話題の周期性に着目した情報要求言語化のためのクエリ拡張手法の提案

大塚　淳史† 　関　 洋平†† 　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
E-mail: †{otsuka,satoh}@ce.slis.tsukuba.ac.jp, ††yohei@slis.tsukuba.ac.jp

あらまし　Web ユーザが生成するコンテンツには，周期的な性質を持つものが多く存在する．本論文では，コミュニティQA の自然言語で投稿された質問記事の中から，周期的な話題に関する質問記事を抽出する手法を提案する．コミュニティQA の質問記事集合から時系列トピックモデルにより，時間追跡可能なトピックを作成する．トピック話題変動の周波数解析することにより得られた特徴量をクラスタリングし，周期的に変化するトピックを特定する．トピックモデルの文書ートピック分布から，周期性を持つトピックに関連の深い質問記事を抽出することで，大量の質問記事の中から，周期的な話題に関する質問記事を抽出する．抽出できた質問記事から拡張クエリを作成することで，季節性の高いクエリを推薦できたので報告する．
キーワード　コミュニティQA，話題分類，周期性，時系列トピックモデル，クエリ拡張

Query Expansion Method based on Periodicity of the Subject for Verbalization Latent Information Needs

Atsushi OTSUKA†, Yohei SEKI††, 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
E-mail: †{otsuka,satoh}@ce.slis.tsukuba.ac.jp, ††yohei@slis.tsukuba.ac.jp

Abstract　Web contents generated by users periodicity. In this paper, we propose extraction method for question articles which contain periodicity from community QA resources written by natural languages. This method create time continuous topic from question article resources, and extract topic variation using JS divergence. We identify periodicity topics by clustering feature quantity made by frequency analysis. finally, we extract periodicity question articles using topics-documents distribution in topic topic model. Our works show that periodicity question articles by our extract method demonstrate high seasonality query expansion.
Key words　Community QA, Topic Classification, Periodicity, Time Continuous Topic Model, Query Expansion

1. はじめに

スマートフォンに代表される情報端末の普及に伴い，Web はより身近な存在となっている．Web ユーザは日常の生活の中で生じた疑問に対して，Web 検索エンジンや，コミュニティQA（CQA）を用いることで，所望の情報を入手している．Web 検索エンジンや CQA で使用するクエリや質問記事作成するため，ユーザは自身の調べたいことである疑問から “調べたいこと” を文章化する．本論文では，これを情報要求の言語化と呼ぶ．そのため，クエリや質問記事は，ユーザが知りたいことである情報要求を強く反映したものになっている．

ユーザの情報要求は時間とともに刻々と変化していくが，その変化には周期性があるとされている．Vlachos ら [1] や，Shokouhi ら [2] の研究により Web 検索エンジンに入力されるクエリには周期性があることを明らかにしている．例えば，“Halloween” というクエリは毎年 10 月に検索頻度が急激に上昇するが，それ以外の期間ではほとんど検索されないクエリである．また，村田ら [3] は，クエリ入力語に最初にクリックされる Web ページを検索意図とみなすことで，Web ユーザの検索意図には周期性があることが明らかになっている．検索の周期性を特定することで，時期に応じたクエリ拡張など，情報検索での幅広い応用が可能になる．Semgstock [4] らはクエリの

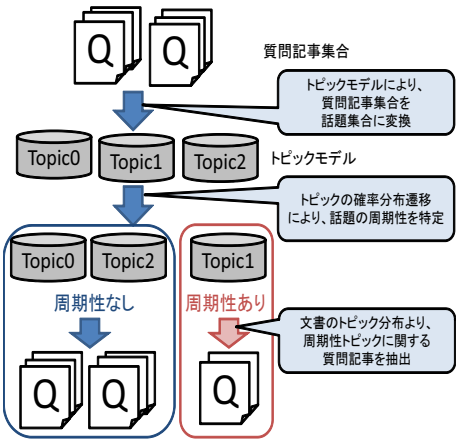

図 1　話題変動のタイプによる質問記事分類の流れ

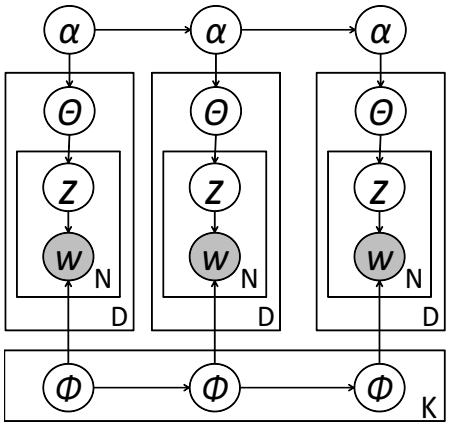

図 2　DTM のグラフィカルモデル (時間分割数 3 のとき)

時間的な周期性に着目して，時刻によって異なる拡張クエリを推薦する手法を提案している．

CQA においても，投稿される質問記事の内容には周期性が存在しているとされている．Shotok ら [5] は，近い内容の質問記事が繰り返し投稿されることに着目し，新しい質問記事が投稿された際に，過去の回答済みの質問記事を用いた自動回答を行う仕組みを提案している．CQA は自然言語で記述した質問記事を入力とするため，キーワード組で表現する検索エンジンのクエリよりも高い表現力を持っていおり，質問記事中の単一のキーワードだけでは，質問記事の主題を特定することは難しい．そのため，質問記事に記述されている内容が周期性を，Web 検索エンジンのクエリのように，単純なキーワードの出現頻度だけで話題の周期性を特定することは困難である．

