多様性を前項と同様の手法で評価する。実験では、1 語の入力クエリから 30 個の拡張クエリを作成する。入力するクエリは各カテゴリで出現率の高かった語の中で、10 カテゴリすべてに存在する 10 語を用いる。入力クエリとそれに代表されるカテゴリを表 6 に示す。比較対象として Yahoo! 関連検索ワード API ^{*1} によって抽出した拡張クエリを用いる。以下の cqa と yahoo の 2 つの手法を比較する。

- cqa: 各カテゴリの第 1 段階拡張クエリ上位 3 語を 10 カテゴリまとめた 30 個
- yahoo: Yahoo! 関連検索ワード API によって推薦された拡張クエリ 30 語

例として、入力語“ソフト”の拡張クエリの一部を表 7 に示す。作成したは前項のキーフレーズ抽出手順に従い、キーフレーズが抽出される。

各入力語に対して、cqa と yahoo の 30 個の拡張クエリにより検索した Web ページから抽出したキーフレーズの合計を図 7 に示す。横軸は各入力語、縦軸は Web ページから抽出できたキーフレーズの合計数である。“ソフト”以外の入力語に対して、提案手法が既存手法を上回っている。

次に、キーフレーズ数の増加の推移のグラフを図 8 と図 9 に示す。図 8 が“日本語”、図 9 が“ソフト”の結果である。横軸は拡張クエリ、縦軸は抽出したキーフレーズ数の累計

*1 <http://developer.yahoo.co.jp/webapi/search/assistsearch/v1/webunitsearch.html>

8 情報要求の言語化を支援するクエリ拡張型 Web 検索システムに関する一検討

表 6 入力語とそのカテゴリ

Table 6 Input queries and categories.

カテゴリ	入力語	カテゴリ	入力語
seiji	中国	health	検査
renai	友達	travel	東京
pc	ソフト	tv	番組
auction	メール	baseball	選手
bukuro	質問	kotoba	日本語

表 7 入力語“ソフト”の拡張クエリ

Table 7 Expansion queries for “Soft”.

cqa	yahoo
ソフト フリー	ソフトバンク
ソフト 読み上げる	DS ソフト
ソフト シェア	マイクロソフト
ソフト ゲーム	wii ソフト
ソフト os	フリーソフト
ソフト 使う	PSP ソフト
ソフト ハード	PS3 ソフト
ソフト 使い捨て	ソフトバンクホークス
ソフト コンタクトレンズ	解凍ソフト

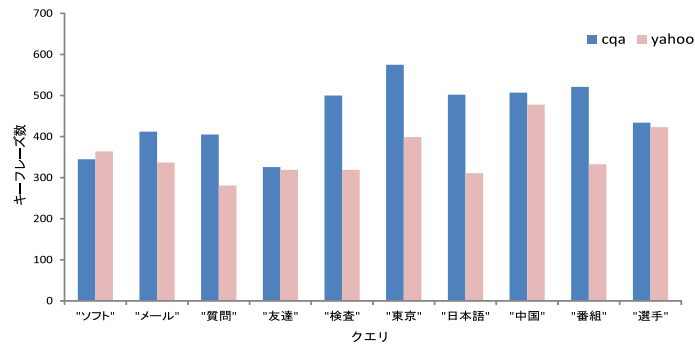


図 7 提案手法 (cqa) と既存手法 (yahoo) のキーフレーズ抽出数

Fig. 7 Number of keyphrase of our system and baseline.

である。“日本語”では、クエリ数が多くなるに従い、提案手法 (cqa) と既存手法 (yahoo) のキーフレーズ抽出数の差が、大きくなっていくことが分かる。一方で、“ソフト”は、提案手法、既存手法ともにクエリ数に比例して同様にキーフレーズ数が増加している。

5.2.3 抽出カテゴリ数と多様性の評価

5.2.2 項で用いた入力クエリは、提案手法において、用意した 10 カテゴリすべてで拡張クエリが推薦されている。しかし、実際には、10 カテゴリすべてのカテゴリにおいて拡張クエリが作成されるとは限らない。そこで、より一般的な入力クエリで拡張クエリを作成し、キーフレーズ抽出数について評価する。入力クエリには 2004 年と 2005 年の Yahoo! JAPAN の検索ワードランキング*1 の総合ランキングから“ANA”，“dell”，“hotmail”などの企業

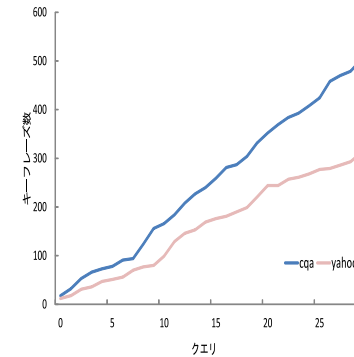


図 8 “日本語”でのキーフレーズ抽出数の推移

Fig. 8 Transition of keyphrase for query “Japanese”.

