コミュニティ QA を用いた
情報要求の言語化支援に関する研究

筑波大学
図書館情報メディア研究科

2013 年 3 月

大塚淳史

目次

図目次

表目次

第 1 章

序論

1.1 背景

インターネットの普及に伴い，World Wide Web（Web）に蓄積されるデータは増加の一途をたどっており，蓄積される情報もより多様なものとなっている．そのため，大量の情報が蓄積されている Web 上から所望の情報を入手するための，Web 検索の重要性はますます高まってきている．ユーザは，自らが調べたいことである情報要求を “疑問” として想起し，言語化することで検索クエリを作成する．検索クエリを作成することができれば，Web から所望の情報を入手することができるが，ユーザが想起できない” 疑問” に対しては，情報要求を言語化することができず，クエリを作成することができない．ユーザ個人が考えうる “疑問” の範囲には限りがあるため，自由入力型のクエリ検索では，ユーザは，自身が想起できる情報要求の範囲の情報しか入手することができず，Web 情報に対して大規模なインデックスを持つ Web 検索エンジンを最大限に活用することができないという問題がある．

現在の商用の Web 検索エンジンは，ユーザのクエリ作成を支援し，ユーザが想起できない情報要求に対する Web 検索を実現する手段として，クエリ拡張機能を提供している．クエリ拡張では，検索エンジンに蓄積されているクエリログデータから，ユーザの検索意図を推測し，ユーザが入力したクエリのキーワードの変更，又は新規キーワードを追加した新たなクエリを提示する．ユーザが入力したクエリが単一のキーワードから構成される単純なものであっても，検索エンジンが提示した拡張クエリを選択することで，的確な 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，ユーザの情報要求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．同音異義語のように，語には，様々な観点が含まれる，複数の観点からクエリが推薦されることは，情

報検索の多様性の観点から見れば好ましいことである．しかし，観点の違いをユーザが認識していない場合，ユーザの混乱を招く結果となる恐れがある．

本論文では，コミュニティ QA（CQA）の質問記事により Web 検索の支援を行う．コミュニティ QA は，疑問を質問記事として自然言語で記述し，投稿することで他ユーザの回答を得ることができる知識共用サービスである．代表的な CQA の例として，Yahoo!知恵袋[*1]や，教えて!goo[*2]が挙げられる．自然言語で記述された質問記事は Web ユーザの “疑問” そのものであるといえる．質問記事を Web ユーザの情報要求とみなすことで，潜在的な情報要求が明確に言語化できる．そこで，質問記事を検索支援に用いることで，情報要求を意識した “疑問ベース” の Web 検索が実現できると考えている．

以下では，本論文での取り組みについて説明する．まず 1.2 節で，周期性に着目した，コミュニティ QA の話題変動の抽出について述べる．次に，1.3 節で，CQA を用いた Web 検索支援として開発した，コンテキスト切り替え型のクエリ拡張システムの概要を説明する．

1.2　周期性に基づくコミュニティ QA の話題分析

Web 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．商用の Web 検索システムにおいても，検索エンジンに入力されるクエリには，時間的な周期性（季節性）を持つことが知られている [17][15]．CQA においても，Web 検索エンジンと同様，投稿される質問記事の内容には，周期性が存在しているのではないかと考えることができる．周期性の持つ話題を的確に抽出できれば，より時期依存性の高い拡張クエリを作成することが可能になる．

CQA の質問記事は自然言語で記述されているため，同一のキーワードが使用されている質問記事であっても，質問で議論されている話題は異なる場合が多い．例えば，PC カテゴリに投稿された “年賀状” という単語が含まれる質問として，“PC による年賀状作成に最適なソフトは?”，“年賀状の CD が入ったまま取り出せなった。どうすればいいですか？” という質問記事が存在する．前者は年賀状が主題の質問である．後者は年賀状という単語は入っているが，主題は CD ドライブに関する質問である．このように，同じ “年賀状” が使われている質問記事であっても質問記事で議論されている内容は全く異なる場合が多く存在する．そのため，Web 検索クエリのように単純なキーワードの出現頻度では，質問内容の周期性を詳細に分析することができない．

本論文では，CQA でのユーザの情報要求の変化をより詳細に分析するために，時系列トピックモデルを用いた話題変動の抽出，分類を行う．トピックモデルは，文書と単語の間には

[*1] http://chiebukuro.yahoo.co.jp/
[*2] http://oshiete.goo.ne.jp/

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．このため，同じ話題で使用される単語は，確率分布内で，近い確率を持つようになるという特徴がある．トピックモデルによって生成される “トピック” をユーザの情報要求の “話題” として扱う．CQA の質問記事は自然言語で記述されており，Web 検索クエリよりも，多くのキーワードを含むため，トピックを形成することによる “話題” の抽出に適している．また，時系列トピックモデルでは，同一トピックであっても，分割時刻毎に異なる単語の確率分布を持つ．そこで，トピックの確率分布が時刻毎にどの様に変化していくのかを追跡していくことで，話題の変化を追跡する．抽出した話題変動に対して，周期性に着目し，フーリエ変換を用いた周波数解析手法により周期性の観点から話題変動のタイプを分類する．

1.3　コミュニティ QA を用いたクエリ拡張型 Web 検索システム

本論文では，ユーザが想起できない情報要求を言語化し，多様な Web 検索を実現するものとして，自然言語で記述した情報要求を付与した “質問記事付き拡張クエリ” を提案する．自然言語はユーザにとって理解しやすいという特徴がある．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，“言語化された情報要求” 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．

CQA は気軽に誰でも質問記事を投稿できる仕組みを提供しており，Yahoo!知恵袋では，2004 年 4 月から 2009 年 5 月までの 5 年間で 16,257,413 件の質問記事が投稿されている．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，ファセット検索型の Web 検索システムを提案する．ファセット検索は，Web 検索結果を時節やドメインなどの様々な “切り口” であるファセットごとに分類して提示する検索手法である．ファセット検索では，情報がファセットによって整理された状態で提供されているため，多様な情報であっても，ユーザは混乱することなく，自身の検索意図に近いファセットでの検索結果を選択することができる．本論文では，質問記事付き拡張クエリの提示にファセット検索を適用し，ファセットごとに異なる拡張クエリを提示する．ファセットには，CQA に付与されている話題の “カテゴリ” と質問記事の投稿日時による “投稿の季節” を用いる．CQA では，質問記事を投稿する際，必ず一つのカテゴリを選択する必要がある，また投稿時期は質問の内容に影響していると考えられるため，カテゴリと季節は検索のコンテキストであるといえる．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．

CQA の質問記事を用いることで，従来の拡張クエリの課題であった，拡張クエリの検索意図の特定と，多様性を両立する．“質問記事拡張クエリ” では，クエリの背後に存在する情報要求を，自然言語の質問記事により言語化する．また，CQA の特徴であるカテゴリや季節性を用いて，コンテキストごとに異なる拡張クエリを提示することで多様性を実現している．

1.4　本論文の構成

本論文では，まず 2 章で本論文に関連する先行研究について述べ，本研究の位置づけを明らかにする．3 章では，コミュニティ QA に投稿される質問記事を周期性の観点に基づいて分析する．CQA の単語の出現確率を用いた手法，時系列トピックモデルを用いた手法により，CQA から周期性を持つ話題を抽出する．次に 4 章で，CQA を用いた質問記事付き拡張クエリによる Web 検索システムの，実装と評価に関して説明する．3 章での結果を元に，CQA の季節性を用いて，季節ごとに異なるクエリを提示する．最後に 5 章で，本研究の成果をまとめて，今後の展望を述べる．

第 2 章

関連研究

2.1 Web 検索におけるクエリ拡張に関する研究

本研究では，コミュニティ QA の質問記事を用いたクエリ拡張により，Web 検索支援を行う．クエリ拡張は多くの研究がなされていると共に，商用の検索エンジンでも提供されており，Web 検索エンジンにおいて重要な機能となっている．商用の Web 検索エンジンでは，検索エンジンに入力されたクエリログを元にクエリ拡張を行なっている．クエリ拡張に他の Web ページなどの外部リソースを使用することも有用であることが多くの研究で示されている．Yin ら [18] 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 [30] 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．堀ら [24] 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．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 [26] 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拡張クエリを作成できるとしている．

以下では，本研究に関連と関連の深いコンテキストアウェアな Web アクセスを実現するためのクエリ拡張，拡張クエリの根拠を提示する研究に関して述べる．

2.1.1 コンテキストアウェアなクエリ拡張に関する研究

本研究では，CQA のカテゴリと投稿の季節の検索の切り口（ファセット）として，ファセットを切り替えることにより，様々なコンテキストに対応した拡張クエリを提示している．ユーザのコンテキストを推定し，コンテキストに対応したクエリを提示する手法では，Cao ら [4]

や Semgstock ら [14] の研究がある．Cao ら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ている．

ファセット検索は単純な検索クエリから多様な Web 検索を実現できる手段として，探索的情報検索分野において，多くの研究がなされている．Jonathan ら [9] は，ファセット検索とユーザプロファイルと組み合わせることにより，個人の趣向に合わせたインタラクティブな情報検索を実現する手法を実現している．Hearst[8] は，ファセットを階層構造化させる Web 検索インターフェースを提案している．廣嶋ら [27] 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る．多義的なクエリを推薦する手法として今井ら [25] 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，話題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．

2.1.2　拡張の根拠を提示するクエリ拡張手法に関する研究

本研究では，CQA の質問記事を拡張クエリと共に提示することにより，拡張クエリの根拠となる情報要求を提示する．クエリの意味や根拠を提示する手法については，Guo ら [6] や，Lin ら [10] 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 [13] 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共に提示する手法を Zha ら [20] 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．

2.2　コミュニティ QA と Web 検索に関する研究

CQA は，ソーシャルベースの新たな情報アクセス手段として注目を集めている．CQA 内で議論される話題に関する研究は，Adamic ら [1] の研究がある．Adamic らは，CQA での話題の最も基本的な単位であるカテゴリに着目し，カテゴリごとにコミュニケーションのタイプが異なることを明らかにしている．また，Miao ら [12] は，CQA の中から新たなカテゴリとなる話題を，トピックモデルによって発見する手法を提案している．

CQA と既存の情報アクセス手段である Web 検索と結びつけることで，より高度な情報検索を実現する研究がなされている．Liu ら [11] は Web 検索ユーザが CQA ユーザとなるまでの経過を分析している．CQA ユーザは質問記事を投稿する際に，Web 検索結果中の CQA ページを閲覧してから利用するケースが多く，また，Web 検索クエリに関しても，より具体的なクエリを投稿する特徴があることを明らかにしている．Yoon ら [19] 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 [22] 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．高田ら [21] 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．

2.3　Web の話題変動と周期性に関する研究

Web 検索におけるクエリの周期性を扱った研究では，Vlacos ら [17] は，クエリログから周波数解析を用いてクエリの出現頻度には周期性が存在することを明らかにし，またクエリのバーストを発見する手法を提案している．村田ら [31] は，検索意図をクエリ入力直後に最初にクリックした Web ページであるとして，クエリに対して，クリックログにより取得した URL の出現頻度の周期的な傾向を分析している．その結果，同じクエリであっても，最初にクリックされる URL は週単位で異なり，かつ周期性が存在していることを示した．Shokouhi ら [15] は，過去のクエリログのデータから，指数平滑法により現時刻でのクエリの出現頻度を予測することにより，時間に適合するクエリ拡張を提案する手法を示した．この手法では，周期性や季節性を持つクエリを効果的に推薦できることを明らかにしている．Efron[5] は，時系列情報を持つ文書コレクションに対して，TF・IDF の拡張として文書の重みに時間情報を組み込む手法を提案している．

時系列上での話題の変化を扱った研究も知られている，芹澤ら [28] の研究は，ニュース記事を日時単位で分割し，LDA[3] によってトピックを抽出，コサイン類似度によってトピック間の類似度を求めている．それにより時間単位でのトピックの話題変化を追跡できるとしている．岩田ら [23] は，購買情報のデータから，時間変化するユーザの興味や流行などを発見するための時系列上でのトピック追跡モデルを提案している．

2.4　本研究の位置づけ

本研究では，クエリ拡張のための外部リソースとして，CQA の質問記事を用いる．質問記事本文を拡張クエリの根拠となる情報要求とみなし，“質問記事付き拡張クエリ”として質問記事と拡張クエリをセットで提示する．また，CQA に投稿される質問記事の話題変化の周期性を特定し，季節ごとに異なるクエリを提示する．

検索のインターフェースとして，CQA のカテゴリと投稿の季節をファセットとし，2 次元のタブにより，カテゴリと季節を切り替えることにより多様なクエリを提示するファセット検索型のクエリ拡張インターフェースを提案する．

以上のように，本研究では，CQA リソースを用いて，多様な検索を実現するためのファセット検索，クエリの理解を支援するためのクエリの意味提示を実現させることで，クエリ拡張において重要な多様性とユーザ理解を両立させている．一つの情報リソースから複数の役割を持つクエリ拡張手法は，従来研究では提案されておらず，本研究の特徴的な点である．