本論文では，CQA から，繰り返し議論される話題に関する質問記事を抽出する手法を提案する．トピックモデルを用いて，質問記事集合を “話題” の単位となるトピック集合に変換し，トピックの時間変化を追跡することで，変動に周期性が見られるトピックを特定する．周期的な話題変動をするトピックに関連の深い質問記事を抽出することで，CQA の質問記事集合の中から，周期的な話題に関して議論される質問記事を抽出する．本論文では，抽出した周期的な話題に関する質問記事集合をリソースとして，情報要求の言語化を支援するクエリ拡張を実装した．質問記事を Web ユーザの情報要求とみなし，季節ごとに異なる質問記事と拡張クエリを提示することで，クエリに関するユーザの情報要求の時間変化を追跡しながら，多様な Web 検索を実現している．

本論文の構成は以下の通りである．まず，2 章で CQA から周期的に議論される話題に関する質問記事の抽出手法について説明する．3 章では 2 章の手法により抽出した質問記事を用いたクエリ拡張システムについて説明する．4 章で，クエリ拡張システムの出力結果を用いた，周期性に関する評価実験を行う．5 章で考察し，6 章でまとめと今後の課題について述べる．

2. 話題の周期性に着目した質問記事分類

本章では，CQA の話題変動の周期性に基づく質問記事分類手法について説明する．図 1 に提案手法の流れを示す．質問記事を “話題” 単位となるようにトピック化し，トピックの中から周期的な変動をするトピックを特定することで，周期的な話題を発見する．最終的に，周期性のあるトピックに関連する質問記事を抽出することで，質問記事集合の中から，周期性に関連する話題の質問記事を分類する．

以降は，トピックモデルを用いた CQA 質問記事の話題化とトピックの時間変化による話題変動追跡手法，そして，話題変化の周期性発見手法について詳述する．

2.1 時系列トピックモデルによる CQA の話題変動抽出

CQA の質問記事は自然言語で記述されているため，一つの質問記事中に多くの単語を含む．そのため，Web 検索エンジンのクエリ頻度のように，単純な単語の出現頻度では，質問記事の話題遷移を特定することは難しい．そこでまず，質問記事集合をトピックモデルによりトピック化する．トピックモデルは，文書と単語の間には潜在的なトピックがあると仮定するモデルであり，トピックは単語の出現確率分布で表現される．同じ話題で使用される単語は，確率分布内で近い確率を持つ．一つのトピックは一つの “話題” に対応させることができ，トピックの時間変化を追跡することで，CQA の質問記事の話題の変化を追跡することができる．

本論文では，時間追跡可能なトピックモデルとして，Blei ら [6] の Dynamic Topic Model（DTM）を用いる．DTM のグラフィカルモデルを図 2 に示す．z はトピック，w は単語を表し，K はトピック数，N は単語数，D は文書数である．α はハイパーパラメータ，θ は（トピック数×文書数）のトピック比率行列，ϕ は（単語数×トピック数）の単語分布行列である．DTM では，モデル生成に時間情報を用いるため，同一トピックを時間を超えて追跡できるという特徴がある．図 2 では，時間分割数 3 の場合を示しており，トピックの単語分布 ϕ が初期状態から 2 回遷移している．時間分割数を TS とした場合，トピックは $\phi_0, \phi_1, ..., \phi_{TS-1}$ までの TS 個の単語分布を持つ．

トピックの時間変化は，トピック内の確率分布の変化である．DTM の各トピックでは，同一の話題で使用される単語が近い確率を持つ．そのため，各トピックの単語分布の遷移を追跡す

ることで，同一話題の変化を調べることができる．トピック時間変化量は確率分布の類似度を算出する JS ダイバージェンスにより得る．JS ダイバージェンスは，確率分布 P と Q から次式で与えられる．

$$JS(P||Q) = \frac{1}{2}\left(\sum_x P(x)\log\frac{P(x)}{R(x)} + \sum_x Q(x)\log\frac{Q(x)}{R(x)}\right) \quad (1)$$

R は確率分布 P と Q の平均であり，$R = (P+Q)/2$ である．JS ダイバージェンスは $0 \leqq JS$ の値をとり，同じ確率分布 $(P = Q)$ のとき，$JS = 0$ をとる．

時刻間の確率分布の類似度が高い場合は，話題がほとんど変化していないことを表し，類似度が低い場合，時間の遷移によって話題が変化したことを示している．

本論文では，時間分割数 TS の時系列データに対して，データの始点となる時刻を基準時刻 t_0 とし，t_0 と $t_0, t_1, ..., t_{TS-1}$ までの，全ての時刻との JS ダイバージェンスを計算する．計算した JS ダイバージェンス値を時系列上に並べることで，トピック内の話題が基準時刻からどれだけ乖離していくのかを分析する．

2.2 周波数解析による話題変動の周期性特定

JS ダイバージェンスにより得た，トピックの時間変化に関する時系列データを離散フーリエ変換（DFT) を用いて検出する．フーリエ変換は，時系領域で表現されているデータを周波数領域で表現するための手法である．$n_0, n_1, ..., n_{N-1}$ までの N 個のデータによって表現される時系列データ x から，離散フーリエ変換によって得られる周波数 k のスペクトルは以下で与えられる．

$$X[k] = \sum_{n=0}^{N-1} x[n]\mathrm{e}^{-j\frac{2\pi}{N}kn} \quad (2)$$

ここで，指数関数部分はオイラーの公式により実数部と虚数部に分割した以下の式を得る．

$$X[k] = \sum_{n=0}^{N-1}\left\{(x[n]\cos\frac{2\pi}{N}kn) - j(x[n]\sin\frac{2\pi}{N}kn)\right\} = Re[k] + Im[k] \quad (3)$$

