コンテキスト切替による多様な情報要求に対するWeb検索手法の提案

大塚　淳史† 　関　　洋平†† 　神門　典子††† 　佐藤　哲司††

† 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
†† 筑波大学図書館情報メディア系　〒 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: †{aotsuka,yohei,satoh}@slis.tsukuba.ac.jp, ††kando@nii.ac.jp

あらまし　本論文では，ユーザの情報要求と Web 検索システムが提供する拡張クエリの検索意図を補完する質問記事付き拡張クエリを提案する．コミュニティQA に投稿される質問記事は自然言語で記述されているため，質問記事を拡張クエリの根拠としてすることができる．大量に蓄積されている質問記事からユーザの検索意図に近い質問記事付き拡張クエリを提示するため，カテゴリと季節を用いたコンテキスト切り替えによる検索インターフェースを実装した．ユーザは自らの情報要求に近いカテゴリと季節を選択することで，自身の興味のある話題に関する拡張クエリを参照することができる．また，コンテキストを切り替えることで，ユーザの疑問の範囲を広げる多様な Web 探索が実現できる．提案法を実装したシステムにおいて，季節が拡張クエリを切り替えるコンテキストとして有用であることを実証した．
キーワード　Web 検索，クエリ拡張，コミュニティQA，コンテキスト，季節性

Diversified Query Expansion Methods using Context-Switching for Exploratory Search

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

† Graduate School of Library, Information and Media Studies, University of Tsukuba
1–2,Kasuga,Tsukuba,Ibaraki,305–0855 Japan
†† Faculty of Library, Information and Media Science, University of Tsukuba
1–2,Kasuga,Tsukuba,Ibaraki,305–0855 Japan
††† National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101–8430, Japan
E-mail: †{aotsuka,yohei,satoh}@slis.tsukuba.ac.jp, ††kando@nii.ac.jp

Abstract　In this paper, we propose faceted query expansion methods using Community Question Answering (CQA) resources. CQA users post questions in natural language. We believe that question articles can replace the information needs of web search users. We propose a diversified faceted navigation system using query expansion approach. Our system aim to match user's intentions and queries by browsing a topic and a time. The question articles deriving query expansion are provided for choosing appropriate terms by users. Our evaluations using actual and long-term CQA resources, such as “Yahoo! CHIEBUKURO,” demonstrate that most parts of the CQA questions are posted in periodicity and in bursts.
Key words　Web Search, Query Expansion, Community QA, Context, Seasonality

1. はじめに

近年の Web の普及により、Web ユーザは多様化し，蓄積される情報もより膨大になっている．所望の情報を Web から入手するためには，Web 検索エンジンを用いることが一般的であるが，Web 検索エンジンでは、ユーザはクエリと呼ばれるキーワード組を自らが作成する必要があるが，ユーザが知りたいことを明確に判断できている場合には、容易にキーワードを入力することができる．ユーザの知りたいことが曖昧である場合や，多様な情報を探索的に入手したい場合などは，ユーザは効果的

なクエリを作成することができない．

現在の商用の検索エンジンは，クエリを作成する手助けとして，クエリ拡張機能を提供している．クエリ拡張は，ユーザが最初に入力したクエリに対して，キーワードの追加や修正を行い、次の検索の候補として提示する機能である．ユーザは漠然としたクエリを入力した場合でも，システムが提示する候補から選択するだけで，より具体的な Web 検索を実行できる．しかしながら，クエリ拡張によって提示される候補は，ユーザの検索意図と必ずしも一致しているとは限らない．最も典型的な例として，同音同字異義語が挙げられる．ユーザがコンピュータ関係の “ウイルス” を検索したいと思ったとしても，検索エンジン側が病気に関する “ウイルス” に関する拡張を行うという場合がある．特に，“RS ウイルス”,“MAC ウイルス” など，一見コンピュータ関係か病気関係か判断できない候補が含まれる場合がある．このような，拡張された候補の中にユーザの知識の範囲外の情報が含まれている場合、ユーザ混乱を招き，クエリ拡張は有効に機能しない．同音同字異義語に限らず，語は幅広い概念を持ち，語の背後にある疑問や知りたいこと（情報要求）は多様である．

本論文では，ユーザの情報要求とシステムが提供する拡張クエリの検索意図を補完するものとして，自然言語で記述した疑問文を付与した拡張クエリを提案する．自然言語はユーザにとって理解しすいため，ユーザにとって未知のキーワードが提示されたとしても，自然言語で記述された疑問を参照することで，提示された拡張クエリがどの様な疑問に対しての検索を実現するものなのか容易に把握することができる．本論文では，自然言語で記述された疑問として，コミュニティQA(CQA) に投稿された質問記事を用いる．質問記事を検索ユーザの情報要求と置き換え，質問記事から，Web 検索クエリとなるキーワード組を作成することで，疑問とクエリのセットを検索ユーザに提示する．また，CQA に投稿されている膨大な質問記事の中から，ユーザの情報要求や検索意図に近い質問記事を提示するために，コンテキスト切り替えによる質問記事提示手法を提案する．質問記事を投稿する際に選択する “話題カテゴリ” と質問記事の投稿日時による “投稿の季節” をコンテキストとしてとらえ，検索ユーザは自らの状況に近いコンテキストを選択することで，自身の情報要求に近い質問記事付き拡張クエリを参照することができる．