第 3 章

周期性に基づく コミュニティ QA の話題分析

3.1 はじめに

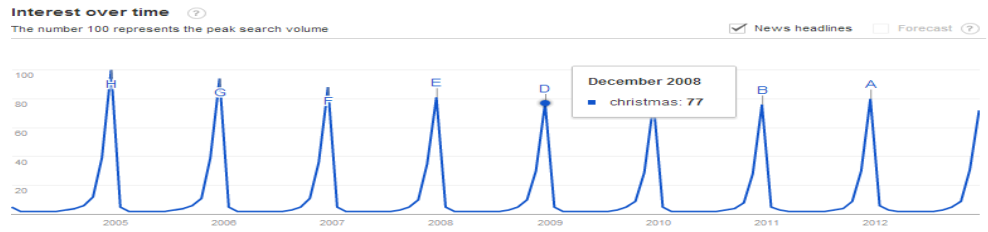
インターネットの普及に伴い，Web ユーザは生活に密着した情報に関しても気軽に検索や質問記事の投稿を行なうようになった．生活に密着した情報に関しては，アクセスする時期や時節の影響を大きく受けることがわかっている．GoogleTrend[*1] では，検索クエリの使用頻度から検索のトレンドの推移を可視化するサービスを提供している．図 3.1(a) は，GoogleTrend で “Chirismas” を検索した時の検索トレンドの推移を示している．クエリ “Christmas” は毎年 12 月に検索頻度が高く，その他の期間はあまり検索されない．このように，Web ユーザの情報要求の変化は周期性を持つことがわかる．

本章では，コミュニティ QA の質問記事の質問内容の周期性を明らかにする．質問内容に周期性がある場合，それらの周期にあった質問記事を用いて拡張クエリを作成することでよりユーザの状況に適合するクエリが推薦できる．本章ではまず，カテゴリ毎に投稿された質問記事内の単語の出現頻度の推移を追跡する．次に，質問記事をトピックモデルにより “話題” 単位にクラスタリングし，話題の変化を追跡する．最終的に，CQA の中から，周期性を持つ話題のみを抽出する手法を提案する．

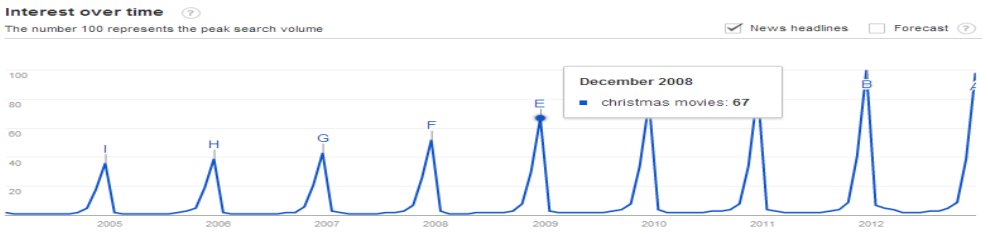
3.2 単語の出現確率推移による話題変動抽出

CQA の質問記事は自然言語で記述されているため，質問記事の話題の変動を確認するには，質問記事中の単語の使用頻度を分析することが最も一般的である．本節では，CQA のカテゴ

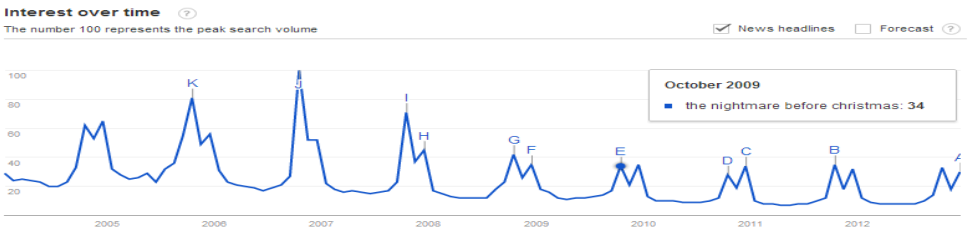
[*1] http://www.google.com/trends/



(a) Christmas



(b) Christmas Movies



(c) The Nightmare Before Christmas

図 3.1　GoogleTrend でのクエリのトレンド変化

リごとに，質問記事中で使用される単語の出現確率の推移により CQA で周期的に発生する話題を抽出する．

各カテゴリにおいて，CQA 質問記事を MeCab[*2]により形態素解析し，月ごとの単語の出現頻度をカウントする．月ごとの投稿質問記事数にはバラつきがあるため，ある月の出現頻度を，投稿された総質問記事数により正規化して出現確率とすることで，単語の使用頻度の推移を評価する．カテゴリ C において，1 ヶ月間に投稿された質問記事のうち，単語 w の出現確率は以下で与えられる．

$$P_{C,w} = \frac{N_{C,w}}{N_C} \tag{3.1}$$

$N_{C,w}$ はカテゴリ C に投稿された質問記事のうち単語 w を含む質問記事数，N_C はカテゴリ C にひと月に投稿された総質問記事数である．

[*2] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

3.2.1　分散とバーストによる周期性を持つ単語抽出

本章では，図 3.1 のように，周期的に使用頻度が変動する単語を CQA において発見することを目的としている．周期性を持つ単語は，ある月では高い出現確率を示すが，他の月では，出現確率が低い単語であると考えられる．そこで，各単語の出現確率の分散を月ごとに計算し，分散の大きい単語を，季節性を持つ単語として抽出する．しかし，出現確率は非常に小さい値をとるため，単純な分散では，単語間の比較を行うことができない．そこで，分散を出現確率の平均値で正規化する変動係数を用いる．元々の出現確率が高い頻出語と稀にしか出現しない単語の分散を変動係数によって比較することで，ある特定の月に偏在する単語を抽出する．変動係数 C.V は以下の式で与えられる．

$$C.V = \frac{\sqrt{\sigma^2}}{\overline{x}} \tag{3.2}$$

このとき，$\sqrt{\sigma^2}$ は出現確率の標準偏差，$\overline{x}$ は平均を示している．変動係数を各単語で計算し，降順に並べたとき，上位の単語を周期性を持つ語として抽出する．

図 3.1 の (a)“Christmas” は，毎年 12 月のみ検索頻度が急上昇し，その他の月では，ほどんど検索されない．このようにある特定の期間のみ，頻度や確率が上昇する現象を，本論文ではバーストと呼ぶ．バーストは，Web 上の様々な情報で発生するとされている．Vlachos ら [17] らは，クエリログから Web クエリのバースト性を検証しており，山家ら [29] は，Vlachos らの手法を用いて，ソーシャルブックマークのバーストとその周期性を発見している．本研究においても，CQA のバーストを発見するために Vlachos らの手法により，単語の出現確率推移からバーストを抽出する．Vlachos らのバースト抽出手法を以下に示す．

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 $std(MA_w)$ から閾値 $cutoff = mean(MA_w) + x * std(MA_w)$ 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

3.2.2　単語の出現確率推移による話題遷移抽出結果

本節では，3.2.1 節で提案した，分散とバースト抽出による周期性を持つ単語抽出手法を実際の Yahoo!知恵袋データに適用した実験結果を示す．まず，提案手法の妥当性を評価するために，PC カテゴリにおいて，以下の 3 つの手法で単語抽出を行い，結果を比較する．

- **Freq** :単語の出現確率の平均値を降順に抽出
- **C.V** :単語の出現確率の変動係数を降順に抽出
- **C.V+Burst** :変動係数（C.V）の結果に，バースト抽出によるバーストの周期性を考慮

期間は，2006 年 1 月から 2008 年 12 月までの 36 ヶ月間で実験を行う．実験で用いたカテゴリは，毎月安定して投稿質問数が多かった PC カテゴリ，旅行カテゴリ，恋愛カテゴリで実験用いる．3.2.1 節で説明したバースト抽出手法のパラメータを，データ系列 t を 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，移動平均の長さ w を季節の区切りと対応させ $w = 3$ と設定した．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．本論文では，周期性のある単語を抽出するため，データセット 3 年間で，毎年同じ月にバーストが発生した単語のみを抽出する．

3 つの手法により，抽出できた上位の単語を表 3.1 から表 3.3 に示す．表 3.1 は，PC カテゴリにおいて出現確率 36 ヶ月間の平均が上位の語である．“教える”，“使う” など CQA 一般に使用頻度の高い語の他，PC カテゴリの特徴的な語として “パソコン” が抽出されている．図 3.2 は，表 3.1 の単語の 36 ヶ月の出現確率の推移である．どの単語もほぼ同じ確率推移をしていることがわかる．2007 年 3 月にすべての単語の出現確率が低下しているがその他は，大きな変動は発生していない．表 3.2 は単語の出現確率の変動が上位の語である．図 3.3 にこれらの単語の出現確率推移を示す．“年賀状” は毎年 11，12 月ごろに高い出現確率を示すが，その他の月ではほとんど出現していない．“ボーダフォン” は 2006 年は比較的出現確率が安定して推移しているが，2006 年 10 月ごろから一気に出現確率が低下し，2007 年以降はほとんど出現しない語となっている．“湿る” は毎年 6 から 8 月にかけて出現確率が上昇するが，“年賀状” よりもゆるやかな変化となっている．“流出” は，2006 年 5 月に一度出現確率が急上昇しているが，その他の期間はほとんど出現しない語である．表 3.3 は変動係数での順位付後，バースト抽出により周期的なバーストが発生している単語のみを抽出した結果である．表 3.2 から “ボーダフォン”，“流出” が表外となり，“除”[*3]，“4 月” が新たに上位で抽出されている．図 3.4 に，C.V + Burst 法により抽出された単語の出現確率の推移を示す．“湿る” と “除” はほぼ同じ確率推移となっている．また，“4 月” は緩やかながら毎年 3 月から 5 月に出現確率

[*3] “除く”，“除湿” などの単語の一部が抽出されたものであると考えられる

が上昇している．

変動係数とバースト抽出による単語抽出手法（C.V+Burst）を他のカテゴリにおいても適用した．ここでは，旅行カテゴリと恋愛カテゴリにおいて抽出できた単語を示す．表 3.4 は旅行カテゴリにおいて抽出できた単語，図 3.5 は，それらの出現確率推移である．“2 月”，“4 月”という特定の月を示す単語の他，“雪”，“GW”といった語が出現している．出現確率推移では，期間はズレているが，“年賀状”と同様，出現確率が急激に上昇するという傾向を示している．恋愛カテゴリでの単語抽出結果を表 3.5，出現確率推移を図 3.6 に示す．抽出できた単語は，バレンタインデーに関する単語であることがわかる．抽出できた全ての単語が毎年 1〜3 月に出現確率が上昇している．

表 3.1　出現確率平均上位の単語（Freq）

順位	単語	出現確率（平均値）
1	教える	0.233
2	できる	0.211
3	使う	0.146
4	パソコン	0.138

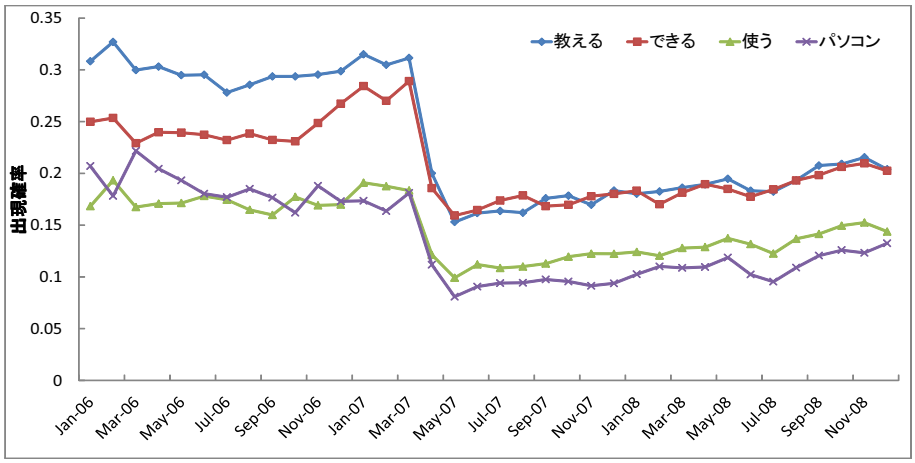


図 3.2　Freq により抽出された単語の出現確率推移

表 3.2　変動係数上位の単語（C.V）

順位	単語	変動係数
1	年賀状	1.93
2	ボーダフォン	1.26
3	湿る	1.20
4	流出	1.17

表 3.3　変動係数とバースト抽出による単語（C.V+Burst）

順位	単語	変動係数	バースト発生月
1	年賀状	1.93	11 月，12 月
2	湿る	1.20	6 月，7 月
3	除	1.13	6 月，7 月，8 月
4	春	0.98	3 月，4 月