各周波数スペクトルの信号の強さは，パワースペクトルを計算することで求めることができる．周波数 k の周波数スペクトル $x[k]$ のパワースペクトル $S(x[k])$ は式（3）を用いて．

$$S(X[k]) = \sqrt{Re[k]^2 + Im[k]^2} \quad (4)$$

パワースペクトルにより，入力した時系列データにどの周波数成分が多く含まれているのかを調べることができる．周波数 k のとき，周期 T は以下の式で与えられる．

$$T = \frac{N}{k} \quad (5)$$

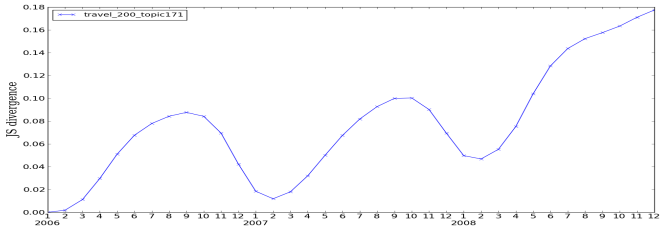

図 4　旅行カテゴリ，トピック 171 の話題変動

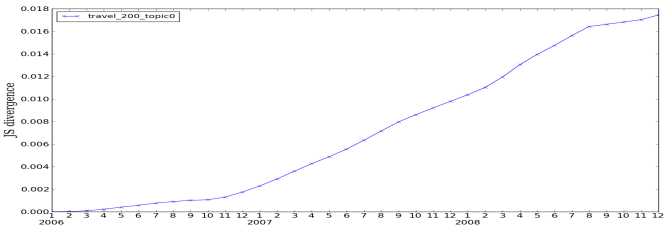

図 5　旅行カテゴリ，トピック 0 の話題変動

離散フーリエ変換により得た話題変動の周波数領域のパワースペクトル分布により，話題変動を周期的な特徴に基いて議論することができる．そこで，パワースペクトル分布を素性としてトピックのクラスタリングを行うことで，質問記事の話題を変動のタイプに基いて分類する．離散フーリエ変換では，サンプリング定理により，周波数 $k = 0, 1, ..., \frac{N}{2} - 1$ までの $N/2$ 個のパワースペクトルを得ることができる．しかし，$k = 0$ は周期 $T = \infty$ となり，周期性に関係しないため除外し，$k = 1, ..., \frac{N}{2} - 1$ までのスペクトルをクラスタリングの素性とする．

2.3 話題変動のタイプに基づく質問記事分類

トピックモデルでは，文書はトピックの出現確率分布により表現されている．トピック数 N のモデルにおいて，文書 D とトピック z_m との関連度は，出現確率 $P(z_m|D)$ により表現され，$\sum_{m=0}^{N} P(z_m|D) = 1$ である．

本論文では，周期的な話題変動をするトピック集合 PZ に対して，これらのトピックの出現確率の和 S_{pz} を以下の式により算出する．

$$S_{pz} = \sum_{z_i \in PZ} P(z_i|D) \quad (6)$$

S_{pz} が設定した閾値を超えた時，文書 D は周期的な話題に関する質問記事であるとして抽出する．

3. 周期性に基づくクエリ拡張

本章では，周期的な話題に関するクエリ拡張システムについて説明する．まず，2 章で提案した手法を実際の CQA リソースに適用した際の結果を示し，その結果を元に作成したクエリ拡張のためのデータセットについて述べる．次に，作成したクエリ拡張システムについて説明する．