コンテキストの妥当性を評価するため，CQA から季節性を持つ語を抽出する実験を行った．CQA の各カテゴリで，ある特定の期間のみ頻繁に使用されている語を特定し，周期性が見られるかどうかの評価を行った．周期的に利用頻度が増加する語は季節性があるものと捉え，CQA 全体で季節性がある語がどの程度利用されているのか検証した．

本論文では，以下コンテキスト切り替えによる Web 検索システムの実装と評価について述べる．2 章で関連する研究について説明し，3 章で作成したシステムのインターフェースについて説明する．4 章でシステムの実装とデータセットについて述べ，5 章で CQA の季節性とシステムの出力結果について評価する．6 章で考察し，7 章でまとめと今後の課題について説明する．

2. 関 連 研 究

コンテキストアウェアなクエリ拡張に関する研究として，Cao ら [1] はクリックログとセッションデータから，現在のセッションに近いクエリをログデータから推薦する手法を提案している．また，Semgstock ら [2] はクエリログから、時間とドメインに依存したクエリを抽出し，時間と場所を変化させることで，拡張クエリが変化するシステムを提案した．

クエリの意味や根拠を提示する手法については，Guo ら [3] が，ソーシャルアノテーションに基づくクエリ拡張を提案している．拡張の際に用いたソーシャルアノテーションをそのまま用いることで，クエリの分類と意味付与を同時に行なっている．自然言語処理の技術を用いてクエリにラベルを付与する手法では，Reisinger ら [4] の研究がある．Reisinger らは Web ページから is-a 関係を抽出することで，クエリに付与するラベルを作成し，確率文脈自由文法を用いてクエリとラベルの関係を紐付けている．また，クエリの意味として文章ではなく画像ファイルをクエリと主に提示する手法を Zha ら [5] が提案している．Zha らは画像共有コミュニティに投稿されている画像と画像に付与されているタグを用いて，画像付きの拡張クエリを提示するシステムを作成している．

情報検索において，多様な切り口から検索候補を提示する手法として，ファセット検索がある．CQA とファセット検索を組み合わせたものとして、Yoon ら [6] の研究がある．Yoon らは、CQA のカテゴリと質問記事中の名詞から，Web 検索結果を分類している．Hearst [7] はファセット検索のインターフェースとしてファセットを階層構造化させる手法を提案している．

本手法では，コンテキストはシステムが自動推定するのではなく，コンテキストの候補を提示し，ユーザが自らに合ったコンテキストを選択する．また，クエリの意味推定では，従来研究では，クエリログのデータを基に追加情報をアノテーションする手法が主であったが，CQA のデータセットのみで拡張クエリと拡張の根拠となる質問記事両方を示すという特徴がある．

3. コンテキスト切り替えによる クエリ拡張システム

本論文では，コンテキストを切り替えることによる，多様な Web 情報検索を実現するシステムを開発する．開発したシステムのスクリーンショットを図 1 に示す．システムのインターフェースは 4 つのブロックから構成される．ファセット検索ブロックでは，2 次元タブとタグクラウドを用いてファセット検索のための拡張クエリを作成する．タグクラウドのキーワードを選択すると，ユーザが最初に入力したクエリに選択したキーワードを追加した “質問検索クエリ” が作成される．“質問検索クエリ” によって、CQA クエリブロックで，CQA から関連する質問記事が検索される．その際に，質問記事からキーワード抽出を行い，“CQA クエリ” が作成される．“CQA クエリ” は質問記事に付与される形でユーザに提示される．Web 検索ブロックは，Web 検索結果が表示される．クエリ表示ブロックで

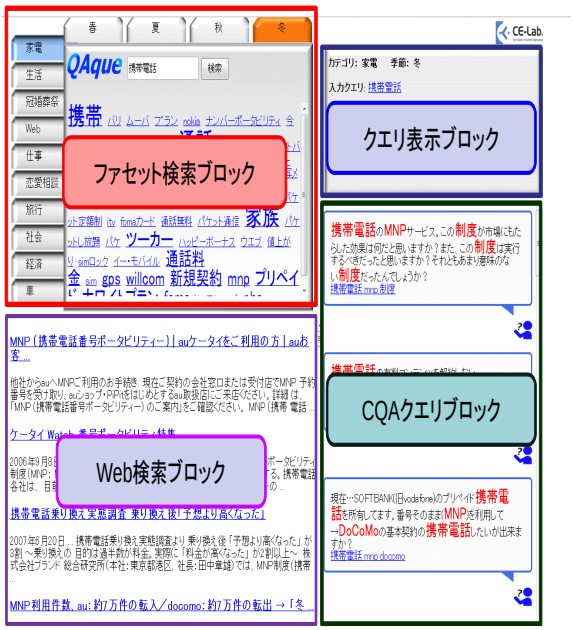

図 1　プロトタイプシステム画面

は，これまでの操作によって作成されたクエリが表示されている．ユーザはクエリの情報を見ながら、ファセット検索ブロックで、タブやタグクラウドを用いたインタラクティブな Web 検索を行うことができる．