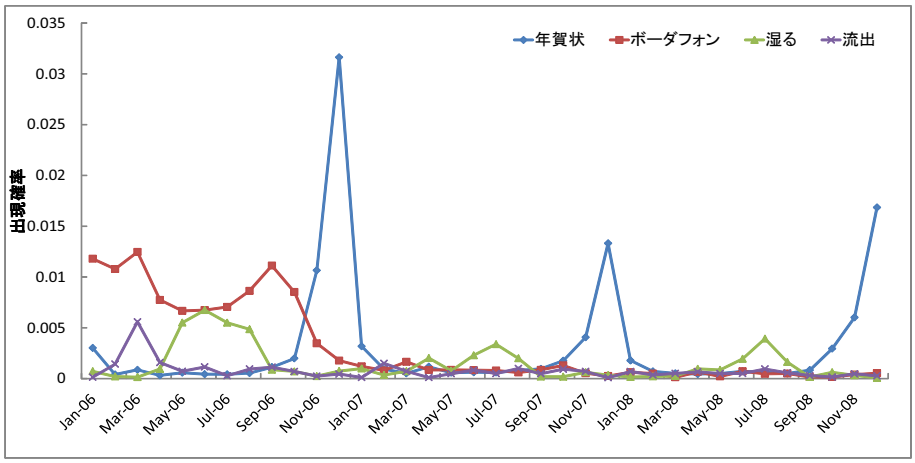


図 3.3　C.V により抽出された単語の出現確率推移

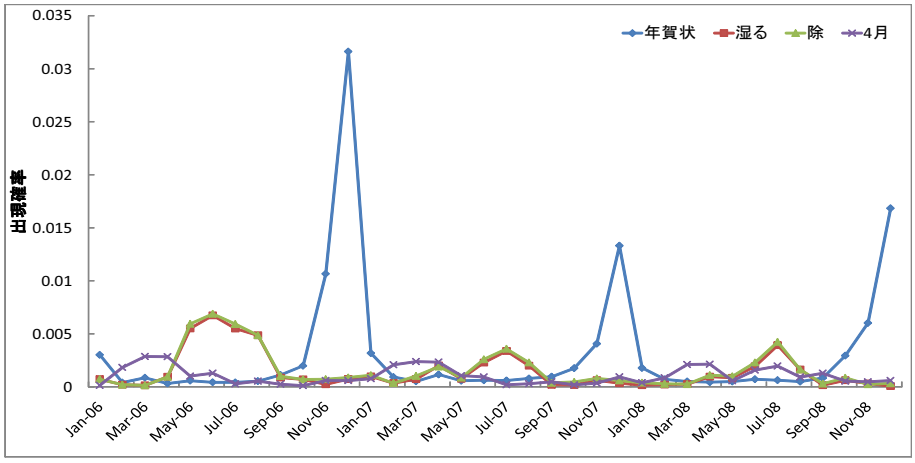


図 3.4　C.V+Burst により抽出された単語の出現確率推移

表 3.4　旅行カテゴリでの単語抽出結果

順位	単語	バースト発生月
1	2 月	1 月，2 月
2	GW	4 月
3	雪	1 月，2 月
4	4 月	3 月，4 月

表 3.5　恋愛カテゴリでの単語抽出結果

順位	単語	バースト発生月
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

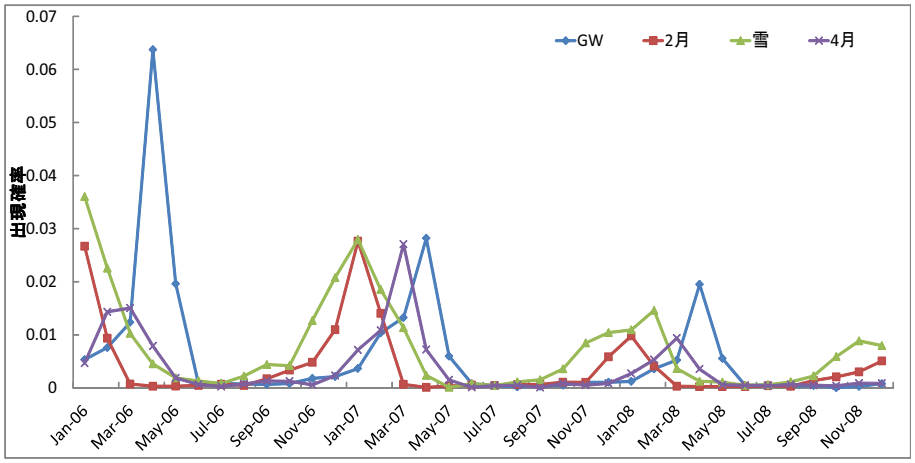


図 3.5　旅行カテゴリでの単語の出現確率推移

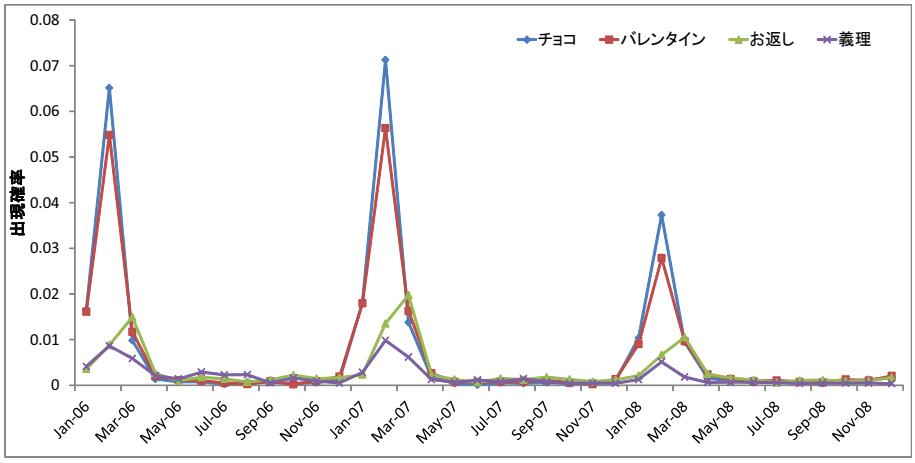


図 3.6　恋愛カテゴリでの単語の出現確率推移

3.3　トピックモデルによる CQA の話題変動抽出

3.2 節では，質問記事を bag of words として見て，bag of words 内の一つの単語のみに注目することで CQA から周期的な話題に関する単語を抽出した．しかし，特定の話題は一つの単語で決定するとは限らない．図 3.1 の例では，(a) “Christmas” というクエリは，毎年 12 月に多く使用されるクエリであることを示しており，“Christmas” は季節性を持つといえる．また，(b) “Christmas Movies” も同様に，毎年 12 月に多く利用されるクエリであるが，その他の月ではほとんど使用されておらず，季節性の強いクエリであることがわかる．しかしながら，“Christmas Movies” の一つである (c) “The Nightmare Before Christmas”[*4] というクエリはこれらの傾向と異なり，1 年間で 10 月と 12 月の 2 回ピークが存在し，その他の月も比較的多く使用されている，比較的周期性が弱いクエリであることがわかる．このように，季節性を持つキーワードが使用されるクエリであっても，キーワードの組み合わせによっては，季節性が弱いクエリとなる場合が存在する．

CQA においても同様の傾向があり，PC カテゴリでは，“年賀状” という単語が毎年 11，12 月に集中して投稿されることが明らかになった．一方で，“プリンタ” に関する話題は一年間を通して数多くの質問記事が投稿されている．しかし，“プリンタ　年賀状” に関する質問記事は，毎年 11，12 月に集中して投稿される．特に，CQA の質問記事は自然言語で記述されているため，Web クエリと違い一つの質問記事に多くの単語が出現する．以上のことより，CQA におけるユーザの情報要求の変化を的確に捉えるためには，単一キーワードに限定せず，キーワード集合からなる “話題” 単位で分析することが有用であると考えられる．

本節では，CQA でのユーザの情報要求の変化をより詳細に分析するために，時系列トピックモデルを用いる．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．このため，同じ話題で使用される単語は，確率分布内で，近い確率を持つようになるという特徴がある．トピックモデルによって生成される “トピック” をユーザの情報要求の “話題” として扱うことで CQA の質問記事を話題単位で分析することができると考える．

以降は，まず 3.3.1 節で，時系列トピックモデルによる CQA 質問記事の “話題化” と話題変動を追跡するための手法について説明し，3.3.2 で話題変動の例を示す．そして，3.3.3 節で，周波数解析を用いることによる CQA からの周期性を持つ話題の抽出手法について説明し，3.3.4 節で評価実験について詳述する．

[*4] 1993 年に公開された映画のタイトル

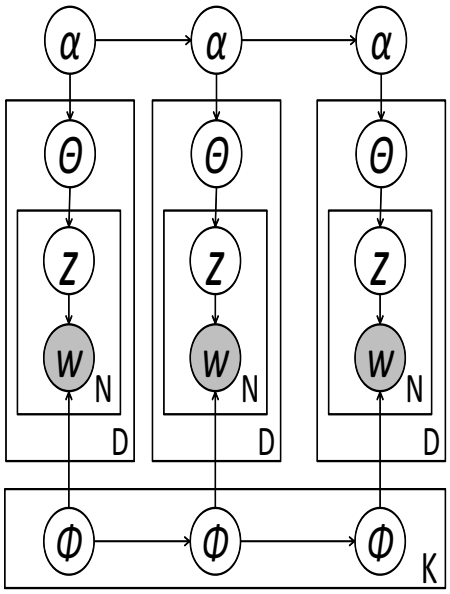


図 3.7　DTM のグラフィカルモデル (時間分割数 3 のとき)

3.3.1　時系列トピックモデルによる話題変動抽出手法

時系列トピックモデル

CQA の話題は，カテゴリにより分類されているが，各カテゴリ内では，より詳細かつ多様な話題が展開されている．そこで，カテゴリ内の話題をより詳細に分析するため，話題の基本単位となるトピックを，トピックモデルによって生成する．トピックモデルでは，文書集合によって作成された文書 - 単語空間を，文書 - トピック空間，トピック - 単語空間に分割する．各トピックは単語の確率分布 (単語分布) で表現されている．本論文では，Blei ら [2] が提案した，Dymamic Topic Model(DTM) を用いる．DTM のグラフィカルモデルを図 3.7 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 3.7 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ことになる．

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．

JS ダイバージェンスによるトピックの話題変動抽出

DTM により生成したトピックの話題遷移を追跡するために，JS ダイバージェンスを用いる．JS ダイバージェンスは，2 つの確率分布同士の類似度を算出する手法であり，確率分布 P と Q から次式で与えられる．

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

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

DTM では，同一トピック内で，時刻毎に確率分布を生成する．時刻間の確率分布の類似度は，時刻間でどれだけ話題が変化したかを表す，話題の変化量と定義することができる．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大きく変化したことを示している．

時間分割数 TS の時系列データに対して，時系列の最初の時刻を基準時刻 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時刻毎にどれだけ変化していくのかを分析する．

3.3.2　CQA の話題変動抽出結果

3.3.1 節で提案した，話題追跡手法を実際の CQA に適用した結果を示す．実験のデータとして Yahoo!知恵袋データを用いた．期間は 2006 年 1 月から 2008 年 12 月までの 36 ヶ月である．実験に使用したカテゴリは，旅行，PC，経済の 3 カテゴリである．DTM のパラメータは以下の通り設定した．

- トピック数 $K = 200$
- ハイパーパラメータ $\alpha = 0.01$
- 時間分割数 $TS = 36$(一ヶ月刻み)

実験によって抽出できた話題変動の例をカテゴリ毎に図 3.8 から図 3.10 に示す．図 3.8 は PC カテゴリでの話題遷移，図 3.9 は旅行カテゴリでの話題遷移，図 3.10 が経済カテゴリでの話題遷移を示している．各図の横軸は 2006 年 1 月から 2008 年 12 月までの 36 ヶ月間の時系列を表しており，縦軸は JS ダイバージェンス値である．各グラフは 2006 年 1 月の確率分布との類似度であるため，基準時刻である 2006 年 1 月は全トピックの JS ダイバージェンス値が 0 であり，JS ダイバージェンス値の上昇は 2006 年 1 月から話題が遠くなっていることを示している．各カテゴリのいずれのトピックにおいても，JS ダイバージェンス値は徐々に上昇していることがわかる．しかしながら，JS ダイバージェンス値の上昇のパターンはトピック毎に異なっている．

