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

大塚　淳史† 　関　 洋平†† 　神門　典子††† 　佐藤　哲司††

† 筑波大学情報学群知識情報・図書館学類　〒305–8550 茨城県つくば市春日 1–2
†† 筑波大学大学院図書館情報メディア研究科　〒305–8550 茨城県つくば市春日 1–2
††† 国立情報学研究所　〒101–8430 東京都千代田区一ツ橋 2–1–2
E-mail: †s0913153@klis.tsukuba.ac.jp, ††{yohei,satoh}@slis.tsukuba.ac.jp, †††kando@nii.ac.jp

あらまし　コミュニティQA サイトに投稿される質問記事は，ユーザの疑問や知りたいことを自然言語で記述したものである．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を複数のカテゴリから抽出することで，検索者の多様な情報要求を満たす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，自身の情報要求を言語化された検索クエリとして確認することができる．大量の質問記事を潜在的意味解析することで，キーワードから多様な質問記事を抽出できることを確認したので報告する．

キーワード　Web 検索，クエリ拡張，コミュニティQA，情報要求，潜在的意味解析

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

Atsushi OTSUKA†, Yohei SEKI††, Noriko KANDO†††, and Tetsuji SATOH††

† College of Knowledge and Library Sciences, School of Informatics University of Tsukuba
1–2,Kasuga,Tsukuba,Ibaraki,305–8550 Japan
†† Graduate School of Library and Information Science and Media Studies,University of Tsukuba
1–2,Kasuga,Tsukuba,Ibaraki,305–8550 Japan
††† National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101–8430, Japan
E-mail: †s0913153@klis.tsukuba.ac.jp, ††{yohei,satoh}@slis.tsukuba.ac.jp, †††kando@nii.ac.jp

Abstract　Question articles posted to the QA community are expressed question in natural language. In This paper, we make query expansion system to support Web search users to generate web query by using QA community resources. This system suggests search keywords and question articles from QA community's category. Users can find their information needs as verbalizing query by browsing question articles. And, we extract diversified question articles from query with latent semantic indexing.

Key words　Web Search，Query Expantion，Commmunity QA，Information Needs，Latent Semantic Indexing

1. はじめに

Web コンテンツの増加により，膨大な Web ページの中から所望のページを探しだす Web 検索エンジンの必要性はますます高まっている．検索エンジンを利用する際，ユーザは自らが知りたいことである情報要求を頭の中で言語化し，検索エンジンの入力となるクエリを作成しなければならない．しかし，ユーザは必ずしも検索に対して適切なクエリを作成できるとは限らない．具体的で明確な言語化が困難な場合，ユーザは一般的な語で検索をせざるを得ない．一般的な語での検索は膨大な数の検索結果となるため，ユーザは満足のいく Web ページを簡単には見つけ出すことができない．

この問題の解決には，検索エンジン側がユーザの情報要求を推定することが有効であるが，検索エンジン側が入手できる情報はクエリだけなので，ユーザの情報要求を一意に特定することは困難である．そこで，ユーザの情報要求を特定するのではなく，検索の候補をユーザに提示することで，検索の支援を行う研究が盛んに行われている．

代表的な検索支援法として，クエリ拡張が挙げられる．Web 検索エンジンにおけるクエリ拡張とは，入力クエリに関連する

複数のクエリを提示し，ユーザに選択させる手法である．拡張クエリは，クエリ内のキーワードの置き換えやクエリに新たなキーワードの追加などによって作成される．ユーザは，提示された拡張クエリの候補中から自分の情報要求に近いと思うクエリを選択する．しかし，キーワード組の拡張クエリでは，拡張されたキーワードだけがユーザに提示されるので，拡張がどのようなコンテキストで行われたのかを把握できない，あるいは，拡張されたキーワードを知らなければ選択できないという課題がある．

本研究では，ユーザがコンテキストを理解できる拡張クエリを作成することを目的として，Yahoo! 知恵袋(注1)や，教えて!goo(注2)に代表される，質問回答サイト（QA サイト）の質問記事に着目した，拡張クエリを作成する手法を提案する．自然言語で記述された質問記事本文も含めてユーザに提示することで，ユーザは拡張されたクエリのキーワードがどのようなコンテキストで使用されているのかを直接読み取ることができる．この結果，自身が期待した情報要求に適合しているかの判断が容易になると考える．質問記事は，ユーザとの親和性が高いフォーマットであり，キーワード組のクエリは，検索エンジンの入力として一般的に使用されている．質問記事とクエリ両方を提示することは，これまでユーザ自身が行っていた “情報要求を言語化しクエリを作成する” という一連のプロセスをシステム側が再現することに相当する．ユーザは拡張クエリのコンテキストを理解した上で拡張されたクエリによって適切な検索を行うことが可能になると考えられる．

QA サイトのカテゴリ分類を利用することにより，多様なクエリの拡張を行う．作成された拡張クエリから検索された Web ページのキーフレーズ抽出実験を行い，推薦された拡張クエリからより多様な Web ページが収集できているか評価を行う．

