# 情報要求の言語化支援のための コンテキスト提示型クエリ拡張法の提案と評価

大塚 淳史[†1] 関 洋平[†1]
神門 典子[†2] 佐藤 哲司[†1]

膨大な Web 情報空間から所望の情報を検索する手段として Web 検索エンジンは幅広く普及している．Web 検索を利用する場面も多岐に渡り，検索ユーザは曖昧な情報要求によって検索を行うことも多くなった．曖昧なクエリでは，的確な検索結果を得ることは難しいが，クエリを拡張しながら検索を繰り返すことで，所望の結果を得られることから，クエリ拡張法に関する研究が盛んに行われている．本論文では，コミュニティQA に投稿された質問記事を使用して，クエリ拡張の候補となる関連キーワードを提示する際に，拡張の根拠となる質問記事を提示するコンテキスト提示型のクエリ拡張法を提案する．ユーザは，コミュニティQA のカテゴリや質問記事を切り替えていくことにより，様々なコンテキストから発生した拡張クエリを選択することができる．潜在的トピックモデルと単語の出現確率を用いて，入力されたクエリに関連するカテゴリから関連キーワードを検索する手法を実現したので報告する．

# Query Suggestion Method considering question's context on Community QA Resource to Verbalize Latent Information Need

Atsushi OTSUKA,[†1] Yohei SEKI,[†1] Noriko KANDO[†2] and Tetsuji SATOH[†1]

The Web search engines are used widely as way to obtain desired information from Web space. With increase using a scene Web search,Web search users often use search engines with ambiguous information needs. It is difficult to obtain desired information with ambiguous query. But, information Needs are verbalized gradually by repeated Web search with expansion query. So,query expansion study is very prosperous. In this paper,we propose query suggestion method considering question's context to verbalize latent information needs. By switching categories and question articles, web search user will find expansion queries relevant to their information needs from different contexts. We also report category select method by using latent topic model and probability of occurrence of words.

## 1. はじめに

情報要求を言語化したクエリを入力とする Web 検索エンジンでは，膨大な Web 情報空間から，所望の情報を検索する手段として幅広く普及している．検索エンジンは，ユーザが入力したキーワードをクエリとして解釈し，適合した Web ページを提示するが，検索結果が 1000 件を超える場合も少なくない．しかし，多くのユーザは，提示された検索結果のごく一部しか閲覧しないことが，オプト社とクロス・マーケティング社が 2006 年に行った調査[⋆1]で報告されている．多くの検索ユーザが，平均的に検索結果を 4 ページ程度しか閲覧せず，その中に満足するページがなければ，約 9 割のユーザがクエリのキーワードを変更，追加を行い再検索している．このことから，検索結果の上位にユーザの要求に合致した結果を提示するだけでなく，検索クエリとなるキーワードの変更や追加を支援するクエリ拡張技術が重要となる．

Web 検索を使用する際，ユーザは自らの知りたいことである情報要求を頭の中で整理し，キーワードの組み合わせとして言語化する．しかし，モバイル端末等の普及により，ユーザは時間や場所を問わず気軽に Web 検索が利用出来るようになったことから，情報要求が必ずしも詳細に言語化されていない状態で検索を行う機会が増加している．例えば，マスメディア等に紹介されたキーワードについて検索を行う場合や，外出先の空き時間た旅行先での情報収集を行う場合には，検索のゴールが明確に定まっておらず，情報要求を言語化させることが困難となる．漠然とした “調べたいこと” を検索エンジンに入力し，得られた検索結果から，興味のある情報を探し出し，そこから次の検索のためのクエリを作成することになる．このようなプロセスを繰り返すことにより，ユーザは最終的に最も知りたいことを序々に明確にしていく．

ユーザの Web ページの再検索を支援する手法として，クエリ拡張がある．クエリ拡張で

†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
⋆1 検索エンジン利用状況実態調査
http://www.opt.ne.jp/news/pdf/pr/20060425_PR_SearchEngine.pdf

は，ユーザが入力したクエリに対して，関連語の追加や変更を行ったクエリを検索結果とともに提示する手法である．ユーザは提示された候補の中からクエリを選択するだけで，再検索を行うことができることから，商用の Web 検索エンジンでも積極的に使用されている．拡張クエリは，検索エンジンのクエリとしてそのまま使用できる形式で提供させるため，キーワードの組で提供される．提示されたキーワードを知らないユーザにとっては，提示されたキーワードの意味やコンテキストを理解出来ないため，それがユーザの情報要求を満足させる拡張クエリであったとしても，ユーザはそれに気がつくことができず，有効に活用されないという問題がある．

筆者らは，拡張クエリのキーワード組と同時にその拡張の根拠を提示し，拡張のコンテキストが理解できる拡張クエリの作成法を提案している[1)]．検索ユーザの情報要求は，ユーザの置かれている環境や状況によって変化する．様々なコンテキストにから発生した情報要求をもとに作成された拡張クエリを提示することで，検索ユーザは，自身のコンテキストと拡張クエリのコンテキストを比較することができ，自身の情報要求に合致する拡張クエリを見つけ出すことができる．本論文では，拡張のコンテキストに対応するものとしてコミュニティQA を使用する．コミュニティQA には，様々な状況における質問記事が多数投稿されている．また，コミュニティQA では，質問を投稿する際，質問の内容にふさわしいと思われるカテゴリを選択する．質問記事とカテゴリはコミュニティQA ユーザの情報要求のコンテキストが強く反映されたものなっているといえる．質問記事と投稿カテゴリを，拡張クエリのコンテキストとしてユーザに提示することで，検索ユーザは自身のコンテキストに近いと思うカテゴリの質問記事と質問記事から作成された拡張クエリによって，情報要求を明確に言語化した検索を行うことが可能になる．