3.1 Yahoo!知恵袋からの周期的な質問記事抽出

提案手法を実際の CQA リソースに適用した結果を示す．国

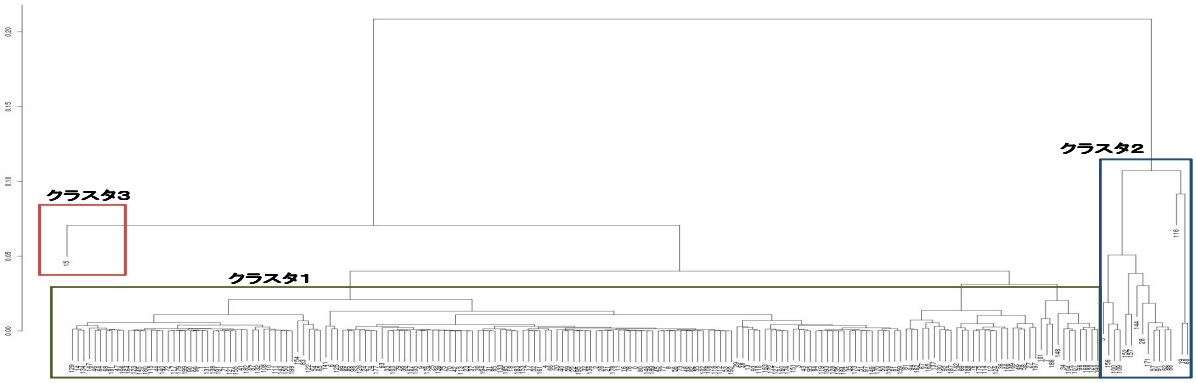

図 3　旅行カテゴリ 200 トピックの周波数解析によるクラスタリング結果

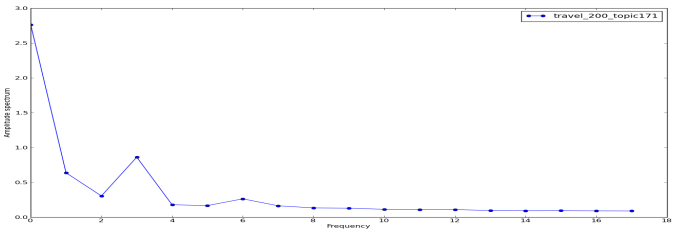

図 6　旅行カテゴリ，トピック 171 のパワースペクトル分布

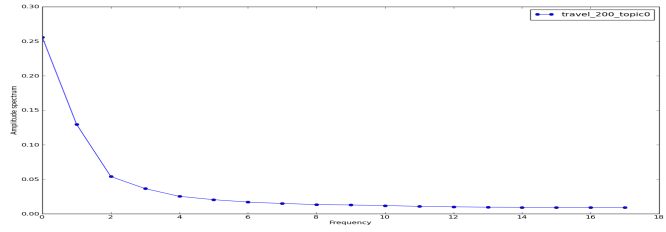

図 7　旅行カテゴリ，トピック 0 のパワースペクトル分布

立情報学研究所提供の Yahoo!知恵袋コーパス(注1)を使用した．データの期間を 2006 年 1 月から 2008 年 12 月までの 12ヶ月間，DTM のトピック数は 200 と設定した．カテゴリは毎月の質問記事の投稿件数が安定している，旅行，PC，健康の 3 カテゴリを用いた．

時系列トピックと JS ダイバージェンスにより，トピックの話題変動を追跡した結果を図 4 に示す．この図は，旅行カテゴリ，トピック 171 の話題変動の様子である．横軸が 2006 年 1 月から 2008 年 12 月までの時系列，縦軸が JS ダイバージェンス値である．JS ダイバージェンスが上昇するほど，2006 年 1 月から話題が遠くなっていることを示している．2 月から徐々に JS ダイバージェンス値が上昇しおり 9,10 月でピークとなり，その後は JS ダイバージェンス値が低下していく話題変動のパターンとなっていることがわかる．2006 年から 3 年間全てで，増減を繰り返す同様のパターンで JS ダイバージェンス値が推移することから，旅行カテゴリのトピック 171 は周期的な話題であると判別することができる．一方，図 5 は，旅行カテゴリ，トピック 0 の話題変動の様子を示している．トピック 0 では，

(注1)：http://www.nii.ac.jp/cscenter/idr/yahoo/chiebkr2/Y_chiebukuro.html

表 1　旅行カテゴリでの階層的クラスタリング結果

クラスタ	トピック数	トピック番号
1	183	下記以外の全トピック
2	16	5, 19, 28, 57, 60, 62, 88, 91, 100, 109, 144, 152, 156, 157, 171
3	1	15

表 2　PC カテゴリでの階層的クラスタリング結果

クラスタ	トピック数	トピック番号
1	199	下記以外の全トピック
2	1	73

表 3　健康カテゴリでの階層的クラスタリング結果

クラスタ	トピック数	トピック番号
1	194	下記以外の全トピック
2	6	1 55 58 123 127 151 177

JS ダイバージェンス値が単調増加となっている．このことからトピック 0 の話題は周期性を持たないといえる．

話題変動を離散フーリエ変換により，周波数領域のスペクトル分布で表現したものを図 6 と図 7 に示す．図 6 はトピック 171，図 7 がトピック 0 である．横軸は周波数，縦軸はスペクトルエネルギーを示している．トピック 0 のスペクトル分布は一般的な指数関数分布となっているが，トピック 171 では，$k = 3$ でスペクトルにピークが立っている．周波数から，トピック 171 の話題変動は，周期 12 の成分を多く含むことがわかる．

周波数スペクトル分布を素性として，階層的クラスタリングを実行した結果を図 3 に示す．階層的クラスタリングから，特徴的な 3 つのクラスタが形成できていることがわかる．トピック 171 はクラスタ 2，トピック 0 はクラスタ 1 に属している．クラスタ 2 に属する 16 のトピックは全て $k = 3$ でスペクトルのピークが立っており，クラスタ 2 が周期的な性質を持つ話題に関するトピックであることがわかる．

PC，健康カテゴリにおいても同様に話題変動の周波数スペクトルによるクラスタリングした結果を表 2 と表 3 に示す．PC カテゴリではトピック 73 からなるクラスタ 2，健康カテゴリでは 7 個のカテゴリからなるクラスタ 2 がそれぞれ周期的な話題を持つトピッククラスタとして抽出できた．

表 4　クエリ拡張のためのデータセット（周期性を持つ質問記事）

カテゴリ	全質問記事数	周期性質問記事数	抽出割合
旅行	217,130	25,780	0.12
PC	516,442	7,206	0.014
健康	471,189	14,401	0.031

3.2　データセット：周期的な話題に関する質問記事の抽出

3.1 節での結果を基に，クエリ拡張システムのためのデータセットを作成する．旅行，PC，健康ではクラスタ 2 が周期性を持つ話題に関するトピック群である．これらのトピックを用いて質問記事抽出を行った．文書分類の式（6）の閾値は 0.3 に設定した．抽出結果を表 4 に示す．全質問記事数は，期間内にカテゴリに投稿された質問記事の総数，周期性質問記事数はトピックモデルから抽出した，周期性に関連する質問記事数である．システムでは，抽出できた周期性質問記事をデータセットとして利用する．