以下では，カテゴリ毎に話題変動の特徴について，詳細に分析する．

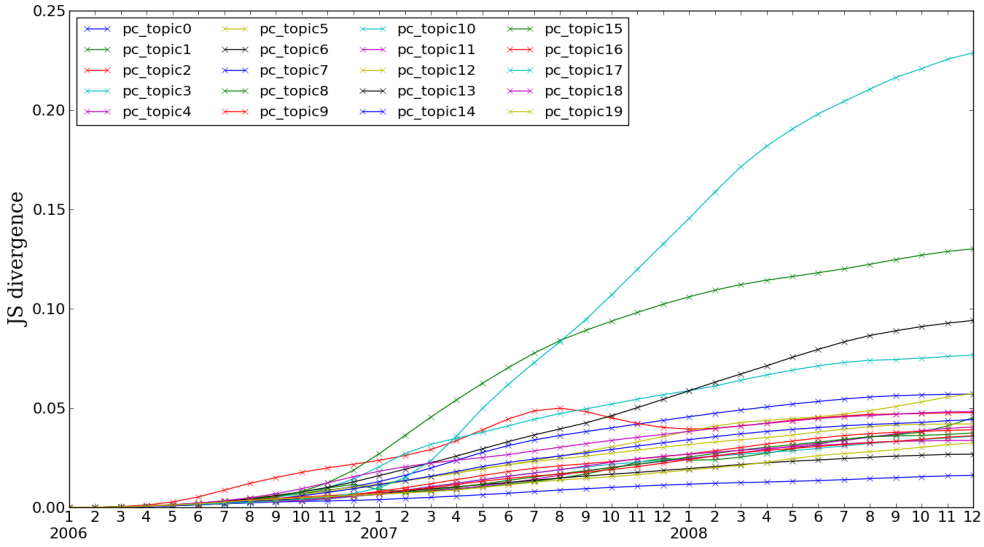


図 3.8　PC カテゴリ全トピックの JS ダイバージェンス（2006 年 1 月基準）

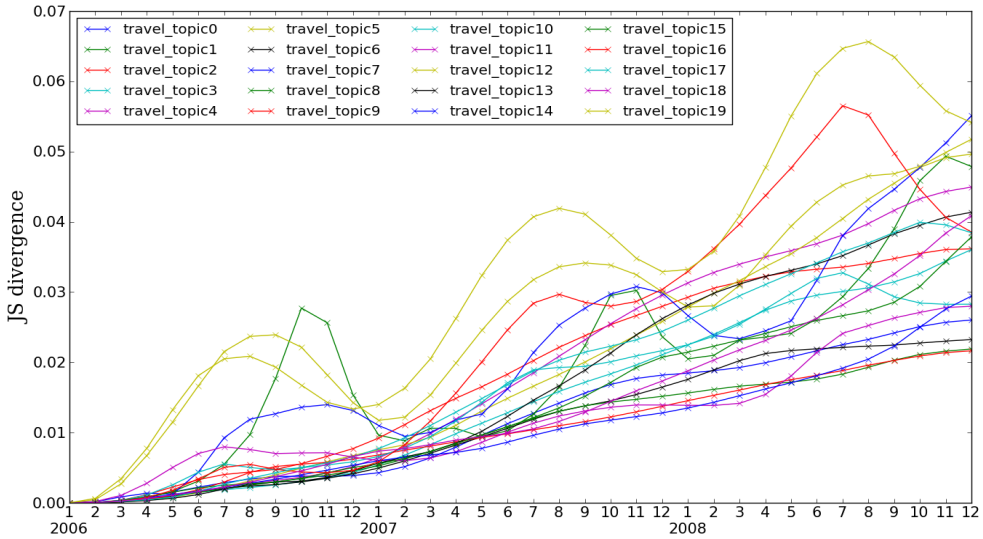


図 3.9　旅行カテゴリ全トピックの JS ダイバージェンス（2006 年 1 月基準）

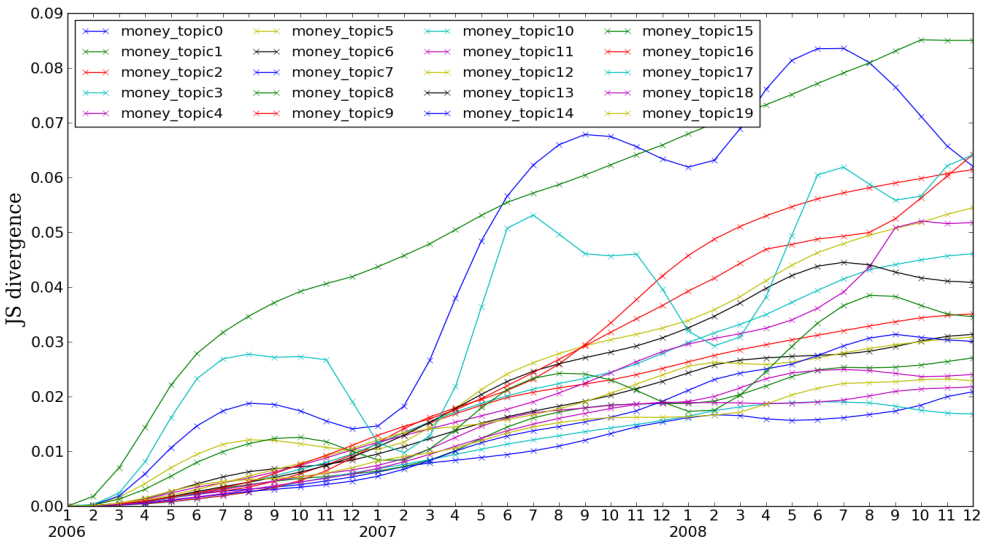


図 3.10　経済カテゴリ全トピックの JS ダイバージェンス（2006 年 1 月基準）

PC カテゴリでの話題変動

PC カテゴリでは，ほとんどのトピックが JS ダイバージェンス値が単調増加のグラフとなっており，JS ダイバージェンスが減少する期間は見られない．図 3.8 の中から，トピック 3 を抜き出したものを図 3.11 に示す．毎月一定の割合で JS ダイバージェンスが増加している．トピックの話題変動の内容を調査するため，各トピックで，時系列上での確率の変動が大きい単語を抽出した結果を表 3.6 に示す．特徴語はトピック全期間の平均確率の上位 5 語，変動語は変動係数 CV の上位 5 語を示している．トピック 3 はパソコン購入に関するトピックであることがわかる．変動語には，“eeepc”，“ミニノートパソコン” などネットブックにカテゴリライズされるパソコンのジャンル名が抽出されている．変動語のトピック 3 内での確率変動の推移を図 3.12 に示す．いずれの語も 2006 年 1 月の時点では，出現確率はほぼ 0 であるが，“ゲーム”，“モニタ”は 2007 年ごろから，“eeepc”，“ミニノートパソコン”，“aspire” は 2008 年ごろから出現確率が上昇している．

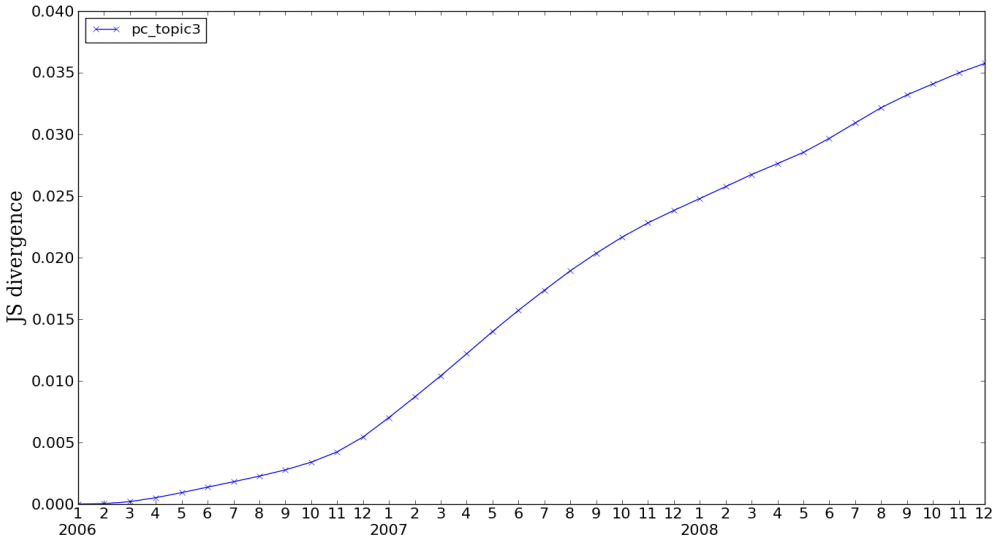


図 3.11　PC カテゴリ，トピック 3 の JS ダイバージェンス推移（2006 年 1 月基準）

表 3.6　PC カテゴリトピック 3 の特徴語と変動語

トピック	特徴語	変動語
3	パソコン 買う ノートパソコン 初心者 新しい	eeepc ゲーム ミニノートパソコン aspire モニター

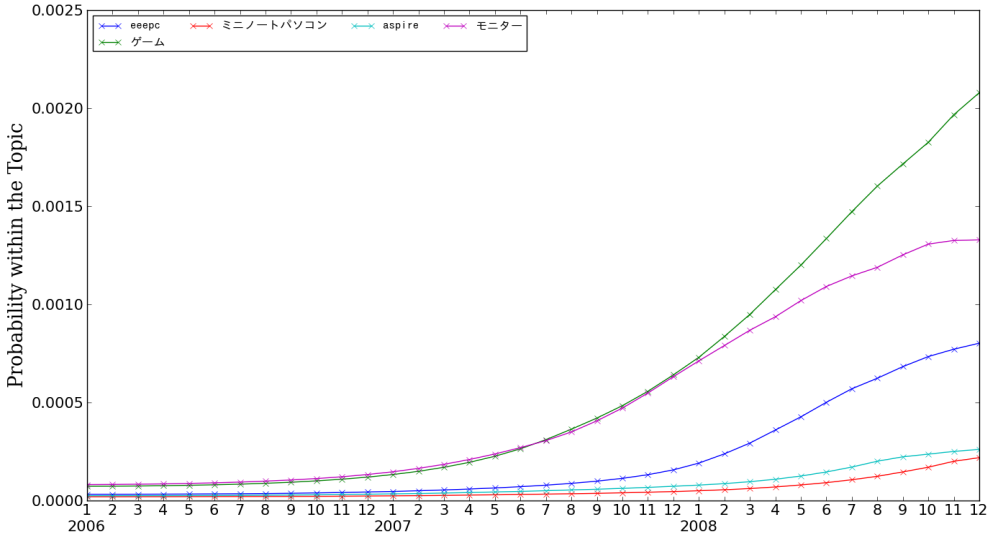


図 3.12　PC カテゴリ，トピック 3 内の単語の確率変動

PC カテゴリでは，ある一時期のみ JS ダイバージェンス値が急上昇する話題変動パターンを発見できた．この話題変動パターンとなった PC カテゴリトピック 2 の JS ダイバージェンス値の推移を図 3.13 に示す．この話題変動では，2007 年 8 月ごろに JS ダイバージェンス値がピークになり，その後は一旦減少した後，緩やかな上昇となっている．トピック 2 の内容語と変動語を表 3.7 に示す．内容語には，“cd”，“音楽”，“ipod” などパソコンで音楽を扱うための用語が抽出されており，変動語には “ニコニコ動画”，“初音ミク” などの語が抽出できた．変動語のトピック 2 内での確率推移を図 3.14 に示す．また，図中に実際に発生した出来事を付与している．“ニコニコ動画” は，ニコニコ動画の実際のサービス開始から発展の過程に対応して出現確率が上昇している．“ニコニコ動画” や “初音ミク” という語において，確率分布の推移とトピック 2 の JS ダイバージェンス推移が対応していることがわかる．

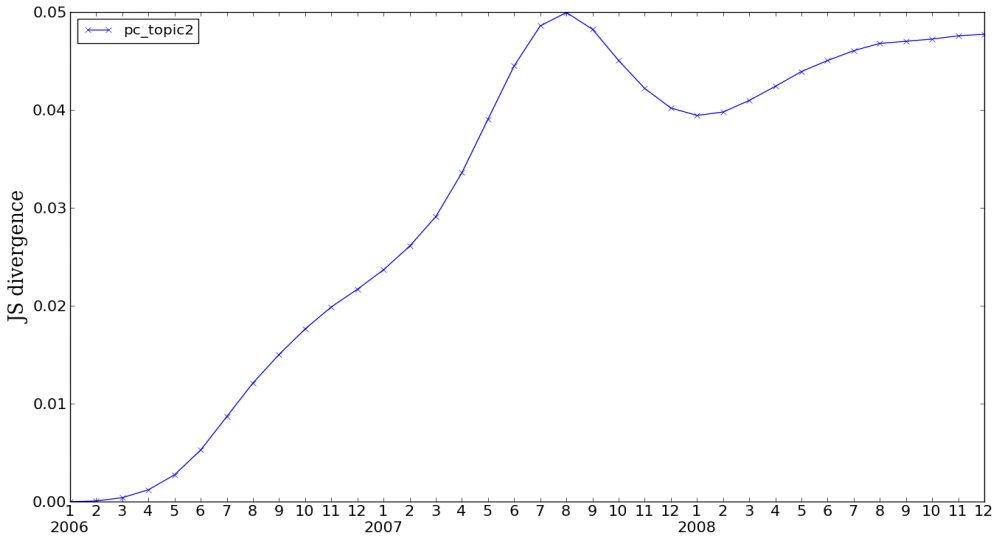


図 3.13　PC カテゴリ，トピック 2 の JS ダイバージェンス推移（2006 年 1 月基準）

表 3.7　PC カテゴリトピック 3 の特徴語と変動語

トピック	特徴語	変動語
2	cd, 音楽, ipod, mp, ニコニコ動画	iphone, 初音ミク, ボーカロイド, ニコニコ動画, dsi

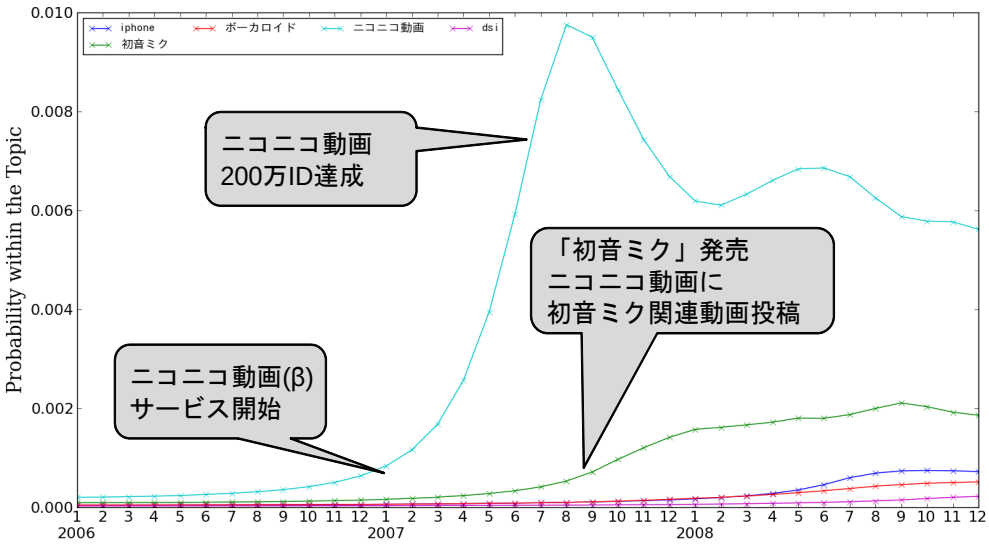


図 3.14　PC カテゴリ，トピック 2 内の単語の確率変動

旅行カテゴリでの話題変動

旅行カテゴリでは，PC カテゴリに見られない話題変動が確認できた．旅行カテゴリトピック 1 の JS ダイバージェンスの推移を図 3.15 に示す．図から，トピック 1 においても，JS ダイバージェンス値は時間の経過とともに上昇していることがわかる．しかし，JS ダイバージェンス値の上昇は単調増加ではなく，増減を繰り返しながら徐々に増加している．JS ダイバージェンス値は話題の近さを表しているため，JS ダイバージェンス値の増減は，2006 年 1 月と近い話題と遠い話題が交互に発生していることを示している．旅行カテゴリトピック 1 の内容語と変動語を図 3.8 に示す．内容語では “京都”，“修学旅行” などの語が抽出されている．変動語では，“紅葉”，“桜” など季節に関連する語が抽出されている．変動語の確率推移を図 3.16 に示す．“紅葉” は毎年 9〜11 月，“桜” は 3，4 月に出現確率が上昇するが，他の季節では低い出現確率になっていることがわかる．