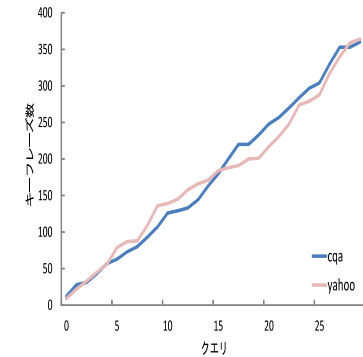


図 9 “ソフト”でのキーフレーズ抽出数の推移

Fig. 9 Transition of keyphrase for query “Soft”.

表 8 全クエリに対する提案手法と既存手法の比較

Table 8 Comparison of our system with yahoo in all queries.

	cqa	yahoo
キーフレーズ抽出数が比較手法を上回ったクエリ数	22	8
平均キーフレーズ抽出数	350.2	312.2

名やサービス名，“天気”や“地図”など一般生活に関連する語などを幅広く 30 語抽出したものを使用する。各入力クエリに対して、30 個の拡張クエリを作成する。今回の入力クエリでは、抽出できる関連カテゴリは 10 カテゴリとは限らない。入力クエリ q に対して、 N_q 個の関連カテゴリが抽出できたとき、1 つのカテゴリ c から抽出する拡張クエリ数 $n_{q,c}$ は $n_{q,c} = \lceil \frac{30}{N_q} \rceil$ 、端数は最も関連の低いカテゴリの拡張クエリを切り捨てることにより調整する。

全入力で拡張クエリを作成し、キーフレーズ抽出を行った結果を表 8 に示す。全 30 クエリのうち 22 クエリにおいて、提案手法が既存手法よりも多くのキーフレーズを抽出できている。また、平均値においても提案手法が多くのキーフレーズを抽出できている。

次に、各クエリにおいて、抽出できたカテゴリ数とキーフレーズ抽出数の関連を図 10 に示す、横軸が抽出できたカテゴリ数、縦軸がキーフレーズの抽出数である。また、回帰分析

*1 <http://picks.dir.yahoo.co.jp/new/review2004/>
<http://picks.dir.yahoo.co.jp/new/review2005/>

9 情報要求の言語化を支援するクエリ拡張型 Web 検索システムに関する一検討

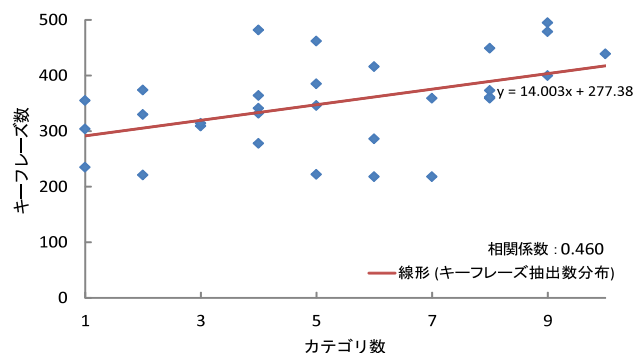


図 10 抽出カテゴリ数とキーワード数の相関
Fig. 10 Correlation of categories and keyphrases.

を行い、最小二乗法による回帰直線を付与した。使用した入力クエリによる抽出カテゴリ数は 1 から 10 まで幅広く分布した。“ANA”、“dell”は抽出カテゴリ数は 1, “hotmail”は 2 であるのに対し, “天気”は抽出カテゴリ数 9, “地図”は 7 と固有名詞よりも一般的な語の方がより多くのカテゴリを抽出できていた。1 カテゴリでも 300 程度のキーワードを抽出できているが、回帰分析の結果、抽出カテゴリ数を多くするにつれ、キーワード抽出数が増加する傾向が見られた。相関係数は 0.46 であり、中程度の正の相関があるといえる。特に、カテゴリ抽出数が 8 以上のクエリについて、多くのキーワードが抽出できていた。

6. 考 察

6.1 拡張の根拠を示すクエリ拡張の有効性

図 5 より, “北海道”という入力語に対して, 多様なクエリ拡張を行えていることが分かる。各カテゴリのタグクラウドでは, 各カテゴリの内容に関連のある語が推薦されていると考えられる。選んだカテゴリがそのまま話題に反映されることから, ユーザは興味のあるカテゴリを選択することで, 自身の興味のある話題についての的確に拡張クエリを使用し, Web 検索を行うことができると考えられる。