以下ではまず，ファセット検索ブロックでのタブとタグクラウドを組み合わせたインターフェースについて説明し，次に CQA クエリブロックでの質問記事付きクエリ（CQA クエリ）について詳述する．

3.1　タブとタグクラウドによる拡張クエリの提示

入力されたクエリの背後には多様な情報要求が存在している．たとえ，入力されたクエリが同じだとしても，情報要求が同じとは限らない．例を挙げると，クエリ “日光” を入力したユーザが，地名の日光を調べたいのか，陽射しに関する日光について調べたいのか判断することは難しい．また，地名の日光だとしても，日光の観光地について知りたいのか，観光地についても夏に代表される観光地と冬が有名な観光地で異なる場合もある．また，観光だけでなく日光の歴史や文化が知りたいのかで，提示すべき拡張クエリは異なる．多様な情報要求に対応したクエリ拡張を行うためには，できるだけ幅広い切り口からのクエリを提示する必要がある．

本論文では、様々な話題の切り口から検索候補を提示するファセット検索型のクエリ拡張機能を提供するために，2 次元タブとタグクラウドを組み合わせたインターフェースを提案する．インターフェースの概要を図 2 に示す．ユーザは検索窓にクエリを入力すると，システムが入力したクエリに関連する話題カテゴリと季節のタブを出力する．タブは 2 次元構成になっており、話題カテゴリと季節をそれぞれ一つ選択する．タブを選択すると，関連語のタグクラウドが表示される．タグクラウドには、フォントサイズの異なる語が出力されるが，フォントサイズは，その関連語を選択することによって検索される質問記事数に依存している．つまり，フォントサイズの大きい関連語を選択すると，CQA クエリブロックでより多くの質問記事が表示される．タブは任意のタイミングで切り替える事ができる．タブはファセット検索における話題の単位に相当するため，タブを切り替えることでタグクラウドの内容は変化する．

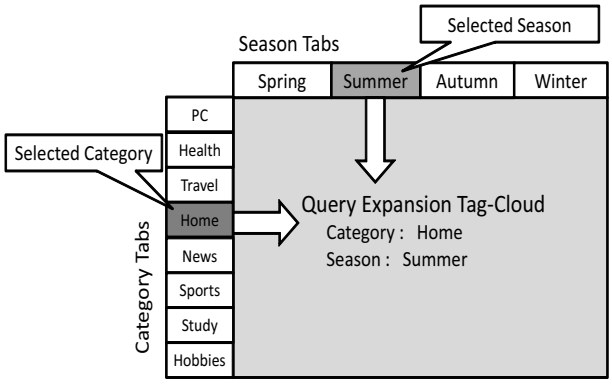

図 2　タブ-タグクラウドインターフェース

ファセットとして使用する話題カテゴリは CQA のカテゴリを使用する．CQA では，質問記事を投稿する際に，必ず一つのカテゴリを選択する必要がある．そのため，コミュニティはカテゴリによって一定の話題ごとの分類がなされている．また，カテゴリ名は短くわかりやすいキーワードが使用されているため，ユーザがどの話題に興味があるのかを選択する際の指標として有用である．また，CQA は誰でも自由な時期に投稿可能なため，内容の類似するの質問記事も数多く投稿されている．内容が似た質問は同時期かあるいは，1 年後などに投稿されているケースが多い．このように，CQA に投稿される質問記事は，時期によって内容が変化し，かつ周期性を持っているといえる．この傾向は Web 検索にも当てはまると仮定し，Web 検索を行った際と同時期の拡張クエリを提示する，また時期が異なることでより多様なクエリを推薦できる．今回は最も一般的な周期である年周期でデータをまとめた．1 年を区切る指標として季節を導入した．

3.2　質問記事付き拡張クエリ（CQA クエリ）

質問記事付き拡張クエリ（CQA クエリ）は CQA の質問記事から作成する．CQA クエリの例を図 3 に示す．ファセット検索ブロックでは，ユーザにとって未知のキーワードが示されることがある．その場合でも，CQA クエリによってキーワードから “具体的な疑問” を表出することができる．そのため，質問記事は拡張クエリの根拠として利用することができる．また，質問記事は個人的かつ詳細的な内容が多いため，同じキーワードが使用されている質問記事であっても質問の内容は全く異なる．そのため，CQA クエリは拡張の根拠だけでなく二段階目の多様な拡張クエリの展開を行なっている．ファセット検索ブロックのタブ・タグクラウドインターフェースと CQA クエリブロックでの CQA クエリを組み合わせて利用することで，ユーザは自らが本当に知りたかった疑問を発見し，より多様な Web 検索を実現することができる．