システムでは，時期ごとに異なる拡張クエリを提示するため，抽出した周期性質問記事を季節ごとに分割する．質問記事集合の中から，3 月から 5 月に投稿された質問記事を春の質問記事集合とする．同様に，6 月から 8 月の間に投稿されたものを夏，9 月から 11 月に投稿されたものを秋，そして 12 月から 2 月の間に投稿された質問記事を冬の質問記事集合として分割する．システムでは，季節ごとにそれぞれ拡張クエリを作成する．

3.3　情報要求の言語化を支援するクエリ拡張システム

周期的な話題に関する質問記事をデータセットとしてクエリ拡張システムを実装した．システムのスクリーンショットを図 8 に示す．システムは，筆者らが提案しているクエリ拡張システム [7] と同様の手法により実装している．システムでは，左上部の検索窓に検索キーワードを入力すると，タグクラウドに拡張クエリで追加する関連語が表示される．タグクラウドから関連語を選択すると，システム右下部に入力語と関連語が含まれる CQA の質問記事とキーワード組が提示される．この質問記事とキーワード組がセットで提示される拡張クエリを，本論文では “質問記事付き拡張クエリ（CQA クエリ）” と呼ぶ．CQA クエリの質問記事は Web ユーザの言語化された情報要求に相当する．キーワード組にユーザが理解できないキーワードが含まれていたり，検索意図が理解できない拡張クエリが提示された場合でも，自然言語で記述された質問記事を参照することで，提示されたキーワード組がどのような情報要求から発生したものなのかを把握することができる．

関連語が提示されるタグクラウドの左辺と上辺には，タブがついており，タブを切り替えることで異なるコンテキストに関する関連語のタグクラウドが表示される．左辺のタブはカテゴリを切り替えるタブであり，上辺のタブは季節を切り替えるタブである．季節タブでは，タブを切り替えることにより，その季節に関連した関連語のタグクラウドが提示される．上辺左端の定番タブでは，季節に関係なく，日常的に投稿される話題に関連する質問記事を用いた関連語が提示される．

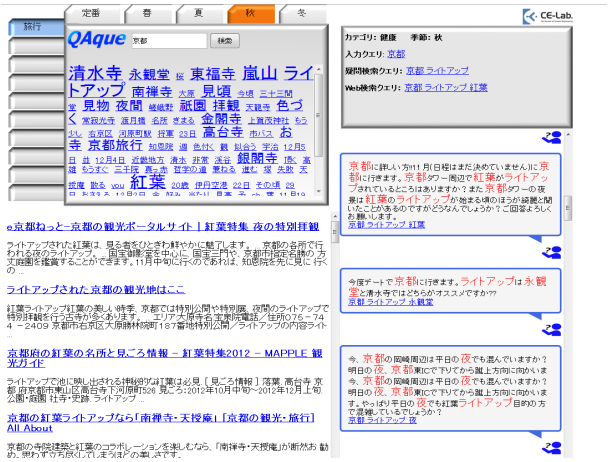

図 8　情報要求の言語化を支援するクエリ拡張システム

4.　評 価 実 験

本章では抽出した質問記事，および質問記事から作成したクエリ拡張システムに関する評価について述べる．まず，話題の変動タイプに基づいて周期的な話題に関する質問記事を抽出した結果について評価実験を行い，次に，実装したクエリ拡張システムにおいて季節タブを切り替えた際の，タグクラウドに表示される関連語について比較，実験を行う．

4.1　周期的な話題の質問記事抽出結果に関する評価

2 章で提案した手法に基いて，Yahoo!知恵袋から周期的な話題に関連する質問記事を抽出した結果を表 4 に示す．その結果，カテゴリによりバラつきはあるものの全体の 1〜10 %程度の質問記事を抽出することができている．本節では，抽出した質問記事の周期性に関して評価を行う．

抽出した質問記事集合で議論されている話題の傾向を把握するため，質問記事を形態素解析し，質問記事に使用されている単語の出現頻度を算出した．表 5 は，旅行カテゴリで抽出された高頻度語を出現頻度順に並べたものである．全質問記事は旅行カテゴリに投稿された全ての質問記事，周期性質問記事は，提案手法によって抽出した質問記事集合である．全体，周期性どちらの質問記事集合ともに最も出現頻度の高い語は “旅行” である．全質問記事集合では，“いい”，“できる”，“お願い” などの一般的な単語が高頻度で使用されているが，周期的な話題に関する質問記事集合では “北海道”，“時期”，“紅葉” などの語が高頻度語として抽出されている．全質問記事集合と周期性質問記事集合での高頻度語の出現順位の変化を図 9 に示す．“行う” という単語は，周期性質問記事集合において 24 番目に多く使用されている単語であるが，全質問記事集合でも 27 番目に多く使用されており，出現頻度の順位にほとんど変化がない．しかし，“雪” という単語は全質問記事集合での順位は 256 位なのに対し，周期性質問記事集合では 23 番目であり，全体に対しての出現頻度が大幅に上昇している．同様に “紅葉” という単語も全質問記事集合では出現頻度 221 位なのに対して，周期性質問記事集合では 13 位と大幅に上昇していることがわかる．