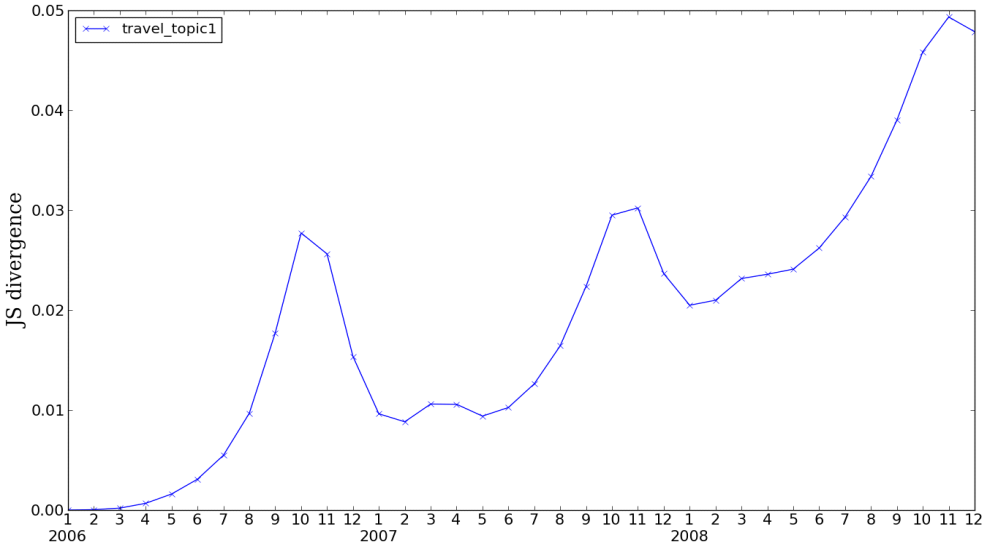


図 3.15　旅行カテゴリ，トピック 1 の JS ダイバージェンス推移（2006 年 1 月基準）

表 3.8　旅行カテゴリトピック 1 の内容語と変動語

トピック	特徴語	変動語
1	京都 見る 修学旅行 桜	紅葉 桜 永観堂 美しい 景色

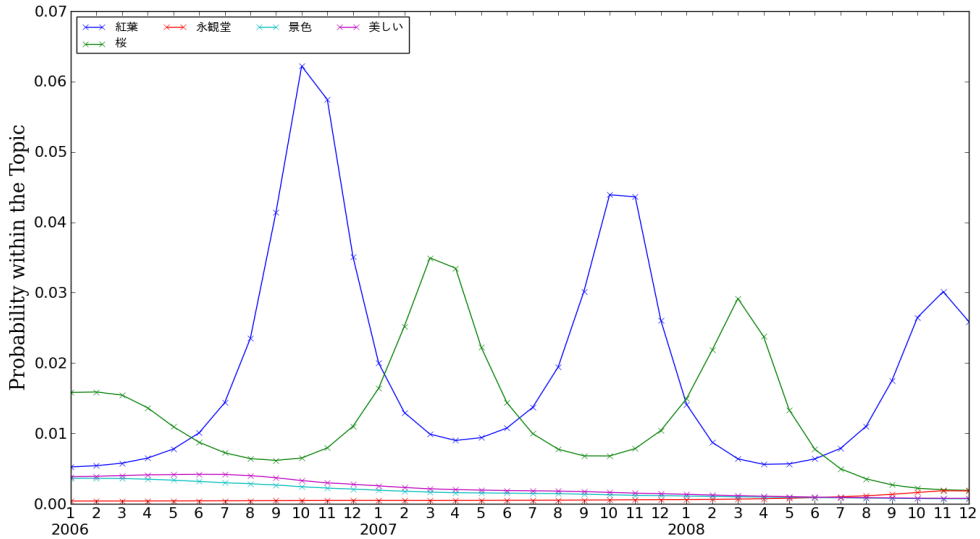


図 3.16　旅行カテゴリ，トピック 1 の単語の確率変動

3.3.3　周波数解析による周期的なトピックの抽出

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パターンを時系列データとして抽出できた．本節では，得られた話題変動パターンから図 3.15 のように周期的な話題変動を発生させるトピックを抽出する手法を説明する．得られた時系列データが周期的な変動を含むデータであることを分析するため，離散フーリエ変換による周波数分析を行う．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の特徴量に変換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．

$$X[k] = \sum_{n=0}^{N-1} x[n] \mathrm{e}^{-j\frac{2\pi}{N}kn} \tag{3.3}$$

ここで，指数関数部分はオイラーの公式を用いて実数部と虚数部に分割でき，以下の式に変換できる．

$$\begin{aligned} X[k] &= \sum_{n=0}^{N-1} \left\{ (x[n] \cos \frac{2\pi}{N} kn) - j(x[n] \sin \frac{2\pi}{N} kn) \right\} \\ &= Re[k] + Im[k] \end{aligned} \tag{3.4}$$

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 $S(x[k])$ は以下の式により計算できる．

$$S(X[k]) = \sqrt{Re[k]^2 + Im[k]^2} \tag{3.5}$$

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．

$$T = \frac{N}{k} \tag{3.6}$$

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 $S(X[k])$ をクラスタリングの素性とする．

階層的クラスタリングは，各要素間の距離行列を使用する．ここで，パワースペクトルを正規化し，パワースペクトル分布を確率分布として表現する．パワースペクトルの正規化は以下の式により実行する．

$$P(X[k]) = \frac{S(X[k])}{\sum_{i=1}^{(N/2)-1} S(X[i])} \tag{3.7}$$

正規化により $\sum P(X[k]) = 1$ を満たすので，トピック間の距離を，式の JS ダイバージェンスによって計算する．

階層的クラスタリングにより，トピックを話題変動の周期的なタイプに基いて分類する．パワースペクトルの確率分布は一般には指数分布に従うため，類似度計算を行った場合，どのトピック間も比較的距離が近い分布となってしまう．そのため，クラスタリングのアルゴリズムは，高い分解能を持つ最長距離法を使用する．最長距離法はクラスタ間の距離を最長となるように要素対を選択していく手法であり，クラスタ A と B の距離は以下の式で定義される．

$$D(A, B) = \max_{a_i \in A, b_j \in B} d(a_i, b_j) \tag{3.8}$$

3.3.4　CQA からの周期的なトピックの抽出実験

Yahoo!知恵袋の旅行カテゴリに，提案手法を適用し，話題変動のパワースペクトル分布をクラスタリングする．データセットは 3.3.2 節と同様とし，DTM のトピック数を 200 に設定して実験を行った．

結果を図 3.17 に示す．また，各クラスタに割り当てられてトピックの詳細を表 3.9 に示す．旅行カテゴリでは，特徴的な 3 つのクラスタに分離された．まず，中央に 183 個のトピックからなる巨大なクラスタ，その右側に 16 個のトピックからなる小規模なクラスタ，そして，最も左側にトピック 15 のみからなるクラスタが存在している．

以下では，それぞれのクラスタについて，特徴的なトピックを例にしてクラスタの特徴を明らかにする．

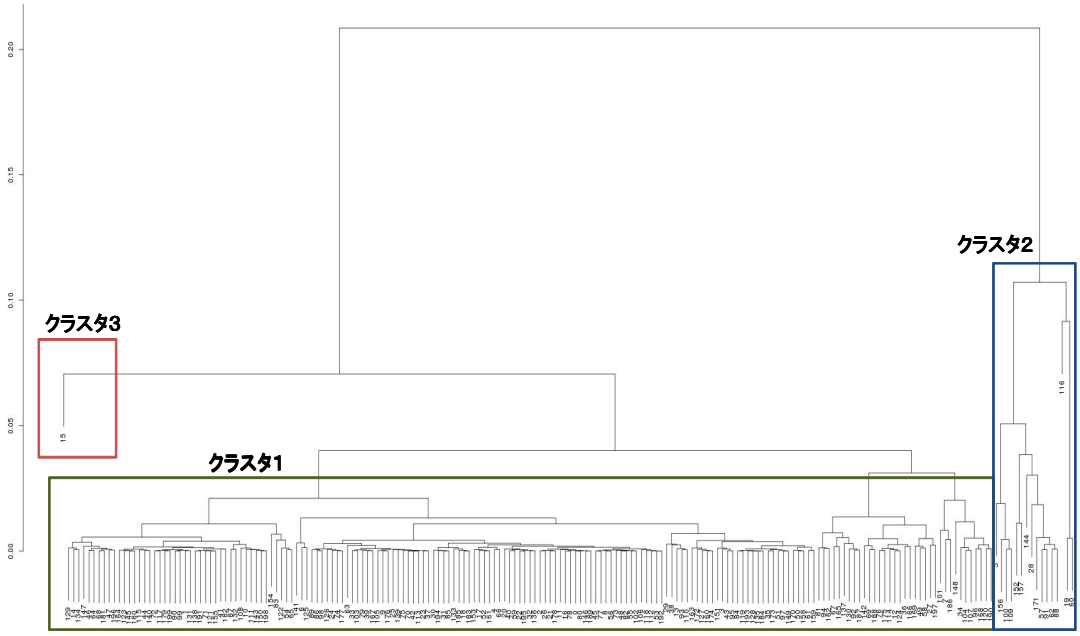


図 3.17　旅行カテゴリ 200 トピックの周波数解析によるクラスタリング結果

表 3.9　旅行カテゴリでの階層的クラスタリング結果（分割数 3）

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

単調変化型トピックの周波数解析結果

旅行カテゴリでクラスタ 1 に属するトピック 0 の話題変動の結果を図 3.18 に示す．横軸は 2006 年 1 月から 2008 年 12 月までの時系列，縦軸は JS ダイバージェンス値である．2006 年 1 月の確率分布を基準としているため，2006 年 1 月の JS ダイバージェンス値は 0 である．その後，時間とともに JS ダイバージェンスが単調に増加している．これは，2006 年 1 月の確率分布から徐々に確率分布が変化してることを示している．図 3.18 の話題変動グラフを離散フーリエ変換により周波数のパワースペクトル分布で表現した結果を図 3.19 に示す．横軸は，$k = 0, 1, ..., 17$ までの周波数，縦軸はパワースペクトルのエネルギーである．スペクトルが $k = 0$ から指数関数に従う分布になっている．

トピック 0 において各月の確率分布の類似度により，クラスタリングした結果を図 3.20 に示す．各要素は，トピック 0 内の月を表しており，括弧内は年を表す．“1(2006)” は 2006 年 1 月の確率分布を表している．どの月においても自身と近い時期でクラスタを形成している．また，同じ月（1 月）であっても，年（2006，2007，2008 年）が異なっている場合，それぞれ遠いクラスタに属している状態になっている．最後に，旅行カテゴリのトピック 0 において出現確率の高い単語の例を表 3.10 に示す．“群馬”，“栃木” など北関東に関連する単語が上位に出現している．