カテゴリは階層関係を持ち，百種類以上のカテゴリを持つ．そのため，すべてのカテゴリで拡張クエリを作成し，一度にユーザに提示することは困難である．ユーザの入力したクエリから，ユーザのコンテキストに関連する適切なカテゴリを選択することが重要となる．本論文では，カテゴリ分類の特徴である階層関係を利用し，それぞれの階層で潜在的ディリクレ配分法（LDA）による，トピックモデルを作成し，出現する関連語の比較を行い，適切なカテゴリの選択方法を評価する．

本論文の構成は以下のとおりである．2 章で先行する関連研究について述べ，3 章でプロトタイプシステムについて説明し，本論文で提案する拡張クエリの作成法について述べる．4 章で評価実験を行い，5 章で考察する．6 章でまとめと今後の課題について述べる．

## 2. 関 連 研 究

クエリ拡張では，様々な外部情報を使用する研究が盛んに行われている．掘ら[2)] は，Web 百科事典である Wikipedeia から作成した拡張クエリと Web 検索結果の擬似適合フィードバックから作成した拡張クエリをユーザ満足度により比較し，Wikipedeia から作成した拡張クエリのほうがより満足度が高くなることを示した．水野ら[3)] は，Web 上では，ユーザが積極的に情報発信しているという特徴を利用し，ユーザが作成した blog やブックマークからユーザの趣向を推定し，ユーザの趣向に基づいた拡張クエリを提案している．より多様なクエリを推薦する手法には，今井ら[4)] の研究がある．今井らは，クエリと URL から 2 部グラフを作成し，クラスタリングを行った結果，意味が偏らないクエリを推薦することができることを明らかにしている．外部情報源の違いは生成されたキーワード組の違いにのみに反映される．その評価はキーワード組によって検索される検索精度の向上や，ユーザ評価の違いによって行われる．それに対して本研究では，情報源である質問記事をキーワード組と共に提示する．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ことを目指す．

クエリの意味を提示する研究には Guo[5)] や，廣嶋ら[6)] らの研究がある．Guo らは提示された拡張クエリに，Web 上で多く利用れているタグによるソーシャルアノテーションを付与することで，提示された拡張クエリの意味を提示する手法を提案している．廣嶋らはユーザが入力したクエリを “グルメ’，“スポーツ”，“企業名” などのタイプに分類し，タイプに応じた Web 検索結果を提示している．クエリの意味を提示する研究では，クエリログやアクセスした URL のログから自動でタグやタイプを推定していた．本研究では，拡張クエリのコンテキストを質問記事に反映させるため，意味推定のための処理を行わない点で従来研究とは異なる．

コミュニティQA と Web 検索を結びつける関連研究として Yoon ら[7)] の研究では，ユーザの要求とコミュニティQA のカテゴリを関連付け，Web 検索結果をコミュニティQA のカテゴリにより分類，再ランキングを行う手法を提案している．山本ら[8)] は，コミュニティQA から形容詞と名詞の組み合わせによる修飾語付き観点を抽出し，タグクラウドとして，ユーザに提示する．修飾語付き観点はユーザがより直感的に分かりやすい表現となっている．実験により，修飾語付き観点は，これまでの検索ではなかなか思い浮かばない意外な組み合わせの語が推薦されることを明らかにしている．本研究では，コミュニティQA の質問記事の一部ではなく，すべてをユーザの情報要求ととらえるという点で従来研究とは異な

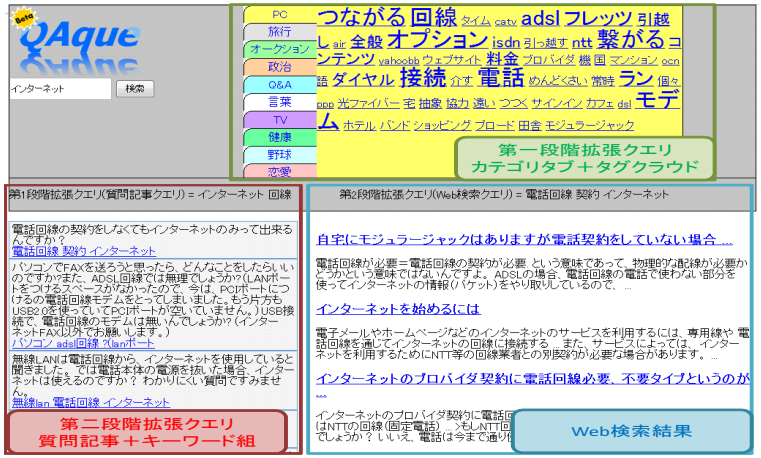


**図 1**　プロトタイプシステム画面

る．また，質問記事は拡張クエリの根拠を提示するために使用するため，質問記事に対する回答記事は使用しない．

## 3. コンテキスト提示型クエリ拡張法の実装

本論文では，情報要求が言語化出来ていないユーザに対して，コンテキスト理解のための，コミュニティQA 質問記事付きの拡張クエリを提示する．3.1 節で筆者らが開発した拡張クエリ提示のためのプロトタイプシステムのインターフェースについて説明する．次に，3.2 節で，コミュニティQA のカテゴリの選択手法について述べ，拡張クエリの関連キーワードの検索法について説明する．