本論文の構成は以下の通りである．2 章で関連研究について述べる．3 章で本研究で提案する拡張クエリについて述べ，4 章で拡張クエリの作成手法を説明する．5 章では提案法を実装したシステムの結果について述べ，6 章で評価実験について説明する．7，8 章で考察とまとめを行う．

2. 関 連 研 究

本研究は，より柔軟な，より高次なクエリ拡張技術と位置づけられる．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]．現在では，Web 上から大量の情報を入手できるようになり，Web ページの情報から拡張クエリを作成する研究も多く行われている．Yin ら [2] は，情報検索システムの精度向上には，システム内文書での適合性フィードバックを行うよりも，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方が効果的であることを示した．

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 [3] は，Web 百科事典 Wikipedia から作成した拡張クエリと，Web 検索結果の疑似フィードバックから作成した拡張クエリとを，ユーザ実験によって比較している．その結果，Wikipedia から作成した拡張クエリの方が疑似フィードバックよりもユーザ満足度が高くなることを示した．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．水野ら [4] はこの特徴を利用し，ユーザの特徴や趣向を反映させたクエリ拡張を行っている．水野らは，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ァイルを情報源とすることで，ユーザの趣向にあった Web ページを検索するための拡張クエリを作成できるとしている．

質問記事には，ユーザの疑問や要求がテキストとして表現されてるため，質問文と Web ページとの検索結果を組み合わせることでユーザの要求と合致した検索結果をユーザに提示できるといえる．QA サイトを用いて検索意図の候補を提示する研究には山本ら [5] がある．山本らは，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．高田ら [6] は QA サイトの質問に対する回答の他に Web ページから別解情報を検索しすることで，Web と QA の相互補完を行っている．

多義的なクエリや Web ページを推薦する手法については，今井ら [7] の研究や Yoon ら [8] 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実証した．Yoon らはユーザの要求を QA サイトのカテゴリに反映させ，QA のカテゴリ分類に対応して Web ページの分類を行うことで，ユーザの検索意図に応じて幅広く Web ページを推薦している．

本論文では，質問記事を “ユーザの情報要求を直接的に表現したもの”であるものとして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す．

3. 質問記事を用いた拡張クエリの提案

本研究で作成する拡張クエリの例を図 1 に示す．本研究では，“話題の多様性”と “要求の詳細さ”という 2 つの視点から，情報要求の候補を提示する手法を提案する．一般に，ユーザの情報要求は，ユーザが検索を行う度に異なったものになる．このため，情報要求の候補を提示する際には，多様な話題を提示することが有効と考える．また，ユーザの “知りたいこと”の詳細さも異なる．具体的なことを知りたい場合と幅広く情報を集めたい場合では，作成するクエリは異なる．

以下の 2 つの視点からの拡張を行い拡張クエリを作成する．

- 情報要求の多様性を展開する拡張
- 情報要求を多段階に展開する拡張

多様性という “広さ”と，多段階という “深さ”を持たせること

(注1)：http://chiebukuro.yahoo.co.jp/
(注2)：http://oshiete.goo.ne.jp/

で，提案する拡張クエリは，従来の拡張クエリとは形式が異なり，より特徴的なクエリとして拡張され，ユーザに提示される．

3.1 情報要求の多様性を展開する拡張

入力されるクエリが同じであっても，その背後にある情報要求は全く異なる場合がある．最も典型的な例は，多義語である．“ウイルス”という語は，病原体の他に，コンピュータウイルスを指す場合もある．このような語がクエリとして入力された場合，入力された語のみで，ユーザの検索意図を反映したクエリ拡張を行うことは難しい．多義語でなくても，ユーザの立場やそのときの状況によって情報要求が異なる場合は少なくない．情報要求の候補としての拡張クエリを提示するには，より幅広い観点から，多様性を持つクエリを作成する必要がある．

本研究では，QA サイトのカテゴリに注目する．QA サイトでは，質問者が質問をする際，質問に合ったカテゴリに記事を投稿している．カテゴリの存在によって，完全でないとしても，質問は体系的に分類された状態になっていると考えることができる．カテゴリ別に拡張クエリを作成し，最後にそれらを統合することで，幅広い観点からクエリを拡張できると考える．

3.2 情報要求を多段階に展開する拡張

ユーザが情報検索を行う際，情報要求の曖昧さに応じて，段階的にクエリを作成する．情報要求が明確であり，具体的な情報を入手したい場合は，クエリのキーワード数を増やして検索を行う．逆に，情報要求が曖昧な段階では，少ないキーワードで幅広く情報を入手しようとする．本研究では，この点に着目し，拡張クエリを段階的に具体化していくことで，多段階の展開を行う．図 1 の一段階目の拡張クエリでは，入力クエリに関連のあるカテゴリと，カテゴリから検索された関連語を一語追加した拡張クエリをユーザに提示する．二段階目の拡張では，一段階目の拡張クエリで提示したカテゴリ内から，質問記事を検索する．検索された質問記事からキーワードを抽出し，拡張クエリを作成する．