表 5　旅行カテゴリでの高頻度語

質問記事分類	高頻度語
全質問記事	旅行，いい，お願い，予定，できる，場所，ない，見る，ホテル，東京，考える，観光，日本，聞く，時間，お勧め，京都，良い，人，安い，おすすめ，お店，オススメ，海外，大阪
周期性質問記事	旅行，見る，予定，北海道，時期，韓国，お願い，いい，場所，考える，できる，今年，紅葉，福岡，香港，今，お勧め，京都，ない，聞く，オススメ，東京，雪，行う，利用

表 6　PC カテゴリでの高頻度語

質問記事分類	高頻度語
全質問記事	パソコン，できる，pc，使う，方法，表示，いい，お願い，出る，ファイル，使用，dvd，cd，ソフト，わかる，出来る，画面，ない，見る，windows，設定，エクセル，入れる，購入，起動
周期性質問記事	印刷，できる，プリンター，用紙，設定，方法，出来る，プリント，使う，年賀状，プリンタ，写真，サイズ，パソコン，紙，ワード，エクセル，いい，お願い，ない，画面，作成，ページ，プレビュー，文字

表 7　健康カテゴリでの高頻度語

質問記事分類	高頻度語
全質問記事	ない，いい，飲む，人，前，病院，痛い，できる，出る，自分，今，最近，薬，痛み，方法，聞く，症状，良い，食べる，お願い，寝る，悪い，見る，病気，気
周期性質問記事	汗，かく，インフルエンザ，刺す，入る，かかる，ない，いい，蚊，出る，暑い，人，寒い，方法，今，夏，予防接種，最近，できる，足，体，良い，寝る，時期，対策

PC，健康カテゴリにおいても同様に高頻度語を抽出した結果を表 6 と表 7 に示す．PC カテゴリでは周期性質問記事集合の高頻度語に “印刷”，“年賀状”，“プリンタ” などの語が抽出されている．また健康カテゴリでは，最も出現頻度の高い語は “汗” であった．その他にも，“インフルエンザ”，“蚊” などの単語が周期性質問記事集合では抽出できていることがわかる．

4.2　クエリ拡張システムに関する評価

本節では，収集したデータセットを基に実装したクエリ拡張システムに関する評価を行う．まず，タグクラウドに表示される入力語の関連語一覧について評価を行い，次に，質問記事付き拡張クエリである CQA クエリについて評価を行う．

4.2.1　タグクラウドの関連語

本システムではデータセットを投稿時期により春〜冬に分類し，それぞれの季節ごとに異なる拡張クエリが提示される．タグクラウド上辺のタブを切り替えることによりタブに対応した季節に関する関連語のタグクラウドが表示される．本システムのタグクラウドには以下の特徴がある．

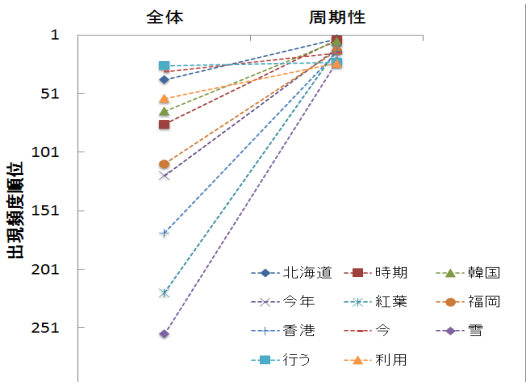

図 9　質問記事集合の違いによる単語の出現頻度順位の変化

（1）　関連語の表示順は，LDA により算出した関連度順となる．

（2）　関連語の文字の大きさは，入力語と関連語を含む質問記事数により決定される．

（3）　関連度順位が上位であっても，入力語と関連語が共に含まれている質問記事が存在しない場合，タグクラウドには表示されない．

本システムのタグクラウドは，質問記事付き拡張クエリである CQA クエリを閲覧するためのものである．LDA を用いた潜在的意味解析による関連語算出では，実際の質問記事で共起しない語同士が高い関連度となる場合が存在するが，今回はそのような単語は除外している．

タグクラウドの評価として，ユニーク語の抽出による評価を行った．ユニーク語とは，一つの季節のタグクラウドにのみ登場する語である．タグクラウドから多くのユニーク語が抽出できれば，システムが季節性を反映できていることが示される．本論文では，旅行カテゴリにて実験を行った結果を示す．タグクラウドを表示させるためには，入力語の入力が必要となる．本論文では，旅行カテゴリで出現頻度の高い語である “京都” と “紅葉” を入力語とした時の結果を示す．図 10 は，旅行カテゴリにおいて，“京都” と “紅葉” を含む質問記事の月ごとの投稿頻度を示したものである．“京都” は毎月ほぼ同じ割合で質問記事が投稿されているが，“紅葉” は毎年 9 月から 11 月ごろに集中して質問記事が投稿され，他の月ではほとんど質問記事が投稿されていないことがわかる．ことのことから，“京都” は周期性のない語，“紅葉” は語自体に周期性を持つ語であるといえる．これらの周期的に異なる性質を持つ語を入力とした時のシステムの出力結果について評価を行う．関連度順に 150 個の関連語を使用した時の，タグクラウドの出力結果について評価実験を行った．

“京都” を入力語としたときの実験結果を図 11 に示す．比較のため，旅行カテゴリの全質問記事集合を対象に同様のクエリ拡張を実施した結果も示す．棒グラフがタグクラウドに表示されたユニーク語数，そして折れ線グラフはタグクラウドに表示された関連語の中でのユニーク語の割合である．“紅葉” では，秋以外の季節で，全質問記事集合により実装したタグクラウドのほうがより多くのユニーク語が表示されている．しかし割