### 3.1 言語化支援のためのインターフェース

筆者らが開発した[1) ]プロトタイプシステムを図 1 に示す．本システムは，第一段階拡張クエリ，第二段階拡張クエリ，Web 検索結果の 3 つの領域から構成される．ユーザは第一段と第二段で提示された拡張クエリを Web 検索結果を見ながらクエリを切り替えていくことにより，情報要求が序々に言語化されていき，最終的に目的を明確にした検索を行うことが可能になる．

(1)　**タブとタグクラウドによるコンテキストの多様性の展開**

検索エンジンに入力されたクエリが同じだったとしても，その背後にある情報要求は

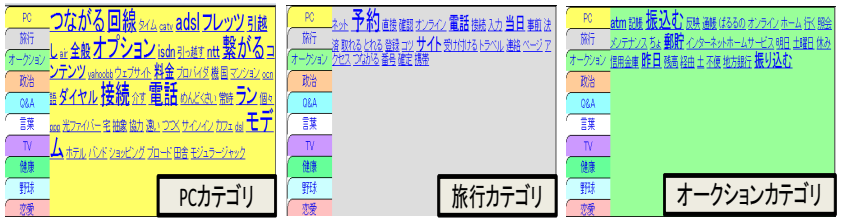


**図 2**　カテゴリごとのタグクラウド

多岐に渡る．特に情報要求が言語化されていない段階では，ユーザはどの様な話題に関して興味があるのか確定していない場合が多い．提示する拡張クエリを幅広い話題から提供することで，より広いコンテキストから発生した情報要求をユーザに提示することができる．本論文では，コミュニティQA のカテゴリ分類を利用し，カテゴリ別に関連キーワードを提示する．関連キーワードの表示方法として，タグクラウドを用いる．カテゴリごとに関連キーワードのタグクラウドが作成され，カテゴリのタブによって，カテゴリを切り替えることができる．図 2 は “インターネット” を入力とした場合にタグクラウドに表示されるキーワードである．PC カテゴリでは “接続”，“adsl” などのキーワードが推薦されているのに対して，旅行カテゴリでは “予約”，オークションカテゴリでは “ATM”，“振込む” などのキーワードが推薦される．推薦されるキーワードはそのカテゴリの内容を反映しているため，カテゴリを切り替えることによって，より多くの話題についての検索を行うことが可能になる．

(2)　**コンテキストの段階的提示による言語化支援**

プロトタイプシステムによって提供される拡張クエリは 1 カテゴリにつき最大 50 件，10 カテゴリ分提示した場合，それぞれ拡張クエリが作成されるため，ユーザに提示される拡張クエリは膨大な数に及ぶ．情報要求が言語化出来ていないユーザにとって，その中から自分の興味のあるキーワードを選ぶことは困難である．そのため，本システムでは，カテゴリタブ，タグクラウド，質問記事付き拡張クエリを段階的に提示することで，具体的な内容の拡張クエリを選択するための支援を行う．本システムの利用手順を図 3 に示す．図 3 はクエリ “京都” を入れた時のシステム探索の例である．本システムでは，以下の 3 つの手順によって拡張クエリが選択される．

(a)　タブの中から自分の興味に合致するカテゴリを選択する

(b)　選択したカテゴリのタグクラウドから自分の興味のあるキーワードを選択する

(c)　質問記事を見てキーワードの根拠を確認，より具体的なクエリで検索を行う

カテゴリ名は簡潔な表記になっているため，ユーザは興味の方向性をここで決定することができる．タグクラウドの中にはユーザにとって未知の語が出現することがある．例えば“トロッコ”は京都に馴染みのないユーザにとっては意味を把握できない可能性が高い．“トロッコ”を選択すると，質問記事付きの拡張クエリが表示される．ここで質問記事を読むことで，“京都の嵐山にはトロッコ列車がある”ということを知ることができ，“京都 トロッコ”という拡張クエリのコンテキストを理解することができる．質問記事にはより具体的なキーワード組による拡張クエリが提示させるため，質問記事の中に自身の情報要求と一致するものがあれば，質問記事のキーワードを使用したより具体的な Web 検索が可能になる．最後に拡張クエリによって検索された Web ページを閲覧し，興味のない話題だった場合はカテゴリやキーワードの選択を行う．この操作を繰り返すことにより，ユーザは序々に自身の興味のある話題が固まっていき，目的が定まった検索を行うことが可能になる．カテゴリタブや関連キーワードは自由に切り替える事が可能なため，ユーザは自身のコンテキストに近いカテゴリや質問記事を切り替えながら検索を行うことができる．