図 6 より, 拡張の根拠を示すことで, ユーザは, 自身の興味に対して体系的に知ることができると考えられる。たとえば, “北海道”に対する旅行カテゴリでは, 北海道の観光地に関する地名などが多く提示されるが, 層雲峡について知らない場合, 通常の Web 検索で層雲峡を検索しても, 層雲峡に関する一般的な観光案内のページしか提示されない。しかし, 提

案システムを使用することで, 層雲峡の見所や層雲峡までの移動手段, 服装や気候などの注意点を調べることが可能になる。特に服装面などは見落としがちな点であると考えられる。提案手法のシステムを使用することで“調べなくてはいけないこと”を一挙に把握できると考えられる。また, 拡張クエリのみで検索しても分かりにくいものに対しても効果的であると考えられる。“北海道 大地”は, 政治カテゴリにあるが, このクエリで検索を行った際に提示される結果は“北海道の高原などの大地”に関する画像のページが主であり, 政治カテゴリと関連する政党大地については, 1 ページ目の検索結果では 1 件しか検索されない。しかし, 質問記事を参照し, 政党大地を認識したうえで, 第 2 段階拡張クエリに提示される“新党大地 北海道 方針”などのクエリを用いることで, 的確に政党大地に関する Web ページを検索できる。興味のある検索語とカテゴリを決定するだけで, ユーザ自身が何に興味があるのか, を認識でき, 次に何を検索すればいいのか理解することができると考えられる。

6.2 拡張クエリの多様性と評価

5.2.1 項の実験では, 検索された Web ページのスニペットからキーワードを抽出することで多様性を評価できるのかを検証した。異なる話題に関するクエリを提供している検索エンジンが, 多くのキーワードを抽出できていたことから, キーワードが多様性の評価に有効に機能していると考えられる。提案手法と既存のクエリ拡張の比較では, ほとんどの場合において提案手法の方がより多くのキーワードを抽出できていたことから, 提案手法は多様性を持ったクエリ拡張が行われていると考えることができる。

図 10 より, 展開されるカテゴリが多ければ多いほど, 多様性を展開できると考えられる。抽出カテゴリ数と, キーワード抽出数の相関実験では, 中程度の正の相関となったが, 実験で使用した 30 クエリに表 6 を含めると, 相関係数は 0.589 となりより強い相関となる。データ数をさらに増やすことでより相関があることが確認できるのではないかと考えられる。また, 図 10 では, 単一のカテゴリのみで拡張クエリを作成してある程度のキーワードが抽出できた。これは, コミュニティ QA の性質によるものであると考えられる。コミュニティ QA では, 一問一答の形式で質問のやりとりを行うため, 個々の質問記事は完結した内容で具体的な内容が記述されていること, コミュニティ QA 内では, 他人の質問, 回答を閲覧できるため, 同様の内容の質問が先に存在している場合, 質問を投稿する必要がなくなる。膨大な数に及び同カテゴリ内での質問記事も内容が細かく異なっているため, 提示される関連語にも多様性が発生したのではないかと考えられる。

7. おわりに

本論文では，ユーザの情報要求の言語化を支援する，コンテキストを提示するクエリ拡張法を提案した．拡張クエリにコミュニティQAのカテゴリ分類を用いることでより幅広い多様な話題に関する拡張クエリを作成した．作成した拡張クエリによって検索されるWebページからキーフレーズを抽出することにより，拡張クエリの多様性を検証した．

提示される拡張クエリのカテゴリから，自身の疑問の方向性に合うカテゴリを選択し，選択したカテゴリ内のタグクラウドの中からさらに興味のある関連語を選択する．最後にクエリと選択した関連語が含まれる質問記事から，最も興味のあるクエリを選択することで，最終的に明確に言語化された情報要求での検索が可能になる．キーワードを閲覧する際に未知のキーワードが提示されても，質問記事やカテゴリ名を見ることで，その語がどのような疑問に使用され何を意味しているのかを判断できる．

今後の課題として，カテゴリの選択があげられる．実際のコミュニティQAでは，階層構造を持つ数百のカテゴリが存在する．それらのカテゴリには内容が近いカテゴリも存在している．そのため，多様性を保ちつつ，入力と関連のあるカテゴリを選択することは重要となる．適切なカテゴリを選択することで，ユーザの情報要求を反映した拡張クエリが提供できるようになると考えている．

謝辞 本研究の一部は科研費(21500091)の助成を受けたものである．本研究の実装・評価に際し，大学共同利用機関法人国立情報学研究所から提供を受けた，Yahoo!知恵袋のデータを利用している．ここに記して謝意を示す．