一段階目の拡張は，ユーザに検索の方向性を示すことを狙いとしている．ユーザはカテゴリ名を見ることで疑問の方向性を把握することができる．図 1 の例では，“海外旅行”は，旅行カテゴリに関係する質問の他に，PC や語学に関する質問もあることがわかる．そして，追加されたキーワード組で Web ページを検索することで，より具体的な情報要求を思いつくことができるようになると考える．

二段階目の拡張は，より具体的に検索を絞り込むための拡張である．質問記事本文が記述されているため，その中で自分の知りたい疑問があれば，そこから作成したキーワード組で Web 検索することで具体的な検索が可能になる．

段階的に情報を増やしたクエリを拡張していくことで，ユーザは混乱することなく，徐々に具体的な検索が行えるようになると考えている．

4. 拡張クエリの作成法

提案法を実現するシステムの処理の構成を図 2 に示す．まず，QA サイトのカテゴリの中から，入力したクエリと関連度の順に高いカテゴリのランキングを作成する．次に，各カテゴリごとに入力クエリの関連語の検索を行う．入力クエリと関連語から第一段階拡張クエリを作成する．第一段階拡張クエリを用いて質問記事を検索し，キーワードを抽出することにより，第二段階拡張クエリを作成する．最後に，カテゴリごとの結果を統合し，関連度の高いカテゴリ順に，拡張クエリをユーザに提示する．また，拡張クエリとは別に検索エンジン API によって Web ページの検索結果を取得し，拡張クエリともに提示する．

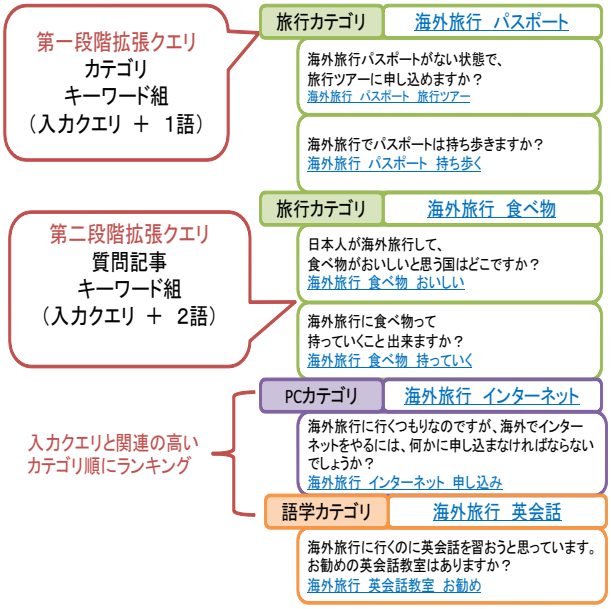

図 1 拡張クエリの構成

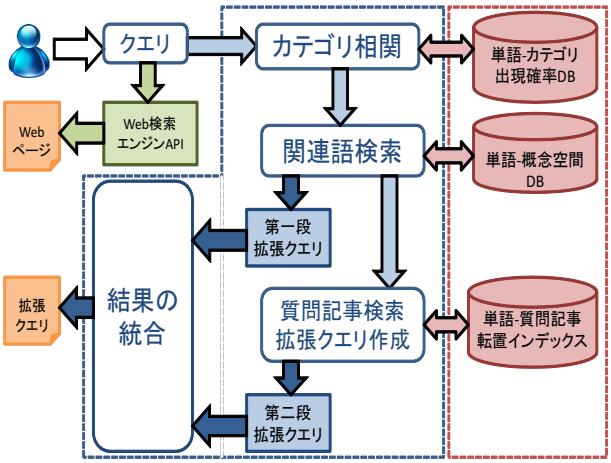

図 2 提案法の構成

本システムは以下の主要なブロックから構成される．

（1） クエリ拡張のための情報源となる質問記事セット

（2） 多様なクエリの作成法

（3） 多段階なクエリ作成法

4.1 拡張のための情報源となる質問記事セット

クエリ拡張の情報源として，国立情報学研究所提供の Yahoo! 知恵袋コーパス(注3)を使用する．今回は，投稿質問数の上位 10 カテゴリをデータセットとし，文書数が 20,000 になるようにランダムサンプリングを行った．取得した文書を形態素解析器

(注3)：http://research.nii.ac.jp/tdc/chiebukuro.html/

表 1　使用カテゴリとキーワード数

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

MeCab(注4)により形態素解析を行い，キーワード抽出を行う．抽出した形態素のうち，以下条件に当てはまる語を検索で用いるキーワードとする．

- 動詞，形容詞，名詞（非自立，接尾，代名詞を除く）
- 二文字以上で構成されている（漢字は一文字でも可）
- MySQL のストップワードリスト(注5)に入っていない
- 3 文書以上に含まれる

使用した Yahoo!知恵袋のカテゴリと抽出したキーワード数は表 1 となる．カテゴリ名は括弧内のローマ字表記を使用する．

4.2　多様なクエリの作成法

多様なクエリのための話題の単位に，Yahoo!知恵袋のカテゴリを用いる．カテゴリ別で拡張クエリを作成し，結果をまとめることで，ユーザに多様な話題からの拡張クエリを提供する．