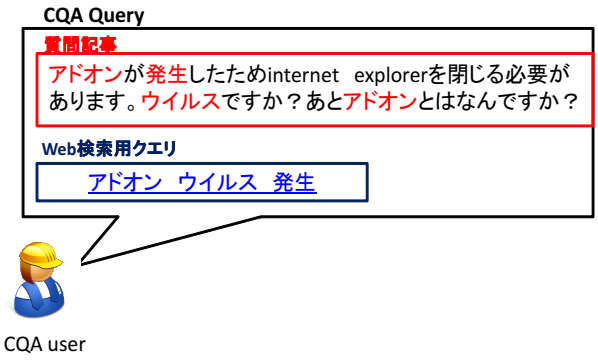

図 3　質問記事付き拡張（CQA クエリ）

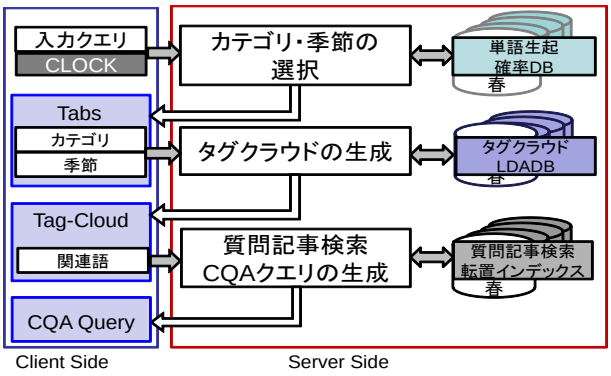

図 4　システム構成図

4.　クエリ拡張システムの実装

本システムの構成を図 4 に示す．本システムでは，主に 3 つの処理を行う．まず，ユーザが入力したクエリから，関連するカテゴリと季節を決定する．ユーザがタブからカテゴリと季節を選択することで，タグクラウドが出力される．最後に，タグクラウドの関連語を選択すると関連する質問記事とその質問記事から作成した拡張クエリが出力される．

以下で，それぞれの処理の実装について説明する．

4.1　データセット

本システムで使用する CQA コーパスとして，本論文では，国立情報学研究所の提供する Yahoo!知恵袋(注1) コーパスを利用する．コーパスには、2004 年から 2008 年の Yahoo!知恵袋のデータが収録されている．本論文では，Yahoo!知恵袋が正式サービスとなった 2006 年から，2008 年までの 3 年分のデータを使用する．Yahoo!知恵袋は数百のカテゴリが 3 階層構造で存在している．すべてのカテゴリ別にデータベースを作成した場合，似たカテゴリの話題が関連カテゴリとして抽出されてしまい，話題の切り替えとして有効に機能しなくなることから，今回はある程度広い話題ごとにカテゴリが分かれるように，複数のカテゴリを集約する．最下層の小カテゴリは，カテゴリによって質問記事数に差があるため，今回はどのカテゴリにおいても比較的多数の質問記事が投稿されてる中階層のカテゴリのデータのみを使用し，中カテゴリの質問記事を階層構造に習った状態で集約していく．例えば，最上位のカテゴリ “インターネット、PC と家電” カテゴリは，“インターネット”，“パソコン”，“AV 機器”，“’ 携帯電話” などのカテゴリを持つが，それらのカテゴリの質問記事を全て一つにまとめて “デジタル・家電” カテゴリという一つのカテゴリとして扱う．また，季節性のために，データセットを期間毎に分割する．2006 年から 2008 年までの各年で同じ月に投稿されたデータを集約する．使用する季節とデータは以下のようになる．

- 春: 各年 3 月から 5 月に投稿されたの質問記事
- 夏: 各年 6 月から 8 月に投稿されたの質問記事
- 秋: 各年 9 月から 11 月に投稿されたの質問記事
- 冬: 各年 12 月から 2 月に投稿されたの質問記事

データセットの一部を表 1 に示す．$NavigationCategory$ は集約したカテゴリ名でシステムにはこのカテゴリ名が表示される．$Second-LevelCategory$ は集約した中カテゴリであり，春〜冬はデータセットに含ませる質問記事数である．冬のデータが最も多くなっていが，これは Yahoo!知恵袋が正式サービスを開始した 2006 年 4 月からデータを採集しているためである．作成したシステムでは，$NavigationCategory$ を 20 カテゴリ，作成し，それぞれを 4 つの季節に分割したため，全 80 個のデータセットが作成される．

4.2　タブによるコンテキスト提示の実装

ユーザがシステムを利用する際，最初に $InititalQury$ としてキーワードを入力する．また，その際ブラウザ上の時間をサーバに送信する．サーバーではユーザが入力したキーワードと時間から，関連のあるカテゴリと季節を選択する．サーバ上に保存されているカテゴリと季節の組み合わせからなる 80 個のデータセットから最も関連のあるデータセットを選択する．季節の関連はユーザが入力をシステムに送信した時間から，対応する月の属する季節を選択する．季節は 4 つで固定のため，ここで決定するのは初期状態で選択される季節のみでユーザはタブ切り替えによって自由に季節を変更できる．カテゴリは，各カテゴリでの入力キーワードの出現確率によって決定する．カテゴリ C において，キーワード w の出現確率 $P_{C,w}$ は以下で計算できる．

$$P_{C,w} = \frac{N_{C,w}}{N_C} \tag{1}$$

$N_{C,w}$ はカテゴリ C でのキーワード w が含まれる質問記事数である．N_C はカテゴリ C での質問記事総数である．出現確率が高いカテゴリは入力キーワードと関連のあるカテゴリであるとみなし，出現確率の降順に最大 10 個のカテゴリをタブに表示する．