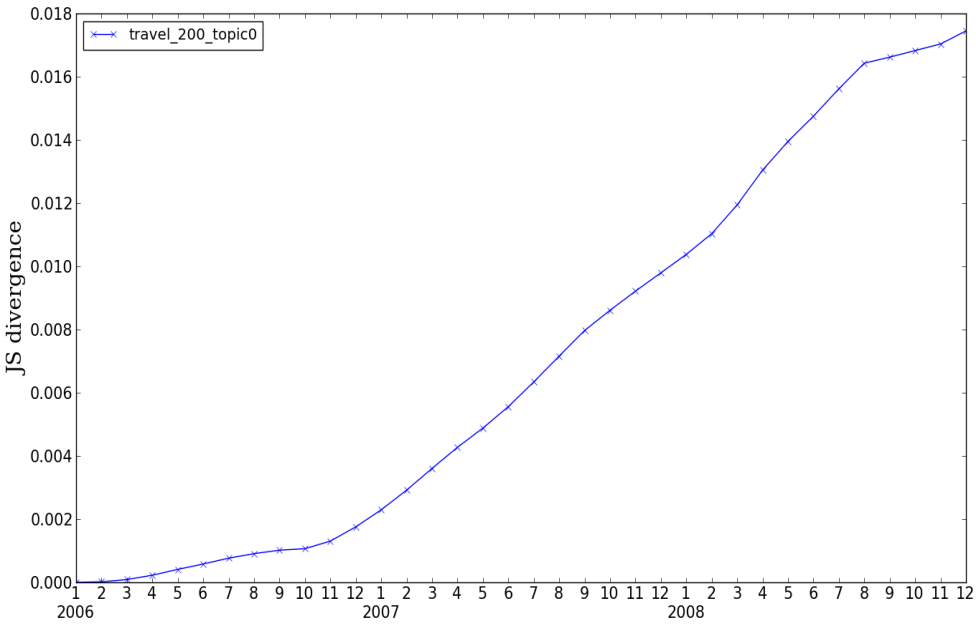


図 3.18　旅行カテゴリ，トピック 0 の話題遷移

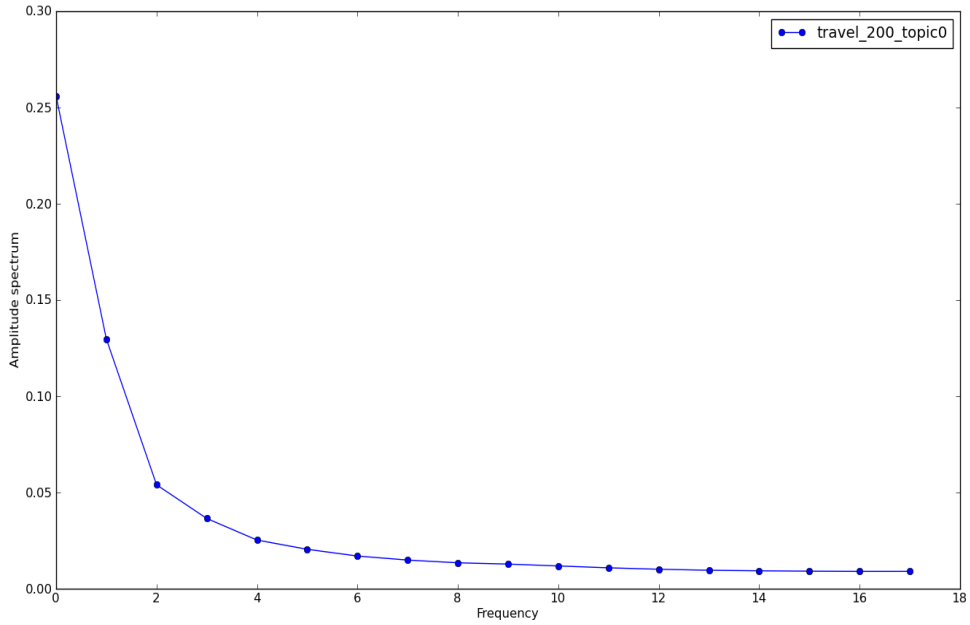


図 3.19　旅行カテゴリ，トピック 0 のパワースペクトル分布

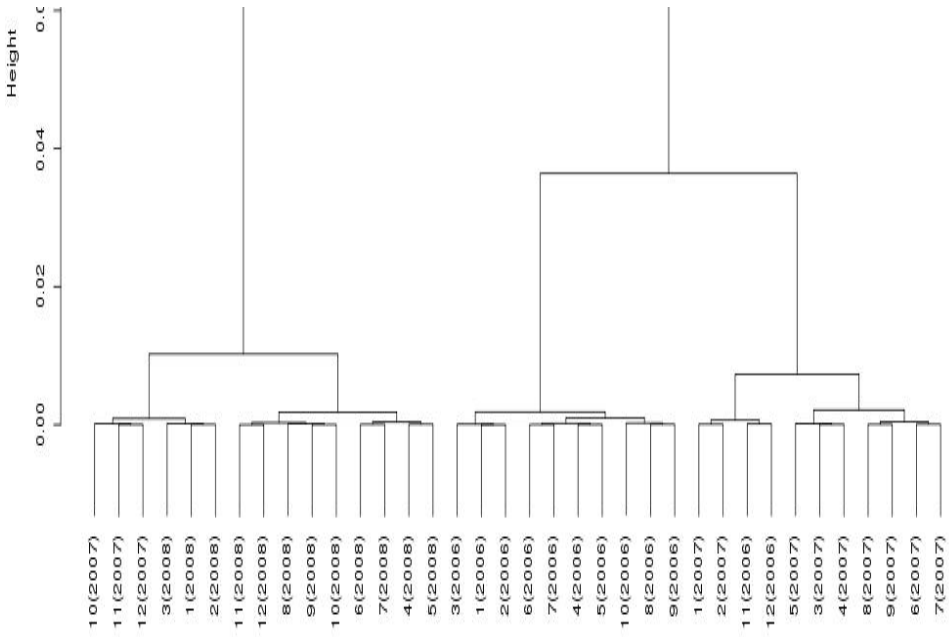


図 3.20　旅行カテゴリ，トピック 0 の各月の単語分布によるクラスタリング結果

周期変化型トピックの周波数解析結果

16 個のトピックから成るクラスタ 2 の例として，トピック 171 の話題変動の結果を図 3.21 に示す．2006 年 1 月から JS ダイバージェンスが徐々に上昇していくが，10 月から低下し，2007 年 2 月まで JS ダイバージェンスが低下した後，再び上昇するパターンを繰り返している．離散フーリエ変換により，図 3.22 のパワースペクトル分布を得る．$k = 3$ の時にスペクトルにピークが立っている．これは，単調変化である図 3.19 とは，大きく異なる特徴である．話題変動は 36 点のデータによって描画されていることから，この話題変動の周期は $T = 36/3 = 12$ である．

トピック 171 の各月の確率分布の類似度から作成したデンドログラムを図 3.23 に示す．2006 年と 2007 年はそれぞれ同じ月は同じクラスタあるいは非常に近いクラスタに属している．トピック 171 において出現確率が上位の単語を表 3.10 に示す．“ヨーロッパ”，“スペイン” など海外に関連する単語の他，“3 月”，“卒業旅行” といった時期を連想する単語が高い出現確率を持っている．

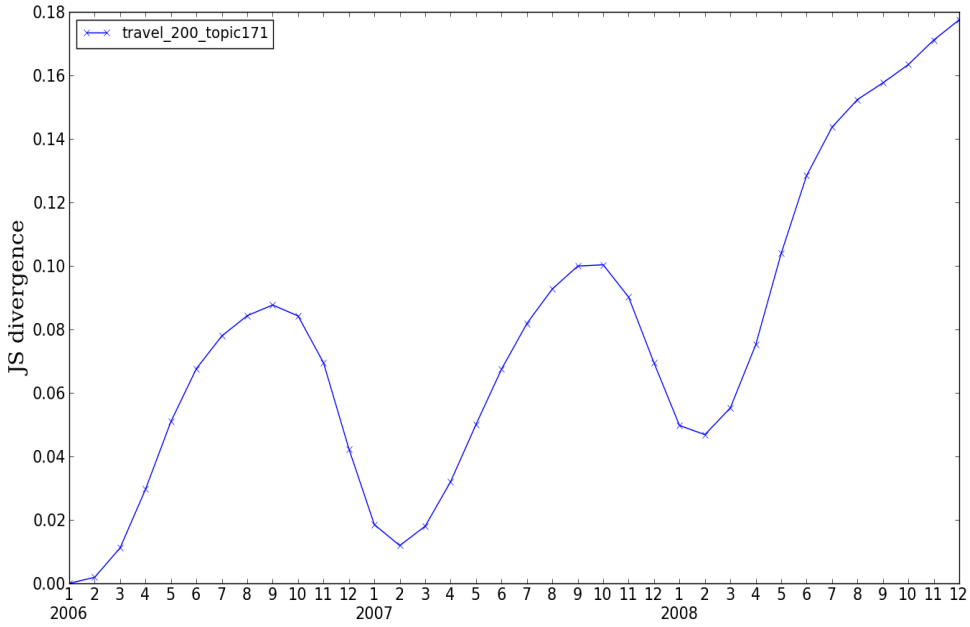


図 3.21　旅行カテゴリ，トピック 171 の話題遷移

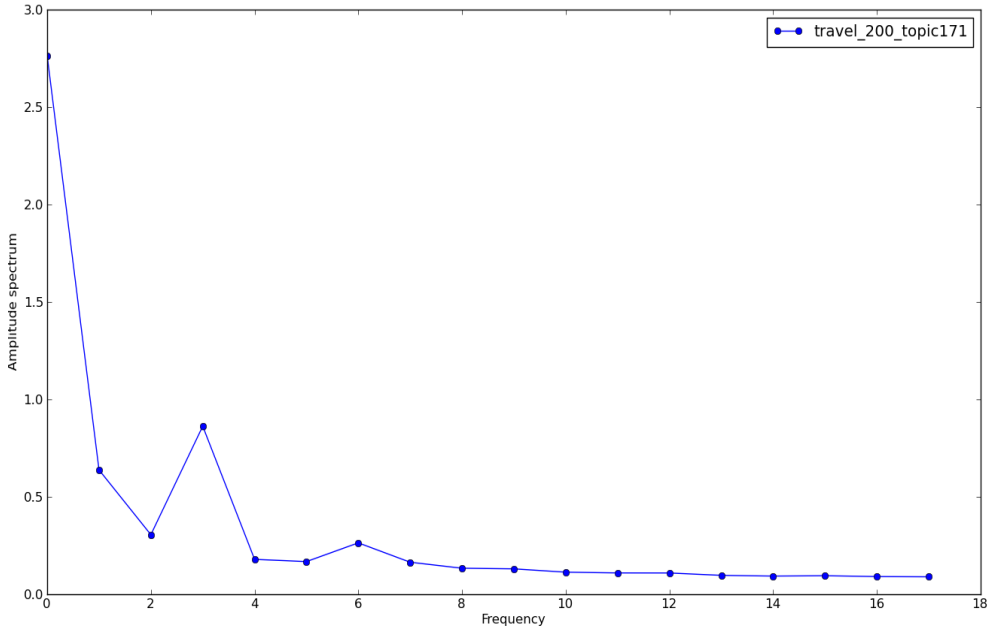


図 3.22　旅行カテゴリ，トピック 171 のパワースペクトル分布

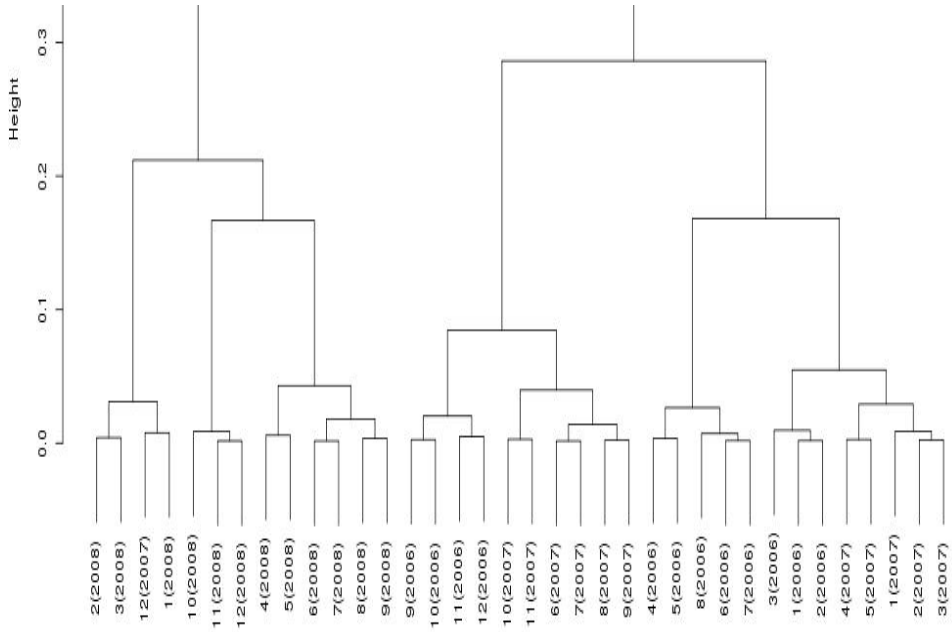


図 3.23　旅行カテゴリ，トピック 171 の各月の単語分布によるクラスタリング結果

バースト型トピックの周波数解析結果

旅行カテゴリ，トピック 15 のみ，他のクラスタから距離のある位置に存在していることがわかる．トピック 15 の話題変動グラフを図 3.24 に示す．2006 年 11 月から 2007 年 4 月にかけて急激な話題変化が見られるが，その後はゆるやかな話題変動となっている．トピック 15 の話題変動から離散フーリエ変換により得られたパワースペクトル分布を図 3.25 に示す．$k = 2, 5$ でピークが発生しており，図 3.19，図 3.22 とは異なるパワースペクトル分布となっている．トピック 15 での各月の確率分布の類似度によって作成したデンドログラムは図 3.26 となる．ここで，2006 年 11 月から 2007 年 4 月までの 7 ヶ月は，独自のクラスタを形成していることがわかる．トピック 15 の代表的な単語は表 3.10 に示す通り，“お薦め”，“名所” といった語が上位にきている．