入力語とカテゴリの関連度の指標として出現率を用いる．出現率とは，サンプリングしたカテゴリ内の全質問記事に対して，入力語を含む文書が何件あるのかを示した割合である．カテゴリ C での入力語 t の出現確率 $P_{C,t}$ は

$$P_{C,t} = \frac{\sum(\text{単語 } t \text{ が出現した文書})}{\text{カテゴリ内の全文書}} \quad (1)$$

関連度によってカテゴリの順位付けを行う．この順位は，拡張クエリをユーザに表示する際の表示順位に使用する．

4.3　多段階なクエリ作成法

一段階目のクエリ拡張では，クエリに関連するカテゴリと，キーワードを一語追加した拡張クエリを提示する．追加するキーワードの検索には，潜在的意味解析（LSI:Latent Semantic Indexting）を用いる．各カテゴリで投稿される質問記事の内容は異なっているため，LSI によって作られる概念空間もカテゴリによって違いが出るといえる．概念空間上で関連語を検索することによって，カテゴリの特徴が現れた関連語が抽出できると考える．

第二段階拡張クエリでは，質問記事本文と，キーワードを二語追加した拡張クエリを提示する．質問記事の検索は転置インデックスを用いることで，高速に検索を行う．キーワード組の作成は複合名詞を用いる．

第一段階目の拡張クエリの作成

データセットの質問記事は 20,000 記事のため，20,000 次元の文書ベクトルを持つ文書・単語行列が作られることになる．文書・単語行列の各要素には，各単語の *tf.idf* 値が入る．*tf* は質問記事に出現する単語数により正規化を行う．正規化を行うことで単語の種類数が少ない簡潔な文書中の単語ほど重視される．文書 D_j で索引語 t_i のスコア $d_{i,j}$ は，

$$d_{i,j} = \frac{\log_2(f_{i,j}+1)}{\log_2(\text{文書 } j \text{ 中の単語の種類数})} \cdot (\log_2 \frac{n}{n_i} + 1) \quad (2)$$

LSI を行うため，20,000 次元の文書ベクトルを 100 次元のクラスに次元圧縮する．次元圧縮には特異値分解（SVD:Singular Value Decomposition) を用いる．特異値分解により $m \times n$ の行列 $\mathbf{D}$ は，以下のように分解できる．

$$\mathbf{D}_{m\times n} = \mathbf{U}_{m\times r}\mathbf{\Sigma}_{r\times r}\mathbf{V}^T_{n\times r} \quad (3)$$

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計算し，類似度が高い順に第一段階拡張クエリに追加するキーワードとする．使用したカテゴリと関連語を第一段拡張クエリとして，ユーザに提示する．

第二段階目の拡張クエリの作成

第二段階目の拡張クエリの元となる質問記事の検索には，転置インデックスを用いる．検索クエリは，第一段拡張クエリを使用する．まず，第一段拡張クエリに含まれるキーワードで AND 検索を行う．次に，AND 検索でヒットした質問記事の中でスコアを計算する．スコアは，質問記事での各キーワードの *tf.idf* 値の総和となる．

第二段階目拡張クエリの作成手順の概要を図 3 に示す．質問記事を形態素解析し，名詞が連続している箇所を接続することで，連結名詞を作成する．形態素と連結名詞のリストを第一段拡張クエリのキーワードと比較し，キーワードを含む連結名詞がある場合，クエリのキーワードを連結名詞に置き換える．これをクエリの全てのキーワードで行う．最後に，転置インデックスを参照し，拡張クエリのキーワード以外で最もスコアの高いキーワードを抽出する．名詞の場合は他のキーワードと同様に連結名詞化し，拡張クエリのキーワード組に追加する．

使用した質問記事と作成した，キーワード組をまとめて第二段階拡張クエリとして提示する．

5.　拡張クエリの実装

Yahoo!知恵袋の質問記事を情報源に，多様性と多段階性を持つ拡張クエリの生成を行う．提案法を実装したシステムのスクリーンショットを図 4 に示す．図 4 の左側は拡張クエリ部，右側は Web ページ検索部である．ユーザが，検索窓に基本となるクエリを入力すると，入力されたクエリから生成した拡張クエリが拡張クエリ部に表示される．また，クエリに対する Web ページ検索の結果を Web ページ検索部に表示される．ユーザは第一段階目の拡張クエリである，カテゴリとキーワード組から興味のある項目を選ぶ．第一段階目の拡張クエリをクリックすることで，第二段階目の拡張クエリである質問記事と更に関連語を追加したキーワード組が表示され，クリックでクエリを

(注4)：http://mecab.sourceforge.net/
(注5)：http://dev.mysql.com/doc/refman/5.1/ja/fulltext-stopwords.html