4.3　タグクラウドの実装

ファセット検索ブロックでは，タブとタグクラウドを出力する．システムは，ユーザが入力したキーワード，タブにより選択したカテゴリと季節から関連語のタグクラウドを作成する．そのため，入力したキーワードの関連語検索を使用する．本実装では，データセットである CQA の特徴を活用し，質問に共に使用される確率が高い語を提示するために統計的手法を用いた確率的トピックモデルによるアプローチによる関連語検索を

(注1)：http://chiebukuro.yahoo.co.jp/

表 1　Navigation Category とデータセット例

Navigation Category	Second-Level Categories	春	夏	秋	冬
趣味	アニメ, コミック, 本, おもちゃ, 占い, くじ, 伝統工芸	46,281	55,793	67,941	89,893
エンタメ	映画, 音楽, 芸能人, ミュージカル, テレビ/ラジオ, 伝統芸能	77,018	85,588	94,995	106,841
デジタル・家電	パソコン, デジタルカメラ, インターネット, ソフトウェア, AV 機器, 携帯電話	92,981	91,920	96,057	110,821
旅行	国内旅行, 海外旅行, 交通案内, 路線案内, テーマパーク	85,900	96,284	99,627	107,634
恋愛	恋愛相談，人生相談	14,9830	16,9239	19,6009	22,0074

実装する．特に，Blei ら [8] が提案した，潜在的ディリクレ配分法（LDA）を用いる．LDA では，文書ー単語行列を潜在的なトピックを用いて，文書ートピック行列，トピックー単語行列に分割することができる．トピックは単語の確率分布によって定義される．ここで，各トピックでの出現確率の分布が近い単語は潜在的に関連のある単語であるといえる．

LDA では，教師なし学習によって文書ー単語行列からトピック集合 Z を推定する必要がある．本実装では，崩壊型ギブスサンプリングを用いた推定を用いる．文書 d，n 番目の単語 $w_{d,n} = v$ のトピック $z_i = k$ の更新式は以下で計算できる．

$$P(z_i = k|Z_{-i}, W) = \frac{N^d_{k-i} + \alpha}{N^d_{-i} + T\alpha} \frac{N^v_{k-i} + \beta}{N_{k-i} + W\beta} \tag{2}$$

$-i$ 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．α，β はディリクレ分布のハイパーパラメータである．

トピック分布 Z から文書-トピック分布 θ と，トピック-単語分布 ϕ が得られる．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．

$$\hat{\theta^k_d} = \frac{N^d_k + \alpha}{N^d + T\alpha} \tag{3}$$

$$\hat{\phi^w_k} = \frac{N^v_k + \beta}{N_k + W\beta} \tag{4}$$

トピックは単語の出現確率分布で表現されているため，データセットに含まれる全ての単語はトピック k での出現確率を持っている．各トピック間で似た出現確率を持つ単語同士は実際の質問記事においても関連して使用される確率が高い語である．そこで，各トピックでの出現確率の類似度を計算することで，単語の類似度を計算できる．各トピックを次元圧縮された行列の基底とみなし，コサイン距離をよって類似度を計算する．

$$sim(w_1, w_2) = cos(P_{w_1}, P_{w_2}) = \frac{P_{w_1} \cdot P_{w_2}}{|P_{w_1}||P_{w_2}|} \tag{5}$$

P_{w_1}，P_{w_2} は単語 w_1 と w_2 の各トピックでの出現確率分布である．実際のシステムでは類似度の降順に関連語を提示する．また，タグクラウドでの文字の大きさはその関連語を選択した際に検索される CQA クエリの数によって決定する．検索される質問記事数に比例し文字のサイズは大きくなる．LDA による検索の結果，潜在的に関連があっても，質問記事で入力キーワードと同時に使用されていない関連語については今回はタグクラウドに表示しない．

4.4　質問記事検索と拡張クエリ作成の実装

システムはタグクラウドから選択されたキーワードを基に関連する質問記事を提示する．また，質問記事からキーワードを抽出し，Web 検索用のクエリを作成する．クエリ尤度モデルはクエリ Q が文書 D を発生させる確率 $P(D|Q)$ の最大尤度を発見することにより検索を行う情報検索手法である．この手法では，ベイズの定理により $P(D|Q)$ を以下の式に変形して考える．

$$P(D|Q) = \frac{P(Q|D)P(D)}{P(Q)} \tag{6}$$

$P(Q)$ は文書に対して独立な確率を持つため，定数項として除外できる．また $P(D)$ は文書に対して事前知識が必要となるため省略可能であり，最終的に $P(D|Q) \propto P(Q|D)$ として考えることができる．$P(Q|D)$ は文書 D がクエリ Q を表現する確率である．つまり，文書 D 対して尤もらしいキーワードの組み合わせによって作成されるクエリ Q を発見することができる．これは，本実装において，質問記事からキーワード組を抽出し拡張クエリを作成する際に利用可能である．そのため，本実装ではクエリ尤度モデルを用いた質問記事検索を実装する．