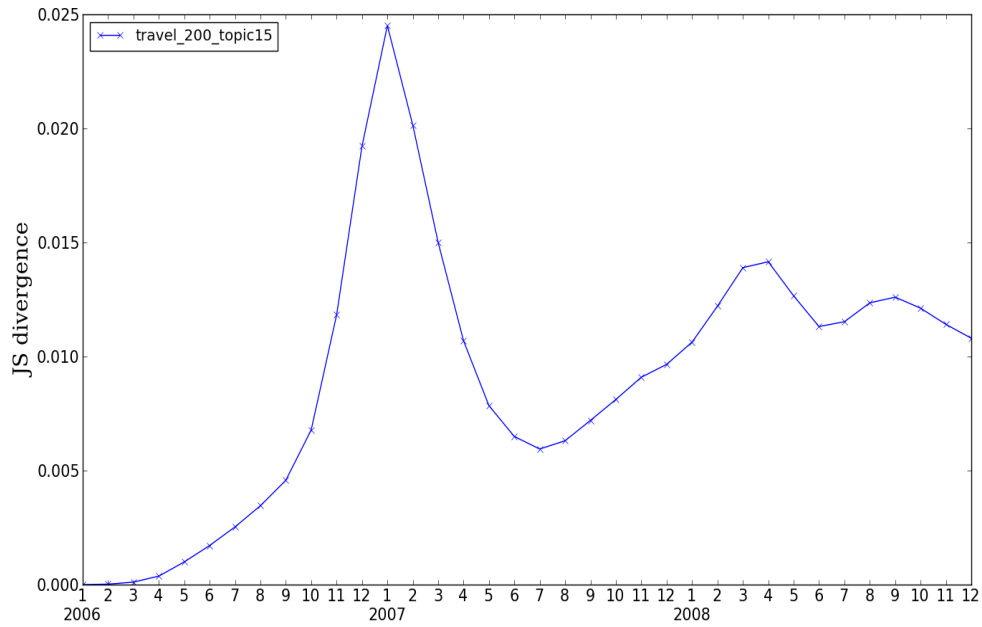


図 3.24　旅行カテゴリ，トピック 15 の話題遷移

表 3.10　旅行カテゴリ各トピックの代表的な単語

トピック	単語
0	群馬，栃木，真ん中，里，湯沢
171	3 月，ヨーロッパ，卒業旅行，スペイン，予定
15	お薦め，名所，新潟，動物園，工事

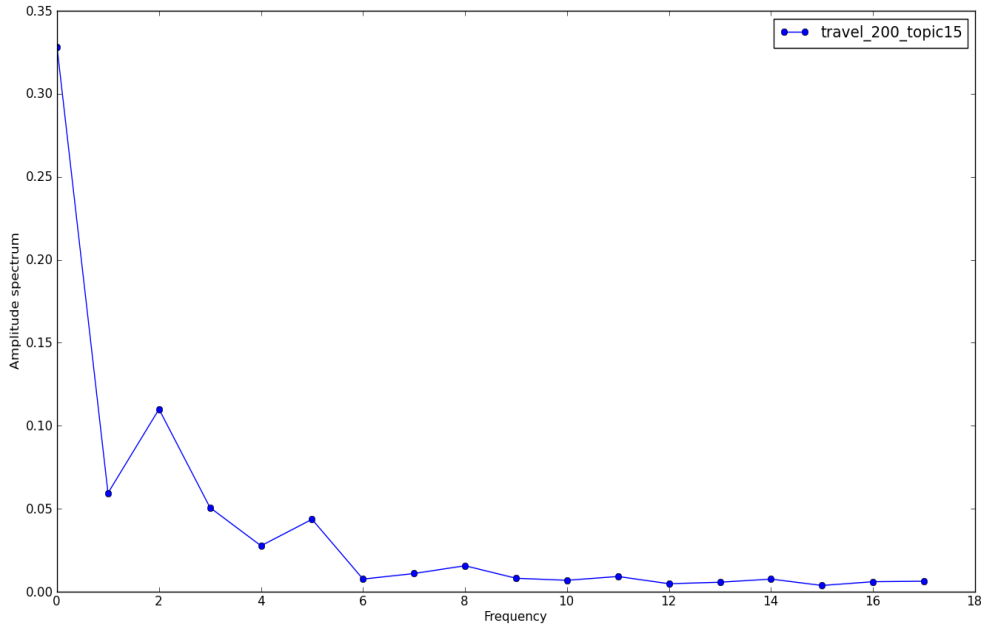


図 3.25　旅行カテゴリ，トピック 15 のパワースペクトル分布

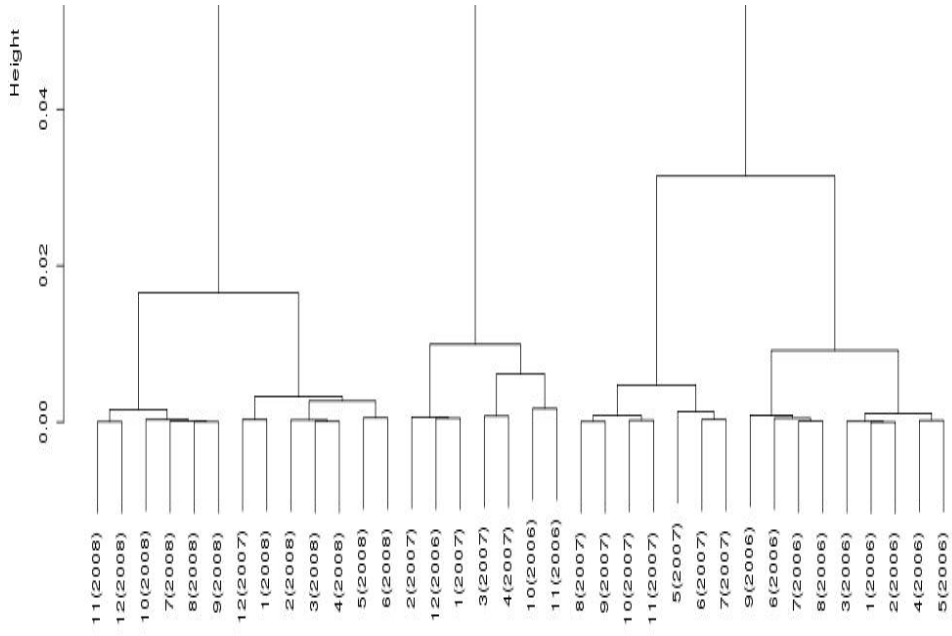


図 3.26　旅行カテゴリ，トピック 15 の各月の単語分布によるクラスタリング結果

その他のカテゴリでのトピックのクラスタ分類結果

“PC”，“経済”，“健康” カテゴリにおいて，旅行カテゴリと同様に階層的クラスタリングを行った．ここで，クラスタ数を 3 とした時の，各カテゴリで作成できたクラスタ内のトピックと，旅行カテゴリのトピック 0，171，15 のパワースペクトル分布の類似度を計算し，クラスタの話題変動のタイプを分類した．旅行カテゴリにおいてトピック 0 が所属するクラスタを “単調変化” クラスタ，トピック 171 が所属するクラスタを “周期” クラスタ，トピック 15 が所属するクラスタを “バースト” クラスタとした．分類を行った結果を表 3.11 に示す．表内の数字は各変動タイプのクラスタに属するトピック数である．スポーツトピックは，3 クラスタに分類したところ，2 クラスタでトピック 171 と最も類似度が高くなるトピックが多数となったため，2 クラスタが周期型クラスタに分類されている．実験の結果，各話題変動タイプに属するトピックの代表として表 3.12 に示す単語を抽出できた．

表 3.11　PC，健康，経済カテゴリでのトピック分類結果

カテゴリ	単調型	周期型	バースト型
PC	167	1	32
健康	190	7	3
経済	177	23	0

表 3.12　PC，健康，経済糧での代表的な単語

カテゴリ	トピック番号	変動タイプ	単語
PC	18	単調変化	mp3，音楽，音質
	73	周期	印刷，できる，年賀状
	105	バースト	ニコニコ動画，ニコ厨
健康	12	単調変化	癌，細胞，禁止
	1	周期	寒い，対策，外
	169	バースト	下痢，ノロウイルス，腸
経済	7	単調変化	現金，振込み，atm
	6	周期	医療費，入院，対象

3.4　考察

3.4.1　時系列トピックモデルによる話題遷移パターン抽出に関する考察

時系列トピックモデルと JS ダイバージェンスを用いた話題変動パターンの抽出では，全てのトピックにおいて，JS ダイバージェンス値が徐々に増加していく傾向が確認できた．これは，全ての話題が徐々に変化していることを示していると言える．PC カテゴリのトピック 3 は，パソコン購入に関するトピックであったが，2008 年頃から “eeepc” や，“ミニノートパソコン” など，ネットブックに関する単語の出現確率が徐々に増加している．これは，CQA の PC カテゴリ内でパソコン購入に関する質問記事の中で，ネットブック購入に関する質問が徐々に増えていっていることを示していると考えられる．PC カテゴリでは，ほとんどのトピックがこのような単調変化の話題変化パターンとなっている．このことから，単調変化型の話題変化や，新製品や新情報により，情報の鮮度が刻々と変化する話題で発生する話題変化のパターンではないかと考えられる．

PC カテゴリでのトピック 2 では，2007 年の 5 月ごろから話題変動量が増加し，2007 年 9 月にピークとなりその後は緩やかな話題変化となっている．このような特異な話題変化は，話題のバーストであると考えられる．トピック 2 で大きく確率変動した単語は “ニコニコ動画” や “初音ミク” といった語であるが，トピック 2 で大きく話題変動している時期と，実際にニコニコ動画や初音ミクが登場した時期が一致している．トピック 2 は PC と音楽関係に関する話題であるが，“ニコニコ動画” や “初音ミク” はどちらもそれらに関係する語である．初音ミクは楽曲データを人口音声で歌唱させることのできるソフトウェアであり、この様子をニコニコ動画に投稿することが 2007 年 9 月ごろにネット上で流行しており，この流行がトピック 2 のバーストに反映されたものであると考えられる．このように，バースト型の話題変動や実世界などでの印象的な出来事を色濃く反映する話題で発生すると考えられる．

旅行カテゴリでは，PC カテゴリに見られない特徴的な話題変動を確認できた．旅行カテゴリのトピック 1 は，“京都”，“修学旅行” といった単語に関連するトピックであるが，“紅葉”，“桜” は毎年見頃となる時期にトピック内の出現確率が上昇している．JS ダイバージェンスの推移もこれらの単語の変動と一致している．以上のことから，旅行カテゴリトピック 1 は京都や修学旅行に関する話題であるが，毎年 3，4 月の春には桜に関する質問記事，9〜11 月の秋には紅葉に関する質問記事が多く投稿されていると考えられる．このように，周期的な変動をする話題は季節的な特徴を持っていると考えられる．

旅行カテゴリトピック 1 では，変動語に “紅葉”，“桜” といった季節性に関連する語が抽出されている．3.2.2 節での単語ベースの話題変動でも，旅行カテゴリでは “雪” といった語が抽出できている．1 単語のみでは “雪” がどのような質問記事で使用されているのかを判別する

ことができない．しかしながら，トピックモデルを用いた手法では “紅葉”，“桜” と言った語は，京都や修学旅行に関する話題で使用され，かつそれらの語は出現確率が周期的に変化することが確認できる．このことから，トピックモデルを用いた話題変動抽出では，“どのような話題が周期的な話題変動を発生されるのか”，また，“どのような単語が話題変動に大きく影響するのか” を明らかにできるといえ，CQA 質問記事の話題遷移の性質をより詳細に把握できるようになったと考えられる．

3.4.2　周波数解析による周期的なトピックの抽出に関する考察

Yahoo!知恵袋を用いたトピック抽出実験

旅行カテゴリトピック 171 の図 3.21 の話題遷移グラフでは，毎年，1 月に JS ダイバージェンス値が最小値になる一年周期の話題変動を起こしている，そして図 3.22 のパワースペクトル分布では，$k = 3$ のときにピークが立つことから周期は 12 ヶ月であり，一年周期で一致している．また，トピック内の月ごとの確率分布を元にクラスタリングをした場合，トピック 0 は，2006 年 1 月と 2 月，2008 年 3 月と 4 月といったように隣合う月同士で結合しながらクラスタを形成しているが，図 3.23 のトピック 171 の結果では，2006 年 1，2，3 月と 2007 年 1，2，3 月といったように異なる年の同じ期間同士が結合していく．確率分布が近いというのは，近い話題であるということを示しており，トピック 171 では，毎年同じ月には近い話題についての質問記事が投稿されていることを示しており，話題が一年周期で変化していることを示している．

旅行カテゴリトピック 15，図 3.24 では，2006 年 11 月から 2007 年 4 月までの間急激な話題変動が発生している．これは，図 3.26 の月間のクラスタ結果でも現れており，2006 年 11 月から 2007 年 4 月までの 7 ヶ月間は他の期間と異なるクラスタを形成している．このことから，トピック 15 は話題がバーストしてると考えることができる．