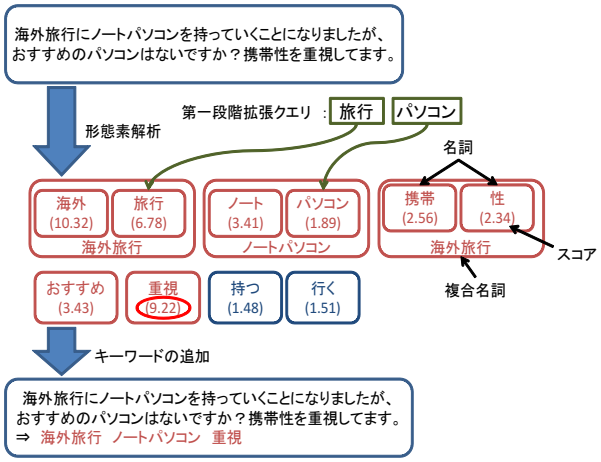

図 3　二段階目キーワード作成

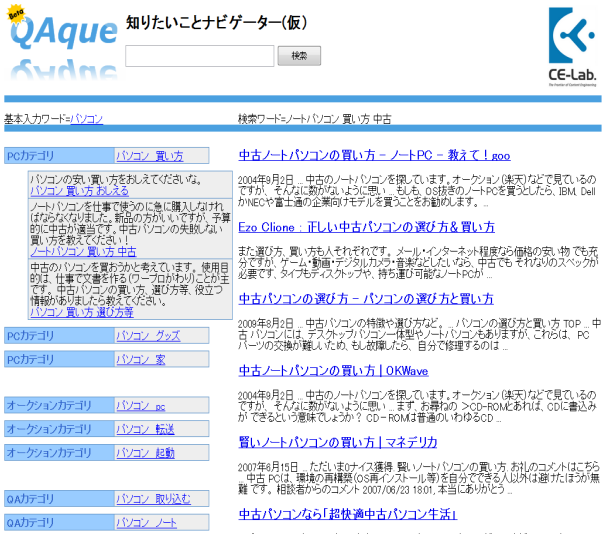

図 4　システム画面

切り替えることができる．拡張クエリ部からキーワード組を選択すると，Web ページ検索部の結果が，拡張クエリ部で選択されたキーワード組での Web 検索結果に切り替わる．ユーザは情報要求に応じて拡張クエリを選択し，Web ページ検索部の結果を交互に見ながら検索を進めることで，ユーザは自らの目的に合致した Web ページを見つけることができる．

5.1　カテゴリによる多様性の実装

Yahoo!知恵袋のカテゴリを用いて多様性を持たせたクエリの拡張を行う．検索の入力語に関連のあるカテゴリを見つけるために，単語の各カテゴリごとの出現率を計算した．

入力語を “旅行” と “ウイルス” としたときのとカテゴリの出現率を順位付けしたものを表 2 に示す．入力語の違いによって出現率のランキングが変化している．“ウイルス” は PC カテゴリが最も出現率が高くなっているが，健康カテゴリでも PC カテゴリに近い出現率となっている．また，“ウイルス” はオークションや知恵袋カテゴリでも出現しているが，全てのカテゴリに出現するわけではないことがわかる．

5.2　多段階による拡張クエリの実装

5.1 で関連があると判定されたカテゴリから，段階的に情報を増やした多段階の拡張クエリを生成する．一段階目で，拡張クエリに追加するキーワードとなる関連語の検索を行い，二段階目で，質問記事を検索し，質問記事からキーワードの生成を行う．語 “ウイルス” を入力したときの，拡張クエリの生成結果の一部を表 3 に示す．一段階目の拡張において，PC カテゴリでは，“汚染”,“検査”．健康カテゴリでは “ノロ”，“感染” と，それぞれ異なる語が追加されている．一方で，オークションカテゴリでも “感染” という語が追加されている．しかし，第一段階の拡張クエリから検索された質問記事は，健康カテゴリとオークションカテゴリでは，全く違う話題であり，そこから生成したキーワード組も異なる事がわかる．

表 2　入力語による出現率のランキング結果

順位	入力語 (出現率)	
	旅行	ウイルス
1	travel(0.102)	pc(0.0252)
2	renai(0.00865)	health(0.00240)
3	seiji(0.00450)	auction(0.00180)
4	health(0.00410)	bukuro(0.00150)
5	kotoba(0.00350)	
6	bukuro(0.00270)	
7	tv(0.002050)	
8	auction(0.00200)	
9	pc(0.000800)	
10	baseball(0.000350)	

6.　評 価 実 験

本研究では，QA サイトのカテゴリ分類を用いた多様性を考慮した拡張クエリの実装を実装した．多様性を持つ拡張クエリが推薦された場合，検索される Web ページもより多様なものになると考えられる．Web ページからキーフレーズを抽出し，抽出できたキーフレーズの数によって，推薦した拡張クエリが多様性を持つのかを検証する．

6.1　Web ページからのキーフレーズ抽出実験

本研究の評価実験として，Web ページからのキーフレーズ抽出実験を行う．検索された Web ページのタイトルとスニペットからキーフレーズを抽出する．キーフレーズはそのページを代表する語であるため，同じキーフレーズを持つページは同じ話題を扱っていることになる．

本実験でのキーフレーズ抽出は以下の手順で行う．

(1)　作成した拡張クエリを Web 検索 API[注6] 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一件ずつタイトルとスニペットを抽出
(3)　タイトルとスニペットからキーフレーズ抽出 API[注7] により，キーフレーズを抽出
(4)　キーフレーズとともに付与されているスコアが閾値以上のものをキーフレーズとする
(5)　Web ページ 10 件でキーフレーズ抽出を行う
(6)　キーフレーズは 2 回以上出現したものをカウント