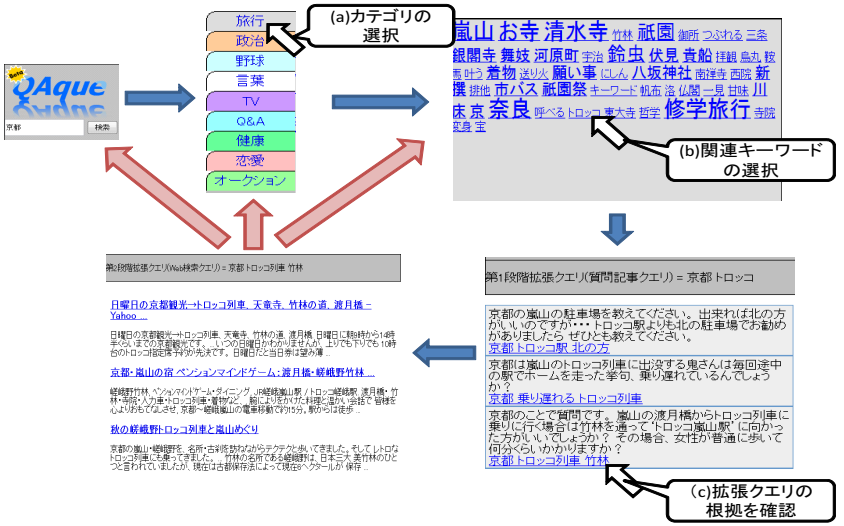


**図 3**　段階的な拡張を行う操作例

本システムではコミュニティQA のカテゴリの選択が非常に重要となる．カテゴリ別の関連キーワードの作成の他，ユーザが自身のコンテキストと拡張クエリを比較する最初の指針となるため，入力クエリに合ったカテゴリを提示する際には，なるべく幅広い内容かつ，入力クエリに関係の強いカテゴリを選択する必要がある．

### 3.2　カテゴリの選択手法

本節では，コミュニティQA のカテゴリから入力されたクエリに対して適切なカテゴリを選択する手法を提案する．コミュニティQA サイトの Yahoo！知恵袋では，数百のカテゴリが存在する．カテゴリ構造の例を図 4 に示す．大カテゴリ“インターネット、PC と家電”には，中カテゴリ“パソコン”，“デジタルカメラ”，“インターネット”などがあり，その下に“windows7”，“デジタル一眼レフ”，“SNS”などの小カテゴリが存在している．

本節では，ユーザのコンテキストに近いカテゴリを選択する手法を提案する．まず，複数の中カテゴリを集約し，大カテゴリを作成することで，カテゴリ構造を再構成する．次に，大カテゴリに潜在的ディリクレ配分法（LDA）を用いて潜在的トピックモデルを構成し，トピックモデルの確率分布から関連キーワードの検索を行う．

#### 3.2.1　大カテゴリの作成

コミュニティQA には数百のカテゴリが存在する．すべてのカテゴリで拡張クエリを作成し，ユーザに提示することは困難なため，内容の近いカテゴリを集約し，カテゴリを再構成する必要がある．本論文では，カテゴリの階層関係を利用する．大カテゴリの下位に存在する中カテゴリの質問記事を集約し，一つの大カテゴリの質問記事空間とする．大カテゴリには，5〜6 の中カテゴリから構成されるカテゴリもあれば，10 を超える中カテゴリから構成される巨大なカテゴリも存在する．カテゴリ数の違いは扱われる話題の広さを反映するものなので，カテゴリ数はできるだけ均一に揃える必要がある．そのため，コミュニティQA の大カテゴリを全て 5 から 8 程度の中カテゴリから構成されるように再構成する．例えば，“エンターテインメントと趣味”という大カテゴリは 13 カテゴリの中カテゴリから構成される．このカテゴリの“テレビ、ラジオ”，“音楽”，“映画”，“演劇ミュージカル、ダンス”，“芸能人”，“伝統文化、伝統芸能”をエンターテイメントカテゴリ，“ゲーム”，“おもちゃ、ホビー’，“絵画、手芸、工芸”，“懸賞、くじ”，“本、雑誌”，“アニメ、コミック”，“占い、超常現象”を趣味カテゴリに再構成する．カテゴリの再構成は，大カテゴリ内のカテゴリ内のみで行い，複数の大カテゴリを跨いだカテゴリの再構成は行わない．

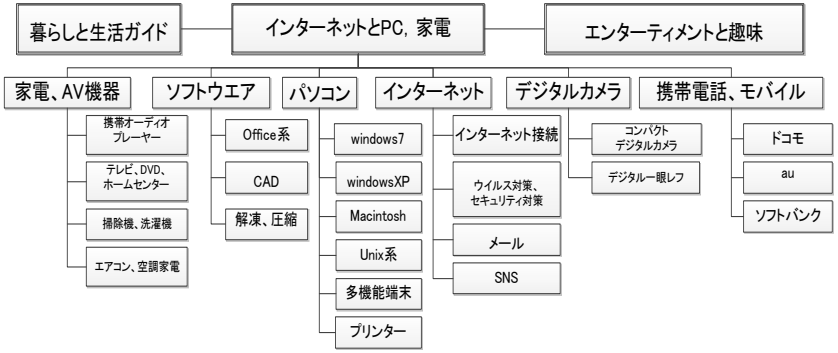


**図 4** コミュニティQA のカテゴリ構造

### 3.2.2 潜在的トピックモデルの作成

潜在的トピックモデルの作成には，潜在的ディリクレ配分法（LDA) を用いる．LDA とは，Blei ら[9] によって考案された，確率モデルを用いた潜在的トピック作成手法である．潜在的トピックモデルでは，文書を複数のトピックからの混合分布であると仮定する．各トピックは単語の確率分布によって表現される．