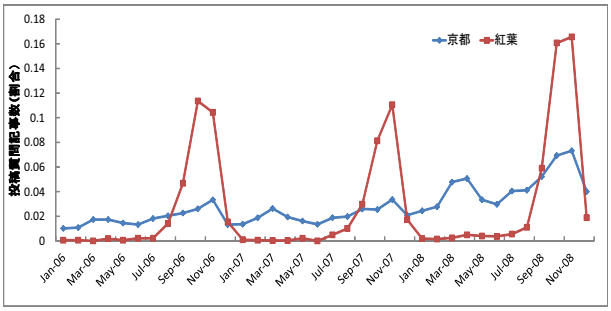

図 10　京都，紅葉を含む質問記事の投稿頻度の推移

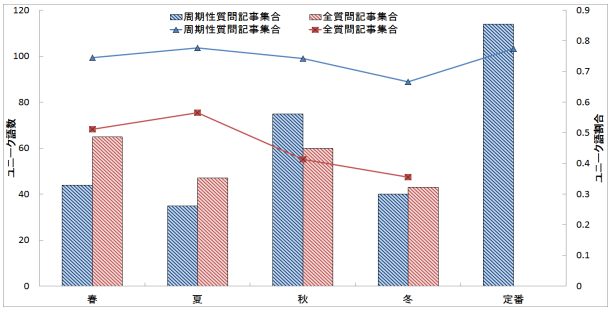

図 11　入力語 “京都” でのタグクラウドのユニーク語

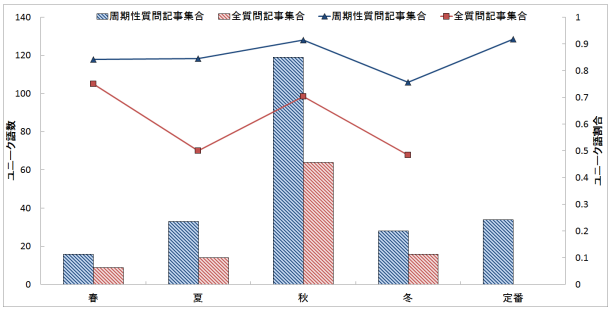

図 12　入力語 “紅葉” でのタグクラウドのユニーク語

合で見ると，ユニーク語はタグクラウド全体の 4〜5 割程度となっている．これは，タグクラウドに表示される関連語の約半数が，他の季節のタグクラウドでも表示されていることを示している．一方，周期性に基いて分類した質問記事集合のタグクラウドでは，どの季節においても，タグクラウドに表示される関連語の約 7 割がユニーク語であり，他の季節では提示されない語となっている．また，定番タブにおいて，多くのユニーク語が表示されているという結果となった．

“紅葉” を入力とした結果は図 12 となる．全ての季節で，周期性に基づく質問記事集合によるタグクラウドの方が多くのユニーク語が提示されおり，タグクラウドに表示される関連語のうち，割合も 8 割から 9 割がユニーク語となっている．秋には 119 個のユニーク語が提示されており，定番を含む他の季節よりも圧倒的に多くのユニーク語が提示されることがわかる．

4.2.2　CQA クエリの季節性

旅行カテゴリで入力語を “京都” とした時，“紅葉” という関連語は夏，秋，冬の 3 つのタグクラウドで表示される．そのため，“紅葉” はある特定の季節のみに出現するユニーク語ではない．タグクラウドから “紅葉” を選択すると “京都” と “紅葉” が含まれる質問記事の CQA クエリが表示される．CQA クエリの例を表 8 に示す．表中の CQA クエリは，“紅葉” を選択した際に，最初に表示される CQA クエリである．夏の CQA クエリでは，11 月に京都旅行に行くための計画に関する質問記事，秋の CQA クエリは，紅葉の見頃に関する質問記事，そして冬の質問記事は，まだ紅葉を見ることができるかに関する質問記事であり，同じ，“京都” と “紅葉” を含む質問記事であっても，それらの文脈や主題は異なっていることがわかる．

5.　考　　察

5.1　周期的な話題の質問記事抽出に関する評価

本論文では，話題変動のタイプに着目し，コミュニティQA から周期的に議論される話題に関する質問記事抽出を行った．時系列トピックモデルと JS ダイバージェンスによりトピックの話題変動を時系列上の変動量に変換し，周波数解析に基いてクラスタリングを行った結果，旅行カテゴリでは 16 個，PC カテゴリでは 1 個，健康カテゴリでは 6 個の周期性のあるトピックが抽出できた．このトピックを用いて質問記事を収集したところ，質問記事集合中に季節に関連する特徴が見られるようになった．

旅行カテゴリでは，カテゴリ全体では “東京”，“観光” などの語が頻繁に使用されている語であったが，周期性トピックから収集した質問記事集合では “紅葉”，“雪” などの単語が高頻度語となっている．“紅葉” は図 10 に示した通り，毎年 9 月から 11 月の秋ごろに集中して投稿される語である．このような語は語自体に非常に強い季節性があり，これらの語が使用される質問記事は周期性，季節性に関連が強い語であると考えられる．そのため，“紅葉” や “雪” がデータセットの高頻度語として抽出できたことから，データセットが周期的な話題に関する質問記事を多数保持していると考えることができる．健康カテゴリでも “汗”，“インフルエンザ” などの語が抽出できていることから，旅行カテゴリと同様の傾向があると考えられる．