参 考 文 献

- 1) Manning, C.D., Raghavan, P. and Schuetze, H.: *Introduction to Information Retrieval*, Cambridge University Press (2008).
- 2) Yin, Z., Shokouhi, M. and Craswell, N.: Query Expansion Using External Evidence, *31th European Conference on IR Research (ECIR2009)*, LNCS 5478/2009, pp.362–374 (2009).
- 3) 村田眞哉，戸田浩之，松浦由美子：検索結果中のアクセス集中サイトを利用したクエリ拡張法の提案，*DBSJ Letters*, Vol.6, No.4, pp.45–48 (2008).
- 4) 堀憲太郎，大石哲也，長谷川隆三，藤田 博，越村三幸：Wikipediaからの拡張クエリ生成によるWeb検索とその評価，*人工知能学会研究会資料*，No.SIG-SWO-A803, pp.13-1–13-7 (2008).
- 5) 水野淳太，村田祐一，勝屋 久：ユーザの嗜好を反映したクエリ拡張を用いた情報検

索・推薦システムの開発，*楽天研究開発シンポジウム 2009* (2009).

- 6) 今井良太，戸田浩之，関口裕一郎，望月崇由，鈴木智也，今井桂子：Web検索サービスにおける多義的なクエリ推薦手法，*DBSJ Journal*, Vol.9, No.1, pp.1–6 (2010).
- 7) Yoon, S., Jatowt, A. and Tanaka, K.: Intent-Based Categorization of Search Results Using Questions from Web Q&A Corpus, *Proc. 10th international conference on Web Information Systems Engineering (WISE2009)*, LNCS 5802/2009, pp.145–158 (2009).
- 8) 高田夏希，大島裕明，田中克己：WebとQAコンテンツの相互補完に基づくソーシャルサーチ，*Webとデータベースに関するフォーラム 2010*，2A-3 (2010).
- 9) 山本岳洋，中村聡史，田中克己：QAコンテンツからの観点抽出とそれにもとづくウェブ検索結果の再ランキング，*Webとデータベースに関するフォーラム 2010*，2A-2 (2010).
- 10) Hassan-montero, Y. and Herrero-solana, V.: Improving Tag-Clouds as Visual Information Retrieval Interfaces, *International Conference on Multidisciplinary Information Sciences and Technologies (InScit2006)*, pp.25–28 (2006).
- 11) Strohmaier, M., Kroell, M. and Koerner, C.: Intentional Query Suggestion: Making user goals more explicit during search, *Proc. 2009 workshop on Web Search Click Data (WSCD'09)*, pp.68–74 (2009).

(平成 23 年 3 月 20 日受付)

(平成 23 年 7 月 7 日採録)

(担当編集委員 戸田 浩之)



大塚 淳史

2011 年筑波大学情報学群知識情報・図書館学類卒業．同年より筑波大学大学院図書館情報メディア研究科博士前期課程に在籍．Web情報検索，情報インタラクションに興味を持つ．日本データベース学会学生会員．



関 洋平 (正会員)

1996 年慶應義塾大学大学院理工学研究科計算機科学専攻修士課程修了。2005 年総合研究大学院大学情報学専攻博士後期課程修了。博士 (情報学)。同年豊橋技術科学大学工学部情報工学系助手。2008 年コロンビア大学コンピュータサイエンス学科客員研究員。2010 年筑波大学大学院図書館情報メディア研究科助教、現在に至る。自然言語処理、意見分析等の研究に従事。ACM, ACL, 電子情報通信学会, 言語処理学会, 日本データベース学会各会員。



神門 典子 (正会員)

1994 年慶應義塾大学大学院文学研究科博士課程修了。博士 (図書館・情報学)。同年学術情報センター助手。1995 年米国シラキウス大学情報学部客員研究員。1996~1997 年デンマーク王立図書館情報大学客員研究員。1998 年学術情報センター助教授。2000 年国立情報学研究所助教授。2002 年より総合研究大学院大学助教授を併任。2004 年より国立情報学研究所ならびに総合研究大学院大学教授、現在に至る。テキスト構造を用いた検索と情報活用支援、探索や学習のための情報アクセス技術、情報検索システムの評価等の研究に従事。ACM-SIGIR, ASIS&T, 言語処理学会, 日本図書館情報学会各会員。



佐藤 哲司 (正会員)

1980 年山梨大学工学部電子工学科卒業。同年日本電信電話公社武蔵野電気通信研究所に入所。以来、論理回路の大規模一括集積技術、データベースマシン、マルチメディアデータベースの研究・開発に従事。1994 年工学博士 (大阪大学) 取得。2007 年 4 月より現職。分散並列処理、情報検索、社会インタラクションに興味を持つ。電子情報通信学会会員。