クエリ尤度モデルでは $P(Q|D)$ が最も高くなる文書が最もクエリに適合する文書となる．$P(Q|D)$ は複数個の単語の組み合わせ $\mathbf{w}$ によって表現される．

$$P(Q|D) = \prod_{w \in Q} P(w|\theta_D)^{c(w,Q)} \tag{7}$$

θ_D は文書 D を構成する言語モデルである．本実装ではユニグラム言語モデルを用いた．$c(w, Q)$ はクエリ Q における単語 w の出現頻度である．$P(w|\theta_D)$ は，言語モデルが単語 w を生成する確率であるが，本実装では，逆頻度要素を持たせるため，ディリクレスムージングを用いた以下の計算を行った．

$$P(w|\theta_D) = \frac{c(w, D)}{|D| + \mu} + \frac{\mu}{|C|(|D| + \mu)} \sum_{D \in C} \frac{c(w, D)}{|D|} \tag{8}$$

$c(w, D)$ は文書 D における単語 w の出現頻度である．$|D|$ は文書 D における総単語数，$|C|$ は文書コレクション C における文書総数である．μ はパラメータであり今回は $\mu = 1$ としている．クエリ尤度モデルを用いて，ユーザが入力したキーワードとタグクラウドから選択したキーワードが含まれる文書をランキング化する．また，質問記事 D_{que} から最も特徴的な拡張ク

表 2　デジタル・家電カテゴリでの単語抽出結果

順位	単語	バーストした月
1	年賀状	11 月, 12 月
2	湿る	6 月, 7 月
3	除	6 月, 7 月, 8 月
4	4 月	3 月, 4 月
5	冷房	6 月, 7 月, 8 月

表 3　旅行カテゴリでの単語抽出結果

順位	単語	バーストした月
1	2 月	1 月, 2 月
2	GW	4 月
3	雪	1 月, 2 月
4	4 月	2 月, 3 月
5	桜	3 月, 4 月

表 4　恋愛カテゴリでの単語抽出結果

順位	単語	バーストした月
1	チョコ	2 月
2	バレンタイン	1 月,2 月
3	お返し	2 月,3 月
4	義理	2 月
5	正月	1 月

エリ Q_{ex} を作成するため，$P(Q_{ex}|D_{que})$ が最も高くなる単語を 3 語抽出する．その際，ユーザが入力した単語とタグクラウドから選択した単語は優先的に抽出される．

5. 評価実験

本システムは，CQA に投稿される質問記事の内容が季節によって変化していることを前提としている．そこで，実際に CQA の季節性を確認し，季節性導入の妥当性を評価する．また，季節性を持つ単語をシステムに入力した際のシステムの出力結果を比較する．

まず，CQA の季節性の評価方法と評価結果について述べ，次に，季節性を持つ語に対して，システムの出力結果について説明する．

5.1 CQA における単語の出現確率推移による季節性評価

CQA は投稿の制限がほとんど存在せず，ユーザは自由に何回でも質問記事を投稿することができる．そのため，流行している話題などは関連する質問記事数が増加するという傾向がある．本実験では，CQA における話題の流行を単語ベースで抽出する．特に，ある期間や時期のみに限定して流行する語を抽出する．ここで，語の流行（バースト）とは各カテゴリにおいて，その期間に投稿された全質問記事に対して，その語が使用された質問記事の割合（出現確率）をみることで判定する．特定の期間のみバーストする語は月毎の出現確率にバラつきがあるといえる．そのため，月毎の出現確率の分散を正規化した変動係数 $(C.V)$ により出現確率のバラつきが大きい単語を抽出する．しかし，変動係数のみでは，突発的なイベントなど一時期限定の単語が上位に抽出されてしまう．本来の目的である季節性のある単語の抽出のため，これらの語を除外する必要がある．

そこで，周期性を評価し，ある一定の期間で周期的にバーストする単語のみを抽出する．単語のバーストの判定には Vlachos ら [9] の提案したバースト判定手法を用いる．バースト判定手法を以下に示す．

（1） データ系列 $t = (t_1, ..., t_n)$ に対して長さ w の移動平均 MA_w を計算

（2） 移動平均の平均 $mean(MA_w)$ と標準偏差 $std(MA_w)$ からしきい値 $cutoff = mean(MA_w) + x * std(MA_w)$ を計算

（3） i 番目のデータに対して，$\{\ MA_w(i) > cutoff\ \}$ の場合バーストと判定

データ系列 t は 2006 年 1 月から 2008 年 12 月までの 36ヶ月分の出現確率のデータである．移動平均の長さ w は季節の区切りと対応させ $w = 3$ に設定している．また，しきい値計算のための重み付けパラメータ x は予備実験の結果 $x = 2.5$ とした．各単語においてバーストする月を抽出し，バースト間隔が一定の単語のみを抽出する．

抽出できた単語の例を表 2，表 3，表 4 に示す．CQA の中でも投稿件数の多いデジタル・家電カテゴリ，旅行カテゴリ，恋愛カテゴリでの結果を示す．“4 月”，“正月”，“GW” など期間や時期そのものを指す単語の他に “チョコ”，“湿る”，“桜” といった直接時期を指していない語が含まれていることがわかる．“チョコ” は恋愛・人間相談カテゴリでは，バレンタインデーに関する投稿が増加するため，1 月，2 月にバーストしている．“湿る” は特に高湿な夏の 6 月、7 月にバーストしている．また，“年賀状” は，1 月ではなく，11 月と 12 月にバーストしている．旅行カテゴリでは，直接月を指定する場合のほか，“雪” や “桜” といった観光や気候関係で季節が影響しているものが抽出されている．