PC カテゴリでは，収集したデータセットの高頻度語が “印刷”，“プリンタ” など比較的話題の語が集中的に抽出されていた．これは，周波数解析によるクラスタリングの結果 1 トピックしか抽出できなかったことが影響していると考えられる．PC カテゴリにおいては “年賀状” という語は，質問記事の投稿パターンが旅行カテゴリの “紅葉” と同じ傾向を示す [8] ことから，年賀状に関連のある印刷トピックが抽出されたと考えられる．

5.2　情報要求の言語化を支援するクエリ拡張システムに関する考察

提案手法により収集した質問記事集合を用いたクエリ拡張システムでは，全質問記事を用いた場合と大きな差が現れている．LDA [9] を用いた関連語のタグクラウドでは，全質問記事を用いて実装したほうが，より多くの関連語が表示されている．しかし，その中の約半数は他の季節のタグクラウドでも表示される語であるため，季節タブを用いてタグクラウドを切り替える効果は薄かった．しかし，提案手法により収集したデータセットにより実装したタグクラウドでは，他の季節で表示されない，“その季節独自の語” がタグクラウド全体の 7〜9 割提示されて

表 8　旅行カテゴリでの CQA クエリの例

タグクラウド			CQA クエリ	
入力語	選択語	季節	質問記事本文	キーワード組
京都	紅葉	夏	11 月の中旬くらいに京都に行こうと思ってます。もう紅葉シーズンは終わってますかね？京都の紅葉はいつくらいまでが見ごろなんですか？	京都 紅葉 11 月
		秋	京都の紅葉ってもう見れますか？京都の紅葉ってもう見れますか？	京都 紅葉 見れる
		冬	明日京都へ行くのですが、まだ紅葉が残っているところはありますか？	京都 紅葉 残る

おり，季節性をより強く反映させたものとなっている．“京都” のような一見季節性と関連のない語であっても，季節性のある関連語が各季節ごとに提示されている．また，定番タブで表示される関連語のほとんどが，全質問記事でのいずれかの季節のタグクラウドで提示された関連語であることから，全質問記事で他の季節と重複していた語が定番に集約されていると考えることができる．そのため，定番タブでは，多数のユニーク語を含む大量の語が提示されている．“紅葉” の様に元々季節性が強い語を入力した場合は，語に該当する季節のタグクラウドに多くの語が提示されている．これは “紅葉” を含む質問記事が秋に該当する 9 月から 11 月の間に集中して投稿されていることが影響していると考えられる．

CQA クエリでは，タグクラウドで重複していた語を選択しても，季節ごとに話題が若干異なる CQA クエリが提示されることが明らかになった．これは，高い表現力を持つ自然言語で記述された質問記事をリソースとしている影響が強いのではないかと考えられる．CQA クエリでは，自然言語で記述された質問記事により，検索意図の理解向上につながると考えている．例えば，旅行カテゴリで夏に “京都” の関連語として “紅葉” が推薦されるのは，違和感を感じるが質問記事を参照することで，将来の旅行に向けた準備の一環としての質問であることがわかる．このことから，“京都に秋や冬に旅行をするなら夏の間に考えておく必要がある” という気づきを与えることができる．このような情報検索行動に対する付加価値も CQA クエリの効用であると考えられる．

6. おわりに

本論文では，話題の周期性に着目したコミュニティQA の質問記事分類手法を提案し，分類した質問記事を用いたクエリ拡張を実装した．実験では，周期性に基いて収集したデータセットから実装したクエリ拡張の方が，全質問記事を用いて実装した時よりも，より季節独自のクエリが推薦できることを明らかにした．

今後の課題として，カテゴリの拡大，バーストなどの話題変動への対応がある．CQA には話題毎に非常に多くのカテゴリが存在するため，その中で議論される話題も多様である．また，ある事件や出来事に応じて関連する質問が多く投稿されるバーストは CQA でも多く見られる傾向であり，クエリ拡張などの情報検索においても特徴的な推薦を行うために有用である．今後は，カテゴリを増やしながら，様々が話題変動を発見，分類する手法を提案する予定である．

謝　辞

本研究の一部は，筑波大学図書館情報メディア系プロジェクト研究による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] 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)*, pp. 131–142, 2004.

[2] Milad Shokouhi and Kira Radinsky. Time-sensitive query auto-completion. *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'12)*, pp. 601–610, 2012.

[3] 村田眞哉, 戸田浩之, 松浦由美子, 片岡良治. サーチエンジンにおける検索意意図の周期的変化の検出. 第 2 回 Web とデータベースに関するフォーラム (WebDB Forum 2009), 2009.

[4] 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)*, pp. 265–268, 2011.

[5] Anna Shtok, Gideon Dror, Yoelle Maarek, and Idan Szpektor. Learning from the Past: Answering New Questions with Past Answers. *Proceedings of the 21st International Conference Companion on World Wide Web(WWW '12)*, pp. 759–768, 2012.

[6] David M. Blei and John D. Lafferty. Dynamic topic models. *Proceedings of the 23rd International Conference on Machine learning(ICML'06)*, pp. 113–120, 2006.

[7] 大塚淳史, 関洋平, 神門典子, 佐藤哲司. 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. 第 4 回データ工学と情報マネジメントに関するフォーラム論文集 (DEIM2012), F8-3, 2012.

[8] 大塚淳史, 関洋平, 神門典子, 佐藤哲司. コミュニティQA を用いたクエリ拡張のためのコンテキスト抽出に関する一考察. 日本データベース学会論文誌, Vol. 11, No. 1, pp. 1–6, 2012.

[9] 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.