表 3 “ウイルス”での拡張クエリ作成結果

第一段階拡張クエリ		第二段階拡張クエリ	
カテゴリ	キーワード	質問記事	キーワード
pc	ウイルス　汚染	ウイルスに汚染されていてもリカバーしたらウイルスなくなるんでしょうか？	ウイルス 汚染 ない
	ウイルス　検査	ウイルス検査ができるサイトを教えて下さい。駆除じゃなくて検査です。	ウイルス検査 駆除 サイト
health	ウイルス　ノロ	ノロウイルスってなんですか？	ノロウイルス
		成人のウイルス性胃腸炎の原因のウイルスで１番多いのはノロウイルスですか	ウイルス性胃腸炎 ノロウイルス 成人
	ウイルス　感染	何故、細菌やウイルスによる胃腸炎も食中毒に分類されるのですか？細菌やウイルスは他人から感染するというケースも多いのに	ウイルス 感染 細菌
auction	ウイルス　感染	ここに、あるオークションが紹介されている。 それをアクセスして見る。 すると、何かのウイルスに感染したり、個人情報（ID など）がどこかにわかってしまう。 そういうことって、ありますか？	ウイルス　感染　紹介
	ウイルス　受信	出品者です。私は、yahoo!のウイルスチェックに入っているのですが、落札者様からのメールが、迷惑メールのフォルダで受信されます。落札者様からのメールに添付ファイルなどはありません。これは、先方がウイルスに侵されているのでしょうか？　開いてメールを読んだ私の方は、大丈夫でしょうか？また、先方にお知らせした方がいいでしょうか？	ウイルスチェック 受信 先方

API によって付与されるスコアは 0〜100 に設定されている．今回は 50 を閾値とした．キーフレーズは 2 回以上出現してからカウントをするのは，Web ページのタイトルや日付など，ページの固有表現を除去するためである．

1 つの入力語から拡張クエリを 30 個作成し，それを 1 セットとする．異なる手順で作成された 3 セットの拡張クエリを比較する．入力語は各カテゴリを代表する語を一語ずつ，計 10 語を用いる．使用した入力語とその語に代表されるカテゴリは表 4 に示す．これらの語は，全てのカテゴリで出現するため，10 カテゴリすべてで拡張クエリが推薦される．比較対象となるクエリの作成手法は以下の 3 パターンである．

category　各カテゴリから作成された第一段階目拡張クエリ上位 30 個

total　各カテゴリの第一段階目拡張クエリ上位 3 個を 10 カテゴリまとめた拡張クエリ 30 個

base(yahoo)　Yahoo!関連検索ワード API(注8)によって推薦された拡張クエリ 30 個

例として，入力語 “ソフト”のデータセットの一部を表 5 に示す．category と total については入力語と追加語の AND 検索型のクエリである．base(yahoo) は複合名詞の一部となっているクエリがある．今回の実験では，対象となる Web ページが上位 10 件のみで検索結果数に影響しない点，ベースラインとして現状の検索エンジンのクエリ推薦と比較するという点からそのまま使用する．

(注6)：http://developer.yahoo.co.jp/webapi/search/

(注7)：http://developer.yahoo.co.jp/webapi/jlp/keyphrase/v1/extract.html

(注8)：http://developer.yahoo.co.jp/webapi/search/assistsearch/v1/webunitsearch.html

表 4　入力語とそのカテゴリ

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 5　入力語 “ソフト” の拡張クエリ

category(pc)	total	base(yahoo)
ソフト フリー	ソフト フリー	ソフトバンク
ソフト 読み上げる	ソフト 読み上げる	DS ソフト
ソフト シェア	ソフト シェア	マイクロソフト
ソフト 割れる	ソフト ゲーム	wii ソフト
ソフト ベクター	ソフト os	フリーソフト
ソフト 最強	ソフト 使う	PSP ソフト
ソフト 杜	ソフト ハード	PS3 ソフト
ソフト お勧め	ソフト 使い捨て	ソフトバンクホークス
ソフト 会計	ソフト コンタクトレンズ	解凍ソフト

6.2　キーフレーズ数抽出実験結果

全入力語に対して，total と base(yahoo) の 30 個のクエリでの合計キーフレーズ抽出数をまとめたものを図 5 に示す．横軸には各入力語，縦軸は拡張クエリ 30 個から抽出できたキーフレーズの合計数となっている．入力語”ソフト”以外の全ての入力語に対して，提案法である total が上回っている．また，半数の入力語において，total と base(yahoo) とのキーフレーズ抽出数の差が 100 以上になっている．

次に，最も total と base(yahoo) の差が大きくなった入力語 “日本語”と base(yahoo) が total を上回った入力語 “ソフト”について，category も含めた結果を示す．図 6 が “日本語”，図 7 が “ソフト”での結果である．横軸には各カテゴリ名，縦軸は合計のキーフレーズ数である．横軸の total，base(yahoo) はカテゴリ名でなく，データセット名である．“日本語”では，ど