出現確率の推移グラフを図 5，図 7，図 6 に示す．横軸が 2006 年 1 月から 2008 年 12 月までの期間，縦軸が出現確率を示している．どちらのグラフにおいてもバーストしている月のみ出現確率が高い値になっていることがわかる．同時に，バーストしていない月では，出現確率がほぼ 0 に近い．バーストの傾向も周期性が強く，毎年同じ月に，高い出現確率を示していることがわかる．また，バーストする際には，一月間だけではなく，二から三ヶ月の間バーストが続いていることがわかる．

5.2 システム出力結果における季節性の評価

CQA の季節性評価実験によって得られた単語を用いて，システムの季節性を評価する．表 5 にシステムの出力結果をまとめたものを示す．$InititalQuery$ はシステムに最初に入力したキーワードである．それぞれの入力語でで最も関連のあるカテゴリでの季節ごとのタグクラウドの出力キーワード数をまとめている．出力語数はタグクラウドに表示されたキーワードの総数である．ユニーク語数は，タグクラウドに出力された語のうち，他の季節では出力されず，その季節のみで出力されるユニークな語が提示される数であり，ユニーク率はその割合である．表 2 から表 4 でバーストしている月に対応する季節が最も多くのキーワードが提示され，またユニーク語も多く出力されていることがわかる．ユニークキーワード数の推移をグラフに

表 5　タグクラウドの出力例とユニークキーワード数

Initial Query[カテゴリ]	季節	総出力語数	ユニーク語数	ユニーク率	キーワード例
年賀状	春	35	5	0.14	黒, 送る, 用紙サイズ
[デジタル・家電]	夏	26	6	0.23	レーザ, プリント
	秋	52	14	0.26	dpi, スキャン, 画像
	冬	62	23	0.37	レフィルインク, 筆ぐるめ, プレビュー
桜	春	45	37	0.82	ブルーシート, 花見, カラオケ
[旅行]	夏	13	13	1.00	宮津，船島
	秋	21	18	0.86	紅葉, 祭り
	冬	22	15	0.68	梅林, 庭，河津
チョコ	春	57	21	0.37	お返し, ブランド, ルイヴィトン
[恋愛]	夏	28	8	0.29	あげる, バレンタイン, 渡す
	秋	27	4	0.14	手作り，お菓子, ワイン
	冬	69	30	0.43	本命, ネクタイ, ラブレター

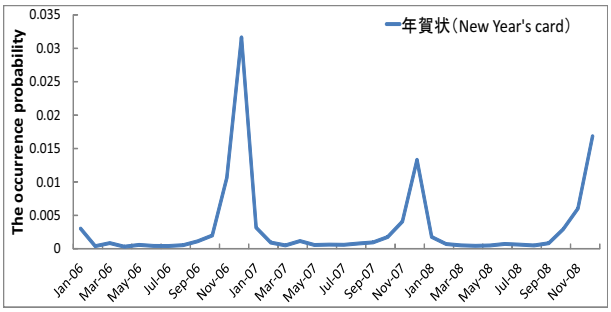

図 5　“年賀状” のデジタル・家電カテゴリでの出現確率の推移

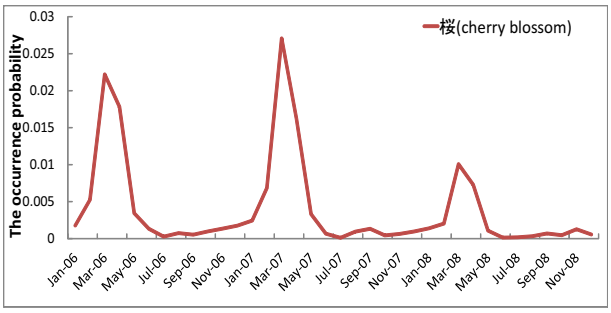

図 6　‘桜” の旅行カテゴリでの出現確率の推移

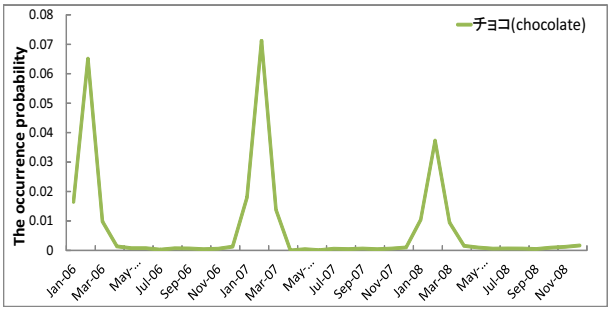

図 7　“チョコ” の恋愛カテゴリでの出現確率の推移

まとめたものを図 8 に示す．横軸は季節，縦軸はタグクラウドのユニークキーワード数である．ユニークキーワード数はバーストしている月に対応した季節が最も多く，その次の季節は極端に少なくなっている．またバーストする季節に向かうに連れ徐々にユニークなキーワードが増加していることがわかる．