各文書はトピック分布 $\theta$ を持ち，単語はトピック $z$ を選択した後，そのトピック $z$ の単語分布 $\phi$ に従い，生成される．ディリクレ事前分布を Dir()，多項分布を Multi() とするとLDA のモデル生成過程は以下のようになる．

(1) 文書 $d$ について Dir（$\alpha$) から多項分布パラメータ $\theta_d$ をサンプリング
(2) トピック $t$ について Dir（$\beta$) から多項分布パラメータ $\phi_t$ をサンプリング
(3) 文書 $d$ に，$N_d$ 個の単語があったとき，$j$ 番目の単語 $w_{d,j}$ について
  (a) Multi($\theta_d$) から $z_{d,j}$ をサンプリング
  (b) Multi($\phi_{t,j}$) から $w_{d,j}$ をサンプリング

LDA では，教師なし学習によって，文書・単語空間からトピック集合 $Z$ を推定する必要がある．推定方法には，差分ベイズ推定法，ギブスサンプリングなどがある．本論文では，崩壊型ギブスサンプリングを用いる．崩壊型ギブスサンプリングを用いたとき，文書 $d$，$n$ 番目の単語 $w_{d,n} = v$ のトピック $z_i = k$ の更新式は以下の通り定義する．

$$P(z_i = k|Z_{-i}, W) = \frac{N^d_{k-i} + \alpha}{N^d_{-i} + T\alpha} \frac{N^v_{k-i} + \beta}{N_{k-i} + W\beta} \tag{1}$$

$-i$ は，トピック集合全体から $i$($d$ 番目の文書の $n$ 番目の単語) 分を除くことを示す．$N^d_k$ は，文書 $d$ において，トピック $k$ が割り当てられた回数，$N^d$ は文書 $d$ において単語が生成された回数，$N^v_k$ はトピック $k$ において単語 $v$ が出現する回数．$N_k$ は，トピック $k$ に出現する単語の総数である．$T$ はトピックの種類数，$W$ は単語の語彙数である．$\alpha$，$\beta$ はディリクレ分布のハイパーパラメータである．

Collased ギブスサンプリングによって得られたトピック分布 $Z$ から文書-トピック分布 $\theta$ と，トピック-単語分布 $\phi$ から文書 $d$ において，トピック $k$ が生成される確率 $\hat{\theta^k_d}$，トピック $k$ から単語 $w$ が生成される確率 $\hat{\phi^w_k}$ は以下の通りである．

$$\hat{\theta^k_d} = \frac{N^d_k + \alpha}{N^d + T\alpha} \tag{2}$$

$$\hat{\phi^w_k} = \frac{N^v_k + \beta}{N_k + W\beta} \tag{3}$$

### 3.2.3 トピックモデルによる関連キーワードの検索

LDA によって文書・単語の出現行列は，文書・トピック分布を表現する行列，トピック・単語分布を表現する行列に分解できる．これは，ベクトル空間モデルにおける潜在的意味インデキシング（LSI）の特異値分解と対応付けて考えることができる．図 5 に，LSI と LDA での行列分解の対応関係を示す．LSI では，文書・単語行列を潜在的な意味を持つ低次元のクラスに次元圧縮するのに対し，LDA では，隠れトピックの確率分布によって次元圧縮を行う．LSI では，単語間の距離をベクトルのコサイン距離によって計算する．コサイン距離が近い単語は，関連する意味を持つ．LDA においても各隠れトピックでの確率分布が似ている単語は，関連する意味を持つ．確率分布の類似度を計算するためには，KL ダイバージェンスを用いる．確率分布 $P$ と $Q$ の KL ダイバージェンスは以下の通り定義する．

$$KL(P||Q) = \sum_{x \in X} P(x) \log_2 \frac{P(x)}{Q(x)} \tag{4}$$

式 (4) より，$KL(P||Q) \neq KL(Q||P)$ であり，$P$ と $Q$ の交換法則が成立しない．本論文では，ユーザが入力した単語の確率分布を $P$ とし，関連キーワードの確率分布を $Q$ とする．

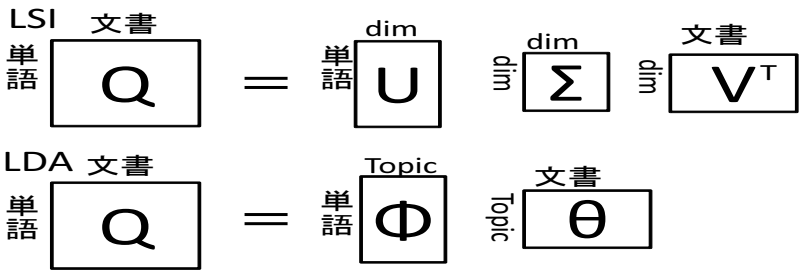


**図 5** LSI と LDA の行列分解の対応関係

## 4. 評　　価

本章では，提案したクエリ拡張法の評価を行う．4.1 節で，評価実験に使用したデータセットについて説明する，次に 4.2 節で，LDA と KL ダイバージェンスから検索された関連キーワードについて，カテゴリごとの出現確率の違いによる抽出した関連キーワードの傾向の違いを明らかにする．4.3 節で，推薦された関連キーワードと中カテゴリの関係を評価する．