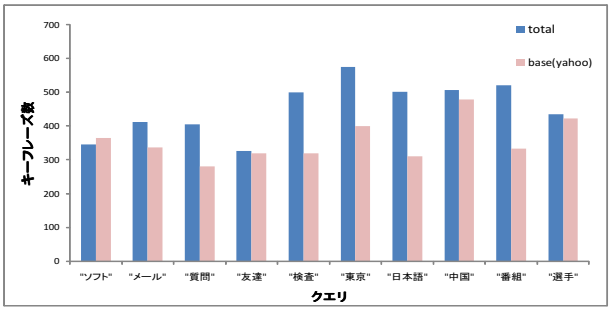

図 5　total(提案法）と base(yahoo) のキーフレーズ抽出数

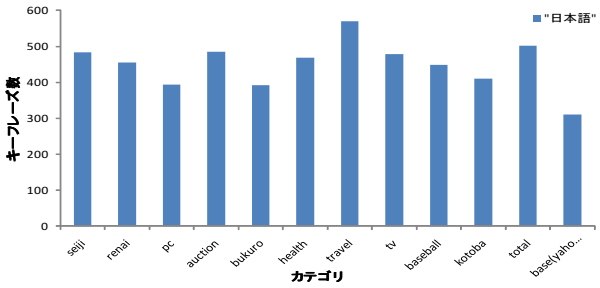

図 6　“日本語”でのキーフレーズ抽出数

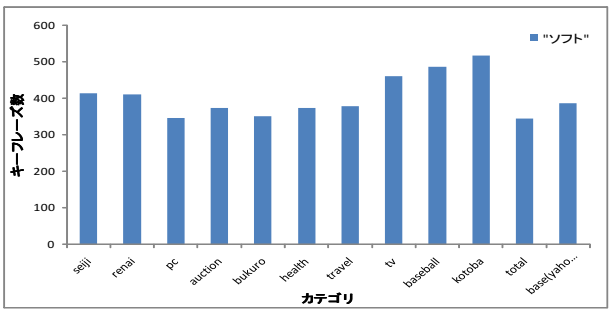

図 7　“ソフト”でのキーフレーズ抽出数

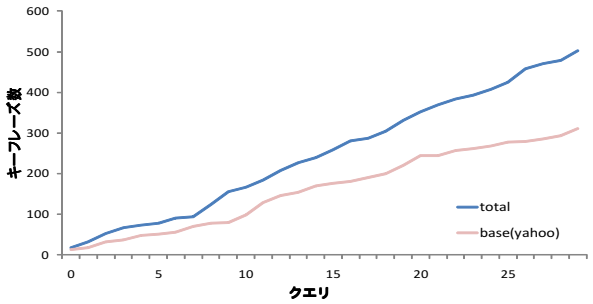

図 8　“日本語”でのキーフレーズ抽出数の推移

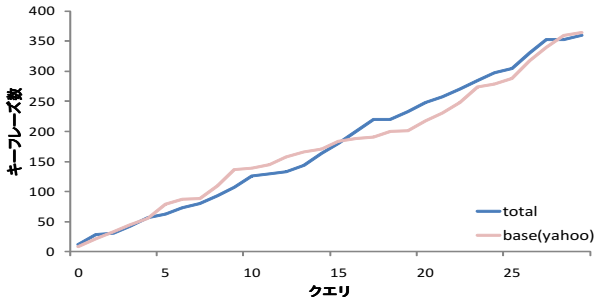

図 9　“ソフト”でのキーフレーズ抽出数の推移 (30 クエリ)

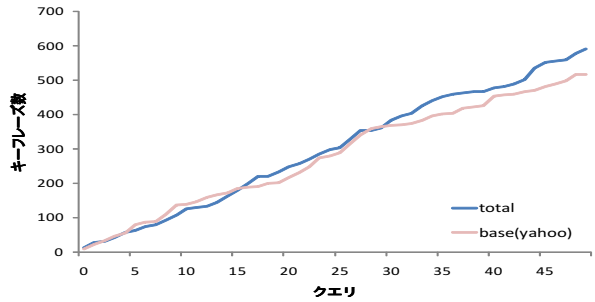

図 10　“ソフト”でのキーフレーズ抽出数の推移 (50 クエリ)

のカテゴリにおいても，base(yahoo) よりも多くのキーフレーズを抽出できている．一方 “ソフト”では，どのカテゴリでも base(yahoo) と同程度のキーフレーズ数しか抽出できていない．また，total はどちらの入力語においても，最もキーフレーズを多く抽出できているわけでなく，全てのカテゴリの平均程度の抽出数になっている．

キーフレーズ数の増加の推移グラフを図 8 と図 9 に示す．図 8 は “日本語”，図 9 が “ソフト”の結果である．横軸はクエリの推移（0〜29），縦軸はキーフレーズ数である．“日本語”では，最初から total の方が抽出数は多いが，後半のクエリに行くに従い，抽出数の差が大きくなっていっている．“ソフト”では，前半は base(yahoo) が多くのクエリを抽出できていたが，15 クエリ目で total が逆転している．最後に base(yahoo) が 1 つのクエリで多くキーフレーズ数を伸ばしており，わずかな差で base(yahoo) が total よりも，多くのキーフレーズを抽出したことになる．最後に，“ソフト”を 50 クエリまで増やして再実験を行った．結果を図 10 に示す．30 件では base(yahoo) の方がキーフレーズ抽出数が多くなったが，その後に total が逆転し，差がつき始めていることがわかる．