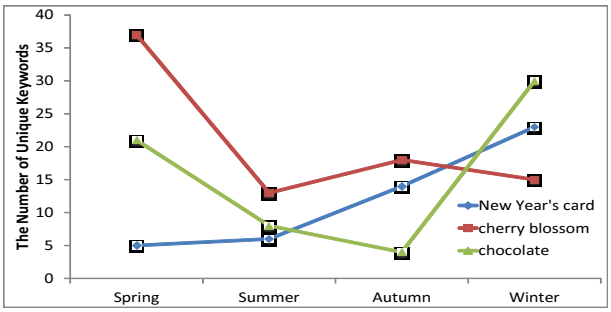

図 8　ユニークキーワード数の季節ごとの推移

6. 考　　察

季節性を持つキーワード抽出実験では，“2 月” などの月そのものを示す単語が抽出されていた．このような時間を示す単語は，Web のクエリとしてはあまり有用ではないが，このような特定の時期を示す単語が上位に出力されていたことから，季節性の抽出手法は妥当であると考えられる．また，今回抽出した “バラつきが大きく周期性のある語” は毎年ある特定の時期に高い出現確率を示し，それ以外の時期は出現確率がほぼ 0 になることがら，これらの語は季節性を持っているといえる．また，CQA の質問記事投稿の傾向にも十分に季節性があると言える．

今回抽出した季節性を持つ語の出現確率の推移を見てみると，バースト期間に幅があることがわかる．例えば，恋愛カテゴリにおける “チョコ” はイベント関連の語であり，バーストとして検出されたのは 2 月のみであるが，1 月と 3 月にも比較的高い出現確率を示している．このように，CQA におけるバーストは実際のイベントから 1 ,2ヶ月の幅を持って発生している．またデジタル・家電カテゴリにおける “年賀状” は 1 月の正月に関する語であるが，実際は 11 月と 12 月にバーストしている．これは実際に年賀状を作成するのが 11 月ごろから始まるためであると考られる．このように CQA の質問の傾向としてバーストにはある程度期間があり，また実際のイベントなどよりも早い時期に質問が多く投稿されるという傾向があることがわかる．投稿時期の分化能を季節というある程度幅を持たせた指標にしたことで，これら投稿の幅や時期のズレに対応することが

できたのではないかと考えられる．

システムの出力結果では，バーストした季節が最もユニークなキーワードが提示されていた．バーストした季節は投稿される質問記事数も多いため，より詳細な内容の質問記事が多いためであると考えられる．このことから，システムは CQA の季節性を反映できていると考えることができる．また，バースト以外の季節においても多くのキーワードが出力されているケースもあるが，他の季節と重複したものが多いという傾向があった．これは，質問の内容がごく一般的なものであることを示しており，季節による差別化が行われているといえる．また，更に投稿質問数が少ない場合は少数の質問記事によりタグクラウドに出力されるキーワードが影響を受けていることがわかる．

7. おわりに

本論文では，コミュニティQA を利用した質問記事付きの拡張クエリを提案した．また，質問記事検索のためカテゴリと季節を用いたファセット検索型のクエリ拡張システムを提案した．システムの妥当性を示す実験として，コミュニティQA の中から季節性を持つ語を抽出し，それらが毎年ある特定の時期のみ多く出現し，その他の時期はほとんど使用されていない語であることを示した．季節性を持つ語を入力とした際のシステムの出力結果をユニークキーワード数によって比較し，該当する季節は他の季節よりもユニークなキーワードが多く推薦されることを示した．

今後は，より季節性やカテゴリの依存度の強い単語が優先してタグクラウドに出力されるように，検索モデルやトピックモデルなどの手法を検討する必要がある．また，質問記事付き拡張クエリについても，ユーザビリティと検索結果の多様性の両面から評価を行う予定である．

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．

文　献

[1] Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, and Hang Li. Context-Aware Query Suggestion by Mining Click-Through and Session ata. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining(KDD'08)*, 2010.

[2] Christian Sengstock and Michael Gertz. CONQUER: a System for Efficient Context-Aware Query Suggestions. *Proceedings of the 20th International Conference on World Wide Web(WWW'11)*, 2011.

[3] Jiafeng Guo, Xueqi Cheng, Gu Xu, and Huawei Shen. A Structured Approach to Query Recommendation with Social Annotation Data. *Proceedings of the 19th ACM international conference on Information and knowledge management(CIKM'10)*, 2010.

[4] Joseph Reisinger and Marius Pasca. Fine-Grained Class Label Markup of Search Queries. *Proceedings of The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT'2011)*, 2011.

[5] Zheng-Jun Zha, Linjun Yang, Tao Mei, Meng Wang, Zengfu Wang, Tat-Seng Chua, and Xian-Sheng Hua. Visual Query Suggestion: Towards Capturing User Intent in Internet Image Search. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 2010.

[6] 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.

[7] Marti Hearst. Design Recommendations for Hierarchical Faceted Search Interfaces. *ACM SIGIR Workshop on Faceted Search*, 2006.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[9] Michail Vlachos, Christopher Meek, Zografoula Vagena, and Dimitrios Gunopulos. Identifying Similarities, Periodicities and Bursts for Online Search Queries. *Proceedings of the 2004 ACM SIGMOD international conference on Management of data(SIGMOD'04)*, 2004.