### 4.1 評価実験のためのデータセット

本論文では，コミュニティQA のデータセットとして，国立情報学研究所が提供する Yahoo! 知恵袋コーパス第 2 弾[⋆1]を使用する．使用した質問記事は，2008 年の 1 年間で投稿された質問記事とする．実験に使用する大カテゴリとそれを構成する中カテゴリ，カテゴリの文書総数を表 1 に示す．大カテゴリとして “インターネット、PC と家電”，“エンターテイメントと趣味” カテゴリを分割し，再構成した “エンターテイメント”，“趣味” カテゴリを使用する．質問記事から MeCab[⋆2] を用いて，形態素解析を行い，単語を抽出した．

全質問記事で 1 回しか出現しない単語はその質問記事固有の表現である．トピックモデルでは，複数のトピックでの出現確率を比較するため，関連語検索にほとんど与えない．本手法では，複数の質問記事で使用されている単語のみを使用した．

### 4.2 カテゴリと出現確率の違いによる関連キーワードに関する評価

データセットに LDA を適用し，KL ダイバージェンスによって関連キーワードの検索を行う．本論文では，LDA のトピック数を 50，$\alpha$ を 50/(トピック数)，$\beta$ を 0.1 にハイパーパラメータを設定する．テストクエリとして Yahoo!JAPAN が提供している 2008 年の検索ワードランキング[⋆3] から，それぞれのカテゴリの内容に近いクエリである “mixi”，“嵐”，“ポケモン[⋆4]” を例として，関連キーワードの検索を行った．

**表 1** 評価実験のデータセットカテゴリ

| 大カテゴリ | 中カテゴリ | 質問記事数 | 単語数 |
|---|---|---|---|
| インターネット、PC と家電 | パソコン，デジタルカメラ，インターネット，ソフトウエア，家電・AV 機器，携帯電話・モバイル | 139,839 | 18,216 |
| エンターテイメント | 映画，音楽，芸能人，演劇・ミュージカル・ダンス，テレビ・ラジオ，伝統文化・伝統芸能 | 158,647 | 19,621 |
| 趣味 | アニメ，コミック，ゲーム，本・雑誌，おもちゃ・ホビー，占い・超常現象，絵画・手芸・工芸，懸賞・くじ | 104,160 | 18,438 |

提示される関連キーワードの例を表 2 に示す．また，クエリに対して各カテゴリでの出現確率でランキングしたものを表 3 に示す．カテゴリ名の横の括弧内の数値は出現確率である．それぞれのクエリに対して，出現確率が高かったカテゴリでは，クエリと関連の高い語が推薦されている．特に，出現確率が 1 位のカテゴリでは，“マイミク”，“ポイント”，“smap”，“news”，“捕まえる”，“lv” などの具体的なキーワードが提示されている．一方，出現確率の低いカテゴリでは，クエリと直接関係の近いキーワードではなく，そのカテゴリの一般的に使用されるキーワードが多く出現している．クエリ “ポケモン” では，趣味カテゴリとエンターテイメントカテゴリで，それぞれ “ゲームのポケモン” と “映画のポケモン” について異なる関連キーワードが推薦されている．エンターテイメントカテゴリでは趣味カテゴリほど具体的な内容ではないが，“シリーズ”，“アニメ”，“ストーリー” など “ポケモンの映画” に関連のあるキーワードが推薦されている．

### 4.3 関連キーワードと中カテゴリの関係性に関する評価

提案法では，複数の中カテゴリを統合し，大カテゴリとしている．そのため，本来中カテゴリによって分類されていた情報が，大カテゴリに再構成した際に，失われてしまう可能性がある．そこで，大カテゴリによって検索された関連キーワードがどの中カテゴリに関連があるか評価を行う．実験として，大カテゴリ “インターネット，PC と家電” カテゴリにお

⋆1 「Yahoo!知恵袋」データの提供について
http://www.nii.ac.jp/cscenter/idr/yahoo/tdc/chiebukuro.html
⋆2 MeCab http://mecab.sourceforge.net/
⋆3 2008 検索ワードランキング http://searchranking.yahoo.co.jp/ranking2008/
⋆4 ランキングでは “ポケットモンスター” であるが，作成したインデックスの関係から今回は “ポケモン” とした