7. 考　　察

ユーザの情報要求の言語化を支援するクエリ拡張システムについて説明してきた．カテゴリごとに分けて LSI を行うことで，検索される関連語は，カテゴリに関連のある語といえる．例えば，“ウイルス” では，health カテゴリの “ノロ” は病原体のノロウイルスを指す語であるのに対して，auction カテゴリの “受信” はコンピュータウイルスを指す語である．このことから，カテゴリを区別することにより，病原体のウイルスとコンピュータウイルスを分けて提示することができるといえる．このことから，ユーザは多義語の混同を避けることができると考えられる．

第二段階目のクエリ拡張では，第一段階目の拡張クエリの欠点を補う形になっているといえる．語 “ウイルス” の第一段階目の拡張クエリである “ウイルス　感染” は，複数のカテゴリで出現している．しかし，第二段階目の拡張クエリでは，質問記事が異なるため，キーワード組もそれぞれ異なるものになっている．“ウイルス　感染” は，従来のキーワード組の拡張クエリであるが，第二段目で全く別のコンテキストから作成され

たものとなっている．このように，質問記事とキーワード組をセットで提示することで，これまではわからなかった拡張クエリのコンテキストが理解できる点は，本手法の最も特徴的な点であるといえる．

また，同じカテゴリ，クエリにより検索された質問記事においても，記事によって話題の違う場合も多い．そこから作成される拡張クエリも，大きく変化する．二段階目の拡張では，カテゴリ内のより詳細な部分で多様なクエリが作成されることになる．このことから，多段階のクエリにより，第一段階目でカテゴリによる大域的な多様性と，第二段階目の質問記事による局所的な多様性の二重の多様性の展開が行われていることになるといえる．

多様性の検証実験より，本研究の提案法は，現在の検索エンジンで用いられているクエリ拡張よりも多様性を持っていると考えられる．特に，推薦するクエリの数が増えると，現状のクエリ拡張では，似た話題に対するクエリが多くなるのに対して，提案法では，カテゴリを横断して話題を集めるため，クエリ数が多くなっても，多くの話題からのクエリ推薦ができたと考えられる．“ソフト”では，Yahoo!API の方が結果がキーフレーズ数が多くなったが，28 クエリ目の急激なキーフレーズ数増加が原因だと考えられる．このときのクエリは “ソフト 99”であった．表 5 のとおり，Yahoo!で推薦されるクエリは “DS ソフト”，“フリーソフト”などソフトウェアに関するクエリが多かったが，“ソフト 99”はカーケア商品を専門とする会社の名前であることからキーフレーズ数が一気に増加したと考えられる．このようにキーフレーズを増加させるには話題を変更するクエリが必要になることから，クエリ数が増加しても安定してキーフレーズを増やすことのできた本研究の拡張クエリは，多様性において十分に有用であると考えることができる．

8. 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のクエリ拡張法を提案した．Web 検索ユーザの情報要求を QA サイトの質問記事と対応させ，質問記事を “情報要求の候補” として検索ユーザに提示すると同時に，質問記事から検索エンジンで使用する拡張クエリを生成する．これにより，ユーザの情報要求を言語化し，キーワード化するプロセスを支援する．拡張クエリの生成は，ユーザの情報要求の多様性と曖昧さの多段階性に着目し，QA サイトのカテゴリ，質問記事，キーワードを段階的に提示する手法を提案した．

提案法によるシステムの実装を行い，多様性に関する評価実験を行った．実験では提案法によるクエリ拡張は既存の拡張よりもより幅広い話題を収集できることを確認した．カテゴリの特徴が現れた語が検索されており，カテゴリ分けを用いることの有用性を確認した．また，拡張クエリに質問記事本文も提示することで，生成された拡張クエリのコンテキストが理解でき，さらに具体的な検索が行えるようになることを確認した．

今後の課題として，本研究の拡張クエリは推薦する語が多いほど幅広い話題を収集できることから，より多くの拡張クエリをユーザに提示するインターフェースを実現する必要があると考えている．また，投稿時期を考慮したクエリ拡張を検討していきたい．質問記事は恒常的に投稿されていることから，投稿時期によって質問の傾向が変化し，季節を分けることで，より特徴的な関連語を検索できるものと考えている．

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．

文　献

[1] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze. *Introduction to Information Retrieval*. 2008.

[2] Zhijun Yin, Milad Shokouhi, and Nick Craswell. Query Expansion Using External Evidence. *31th European Conference on IR Research(ECIR2009)*, Vol. LNCS 5478/2009, pp. 362–374, 2009.

[3] 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.

[4] 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.

[5] 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, No. 2A-2, 2010.

[6] 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, No. 2A-3, 2010.

[7] 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.

[8] 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)*, Vol. LNCS 5802/2009, pp. 145–158, 2009.