**表 2**　関連キーワード例

| カテゴリ名 | mixi | 嵐 | ポケモン |
|---|---|---|---|
| インターネット, PC と家電 | 自分, 知る, 名前, 内容, 登録, url, メール, 何, 心配, 変, 人, 不安, 怖い, 書く, ゆう, 友達, 覚える, ポイント, 配信, 友人, 特定, 怪しい, アドレス, マイミク | ある, なる, する, 思う, ない, お願い, 普通, 使う, いい, 関係, 先日, つける, 今, つく, いう, 教える, 感じ, 違う, 心配, わかる, 全部, 一緒, 先, 低い | する, お願い, ある, 詳しい, なる, 教える, やる, 使う, わかる, 別, 大丈夫, 買う, パソコン, 作業, 前, 方法, 思う, 困る, ノート, 新た, 知識, 上記無知, 自体, 壊れる |
| エンターテイメント | 思う, ない, 多い, ファン, 人気, 日本, 存在, 日本語, 日本人, 評価, 韓国, なる, 海外, 最近, 一般, 応援, 国, アメリカ, 外国, 東方, 残念, 起, 一部, アイドル, 非常 | ジャ, ニーズ, ファン, グループ,news, ジャニ,smap, メンバー,hey, 関,jump,say,kat,tun, 大野, 中居, 赤, 松本, 西,jr, 山田, 亮, 人気, 葉 | 映画, 観る, 作品, 怖い, 面白い, シーン, シリーズ, アニメ, ホラー, 見る, 洋画, 公開, 邦画, 期待, 内容, 監督, 劇場, ポニョ, 感想, 映像, 最高, 苦手, ストーリー, 上映 |
| 趣味 | サイト, する, 書く, 送る, 携帯, かかる, メール, 情報, 方法, 来る, 利用, できる, 電話, 登録, 公式, 確認, 心配, 届く, 無料, 個人, 変える, ブログ, アドレス, 大丈夫,url | 教える, 嬉しい, お願い, 詳しい, いただける, 知る, あと, 頂ける, オススメ, わかる, うれしい, 幸い, ありがたい, いらっしゃる, 助かる,(?), ご存知, ご存じ, よい, 最近, ついで, おる, 分かる, ススメ | hp, 技, 努力, レベル,lv, 覚える, 性格, 攻撃, プラチナ, 育てる, 素早い, 捕まえる, バトル, パール, 進化, 持ち物, 252, 特攻, 特, 防御, 防, 100, ボール, ダイヤモンド |

**表 3**　カテゴリ確率順位

| | mixi | 嵐 | ポケモン |
|---|---|---|---|
| 1 | インターネット ($3.61 \times 10^{-3}$) | エンタメ ($7.31 \times 10^{-3}$) | 趣味 ($1.58 \times 10^{-2}$) |
| 2 | 趣味 ($1.53 \times 10^{-4}$) | 趣味 ($2.40 \times 10^{-4}$) | エンタメ ($2.77 \times 10^{-4}$) |
| 3 | エンタメ ($1.13 \times 10^{-4}$) | インターネット ($1.43 \times 10^{-4}$) | インターネット ($6.43 \times 10^{-5}$) |

いて，関連キーワードを検索し，各関連キーワードごとに，中カテゴリでの出現確率を比較する．最も出現確率の高いカテゴリを，その関連キーワードが所属する中カテゴリとする．50 個の関連キーワードのうち，どの中カテゴリにどのくらいのキーワードが所属しているのかを比較する．入力するクエリは，中カテゴリのカテゴリ名から “インターネット”，“パソコン”，“ソフトウエア”，“デジカメ”，“家電”，“携帯”，“psp” を使用する．

実験の結果を図 6 に示す．入力したクエリに関連する中カテゴリに関連するキーワードが推薦されている．特に，クエリ “デジカメ” では，92% がデジタルカメラカテゴリ，クエリ “携帯” では，88% が携帯電話・モバイルカテゴリに関連の強いキーワードが推薦されている．一方で，他のカテゴリの語も推薦されている．クエリ “インターネット” カテゴリで

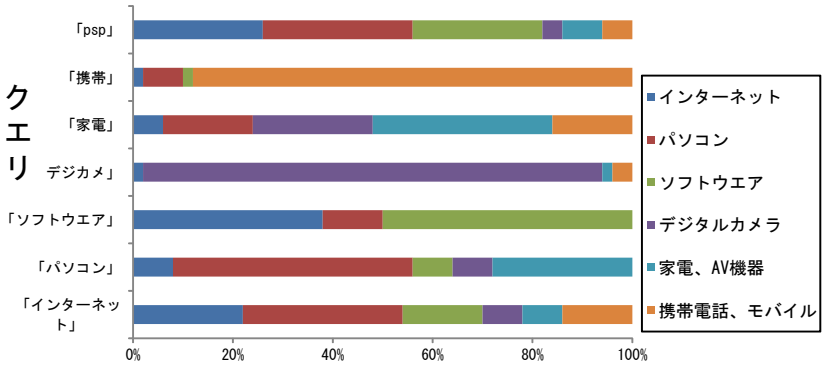


**図 6**　関連キーワードにおける中カテゴリの出現割合

**表 4**　カテゴリ確率順位

| | パソコン | デジカメ | psp |
|---|---|---|---|
| 1 | パソコン ($3.09 \times 10^{-1}$) | デジタルカメラ ($2.33 \times 10^{-1}$) | ソフトウエア ($4.92 \times 10^{-3}$) |
| 2 | インターネット ($9.69 \times 10^{-2}$) | パソコン ($9.81 \times 10^{-3}$) | インターネット ($3.75 \times 10^{-3}$) |
| 3 | ソフトウエア ($8.35 \times 10^{-2}$) | ソフトウエア ($4.10 \times 10^{-3}$) | パソコン ($3.31 \times 10^{-3}$) |
| 4 | デジタルカメラ ($7.86 \times 10^{-2}$) | 家電 ($3.30 \times 10^{-3}$) | 家電 ($2.57 \times 10^{-3}$) |
| 5 | 家電 ($4.81 \times 10^{-2}$) | 携帯電話 ($2.60 \times 10^{-3}$) | デジタルカメラ ($1.38 \times 10^{-3}$) |
| 6 | 携帯電話 ($3.00 \times 10^{-2}$) | インターネット ($6.08 \times 10^{-4}$) | 携帯電話 ($8.97 \times 10^{-4}$) |

は，インターネットカテゴリとほぼ同数のパソコンカテゴリの関連語が推薦されている．他のクエリにおいても自身が所属するカテゴリ以外の中カテゴリからも多くの関連語が推薦されていることがわかる．表 4 に，クエリ “パソコン”，“デジカメ”，“psp” の中カテゴリの出現確率のランキングを示す．図 6 と表 4 を比較すると，カテゴリの出現確率の順位と推薦される関連キーワードの割合がほぼ一致している．“パソコン” の出現確率は，パソコンカテゴリが最も高く，他のカテゴリと比較しても出現確率が 10 倍程度高い値になっている．推薦される関連キーワードの割合もほぼ同様の分布になっており，パソコンカテゴリに関連するキーワードが 5 割近く推薦され，他のカテゴリが同程度の割合で推薦されている．“デジカメ” では，この傾向がさらに強くなっている．“psp” では，出現確率が高いソフトウエア，インターネット，パソコンカテゴリの関連キーワードがほぼ同数推薦されている．

## 5. 考　　察

本論文では，複数のカテゴリを投稿し，LDA により関連語の検索を行った．同一のクエリを複数のカテゴリで関連語検索を行い，推薦される関連キーワードの比較した．推薦される関連キーワードは，出現確率が高いカテゴリほどより具体的なキーワードが推薦される．LDA では，文書中の各単語がトピックに割り当てられる．出現する回数が少ない単語は，トピックを割り当て割れる回数が少ないため，特徴的な関連語を推薦するための確率分布を形成することができなかったためであると考えられる．これは，クエリとカテゴリの距離が十分に遠いといえる．クエリに対して，距離が遠く関連の低いカテゴリはユーザに提示する必要がないため，システムに実装するためには，クエリに対する各カテゴリの出現確率を比較し，出現確率が高いカテゴリのみを選択して，関連語検索を行うことが必要になると考えられる．クエリ “ポケモン” のように同一の話題でもゲームに関する内容，と映画に関する内容で異なる関連キーワードが推薦されることから，ユーザに複数のカテゴリを提示することで，より多様な話題についてキーワードを提示できるようになると考えられる．

複数の中カテゴリを統合し，関連語検索を行ったが，検索される関連語は，中カテゴリの話題に対応することができるといえる．出現確率が高いカテゴリほど多くの関連キーワードが推薦される．これは，大カテゴリにおける出現確率によって提示するカテゴリを選択する操作に対応できると考えられる．大カテゴリでは，タブ切り替え等の操作によってユーザが興味のあるカテゴリを選択するが，中カテゴリでは，LDA によって関連のあるカテゴリが自動的に割り当てられ，それに応じた数の関連語が各中カテゴリから推薦されていると考えることができる．本論文では，中カテゴリは一つの質問記事空間に統合されているが，潜在的には，LDA によって関連する中カテゴリの選択が行われているといえる．

## 6. 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ために拡張のコンテキストを提示するクエリ拡張システムについて提案した．特に，コミュニティQA のカテゴリを利用し，関連するカテゴリから拡張のための関連するキーワードを検索するための手法を提案した．

コミュニティQA の階層構造を利用し，複数の中カテゴリの集合から，大カテゴリを作成することにより，数百のコミュニティQA のカテゴリ全体からクエリに関連するカテゴリを検索することが可能であることを明らかにした．LDA によるトピックモデルの確率分布により関連語検索を行うことにより，カテゴリの話題に関連のあるキーワードの推薦した．

今後は，本論文で提案したカテゴリ選択法をコミュニティQA 全体で実装を行い，評価を行う予定である．また，LDA において “する” 等の一般語が多く出現していたことからより具体性のあるキーワードの抽出法について検討を行う予定である．

**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) 大塚淳史, 関洋平, 神門典子, 佐藤哲司. 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. 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), pp. F6–3, 2011.
2) 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.
3) 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.
4) 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.
5) Jiafeng Guo, Xueqi Cheng, and GuXu. A Structured Approach to Query Recommendation with Social Annotation Data. *CIKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010.
6) 廣嶋伸章, 戸田浩之, 松浦由美子, 片岡良治. 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. 情報処理学会論文誌. データベース, Vol.3, No.3, pp. 33–45, 2010.
7) Sounwood Yoon, Adam Jatowt, and Katsumi Tanaka. Intent-Based Categorization of Search Results Using Questions from Web Q&A Corpus. *Proceedings of the 10th international conference on Web Information Systems Engineering (WISE2009)*, pp. 145–158, 2009. LNCS 5802/2009.
8) 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, pp. 2A–2, 2010.
9) DavidM. Blei, AndrewY. Ng, and MichaelI. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, Vol.3, pp. 993–1022, 2003.