旅行カテゴリ以外のカテゴリでは，生成されたクラスタが旅行カテゴリの，どのタイプのクラスタに近いかを比較することにより，話題変動のタイプごとにトピックを分類した．PC カテゴリでは話題がバーストするトピックが多いがこれは，途中から急激な話題変動を起こすものが多かった．PC 関連は新製品や新サービスの登場で話題が急激に変わるケースが多いのではないかと考えれる．経済カテゴリでは周期的な話題変動のトピックと判定されたが，実際は高頻度でバーストを繰り返すトピックも存在しており，それに周期性があると判断されたと考えられる．クラスタリングの際のクラスタ数を増やすことで，より様々な話題変動を分類できるのではないかと考えられる．

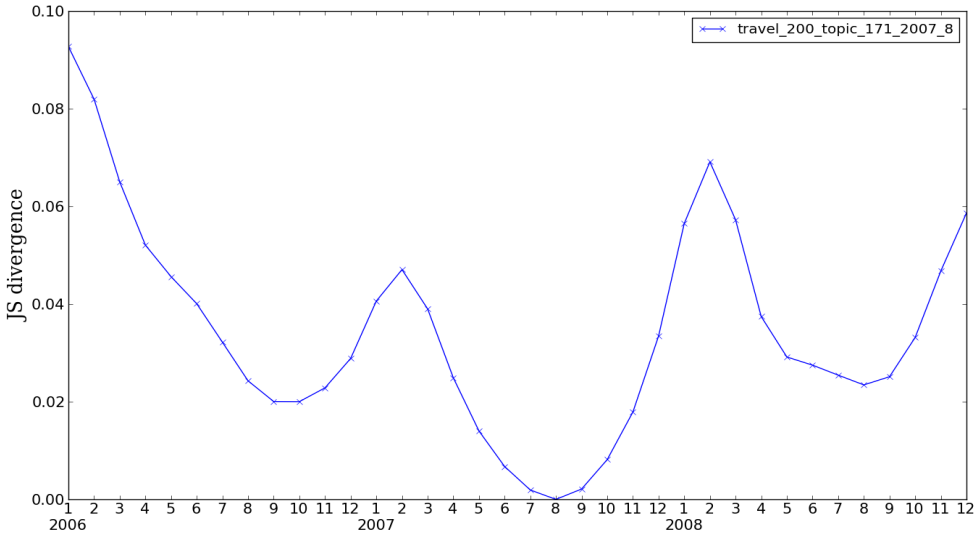


図 3.27　旅行カテゴリ，トピック 171 の話題遷移（2007 年 8 月基準)

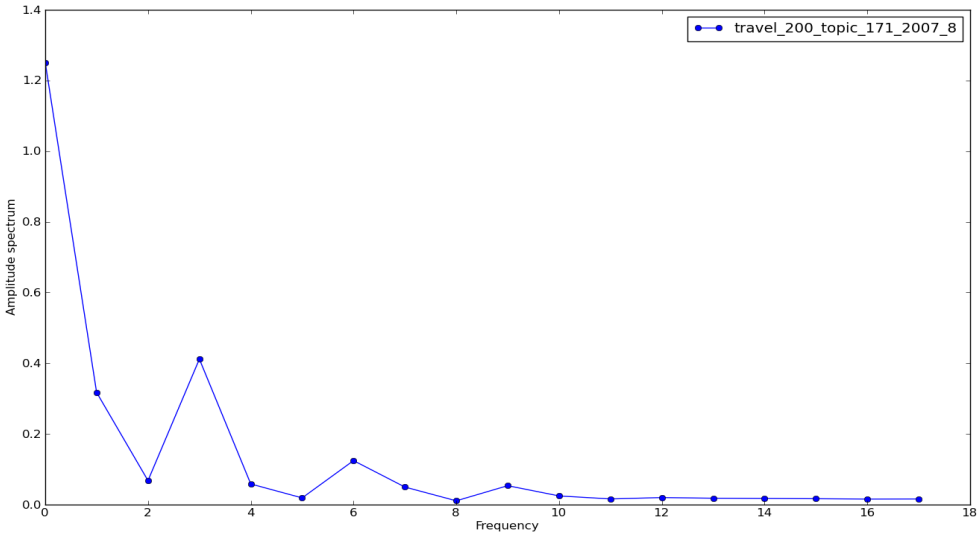


図 3.28　旅行カテゴリ，トピック 171 のパワースペクトル分布（2007 年 8 月基準)

周波数解析によるトピック分類手法

提案手法によりトピックをクラスタリングをした結果，それぞれのクラスタのトピックは全く異なる話題変動をしていることがわかる．また，パワースペクトル分布においても，それぞれ異なる特徴を示していることから，提案手法である周波数解析による話題変動タイプに基づくトピック分類の有用性が示されているといえる．また，実験ではデータセットの基点である 2006 月 1 月の確率分布と他の月の確率分布の類似度を算出することで，話題変化量の時系列データを得ているが，離散フーリエ変換により，周波数領域においてクラスタリングすること

で基準点の違いを解消できると考えられる．図 3.27 は，JS ダイバージェンスの基準を 2007 年 8 月の確率分布にした時の話題変動のグラフである．グラフの形状は 2006 年 1 月基準の図 3.21 と異なっているが，このデータを離散フーリエ変換により周波数領域のパワースペクトルで表現すると図 3.28 となる．2006 年 1 月基準の図 3.22 と比較すると，$k = 3$ でピークが立っており，分布の特徴にほとんど差がないといえる．話題変動のグラフの形状が異なるのは，変動の位相が異なるためであると考えられるが，どちらにおいても，一年周期の話題変動をするトピックであると判別することができる．また，図 3.23 の各月の確率分布の類似度に基づくデンドログラムは，全ての月の確率分布間の類似度を計算することによって作成されている．異なる年の同じ期間同士がクラスを形成していることから一年周期で似た話題となっていることが，一年周期のトピックであることを証明しているといえる．以上のことから，本手法では，JS ダイバージェンスによる話題変化量の時系列データを得ることで，話題の基準点によらず，周波数解析による話題変動のタイプに基づいたトピック分類が可能であるといえる．

3.5　まとめ

本章では，コミュニティ QA（CQA）に投稿される質問記事で議論される話題の周期性について明らかにした．質問記事で使用されている単語の出現頻度の推移から，CQA で使用される単語には周期性があることを明らかにした．

CQA に投稿される質問記事を時系列トピックモデルによって話題単位にクラスタリングし，トピックの確率分布の時間変化から JS ダイバージェンスにより話題変化量を算出し，話題変化を追跡した．その結果，CQA の話題変動には，話題が単調に変化する単調変化形，ある時期に話題が急激に変化するバースト型周期に似た話題が議論される周期型の 3 つの話題変化パターンが存在することが明らかになった．また，それらの話題変動は各カテゴリに混在しており，各カテゴリには短調変化や周期型など，様々な変動タイプの話題に関する質問記事が投稿されていることを明らかにした．

トピックの話題変動量の時系列データから，離散フーリエ変換による周波数解析により，周期的な特徴量データを作成し，それをクラスタリングすることで，話題の周期性に関するトピック分類を実現した．その結果，トピックの中から周期的な話題変動パターンとなるトピックを抽出することが確認できた．

第 4 章

コミュニティ QA を用いたクエリ拡張型 Web 検索システム

4.1 はじめに

本章では，コミュニティ QA（CQA) の質問記事を用いた，情報要求の言語化支援を行う Web 検索システムについて説明する．まず始めに，情報要求言語化のための拡張クエリ “質問記事付き拡張クエリ” とタブとタグクラウドを組み合わせたファセット検索型の検索インターフェースを提案する．

4.1.1 質問記事付き拡張クエリ

本研究では，Web 検索における情報要求の言語化を支援するため，CQA の質問記事と質問記事から作成した拡張クエリをセットで提示する “質問記事付き拡張クエリ（CQA クエリ）” を提案する．CQA クエリの例を図 4.1 に示す．複数のキーワードからなるキーワード拡張クエリには，ユーザにとって未知のキーワードが含まれていることがあるが，CQA クエリでは，質問記事を参照することで，クエリの主題を知ることができるという特徴がある．例では，“ノロウイルス” というキーワードをユーザが知らなかったとしても，質問記事を参照し，“私の友人も感染” などの文章を読むことで，ノロウイルスが人に感染するウイルスであることを把握でき，違和感なく Web 検索でノロウイルスの予防法について調べることができるようになる．このように，自然言語はユーザにとって理解しやすく，疑問が詳細に記述されているため，質問記事を参照することで，拡張クエリから “具体的な情報要求” を表出することができる．

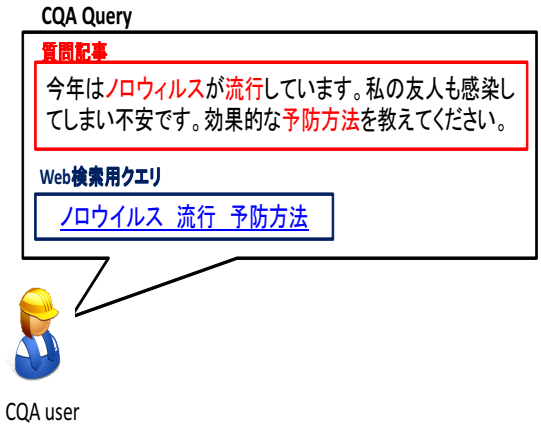


図 4.1　質問記事付き拡張（CQA クエリ）

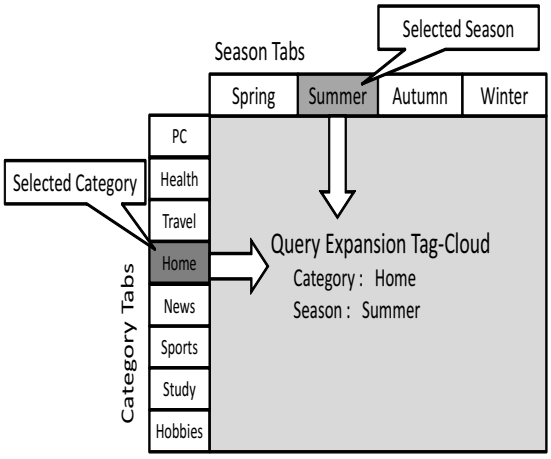


図 4.2　タブ-タグクラウドインターフェース

4.1.2　タブとタグクラウドによるファセット検索インターフェース

CQA には大量の質問記事が投稿されている．質問記事は個人的かつ詳細な内容が記述されているため，同じ主題の質問記事であっても，質問記事の細かい内容によって異なる拡張クエリが作成される．そのため，大量に生成される CQA クエリの中から，ユーザの興味や状況に近いクエリを提示するための仕組みが必要となる．本研究では，様々な切り口から CQA クエリを提示するファセット検索型のインターフェースを提案する．検索のファセットとして，質問記事が投稿されているカテゴリと投稿時期を使用する．CQA では，質問記事を投稿する際，質問の内容に近い一つのカテゴリを選択する必要がある．そのため，CQA の質問記事はカテゴリによって質問内容の分類がなされていることになる．質問記事の内容が異なれば，そこから作成される拡張クエリのキーワードも異なるため，ファセットとして，カテゴリを切り替えることで，異なる CQA クエリを提示できるようになる．2 つ目のファセットとして，季節を導入する，CQA に投稿される質問記事の話題には周期性があることが，3 章での実験の結果分かっている．そこで，質問記事の投稿時刻を用いて，質問記事を季節ごとに分類することで，季節ごとに異なる CQA クエリを提示する．

カテゴリと季節の 2 つのファセットを用いた拡張クエリの提示法として，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用いる．インターフェースのイメージを図 4.2 に示す．縦に配置されているのがカテゴリ選択タブ，横に配置されているのが季節タブである．ユーザは，それぞれのタブから興味のあるカテゴリ，季節を選択することでそれらのコンテキストに関連する CQA クエリを入手することができる．インターフェースの中央部は，キーワードのタグクラウドが表示される．タグクラウドには，クエリ拡張で追加されるキーワード

一覧が表示される．キーワードが選択されると，入力キーワードと選択キーワードを含む質問記事から作成された CQA クエリが表示される．

以後の節では，CQA クエリとファセット検索型インターフェースの実装方法と評価について詳述する．

4.2　提案手法の実装

作成したシステムのスクリーンショットを図 4.3，システム構成図を図 4.4 に示す．本システムは主に 4 つのブロックで構成されている．

ファセット検索部の検索窓にキーワードが入力されると，キーワードに関連するカテゴリタブ，システムにアクセスした時刻から季節タブが初期状態として設定される．カテゴリ，季節それぞれでタブを選択すると，入力キーワード，選択カテゴリ，季節に関連するキーワード一覧がタグクラウドに表示される．タグクラウド中のキーワードが選択されると，システムは選択されたカテゴリ，時期に投稿された質問記事の中から，入力キーワードと選択キーワードが含まれる質問記事を検索し，CQA クエリブロックに表示する．それぞれの質問記事には，質問記事からのキーワード抽出により作成した Web 検索用のキーワード組が付与されている．

以降は，まずシステムの実装に用いたデータセットについて説明し，次に，ファセット検索ブロックでのタブとタグクラウド，CQA クエリブロックでの CQA クエリの実装方法について詳述する．

図 4.3　プロトタイプシステム画面

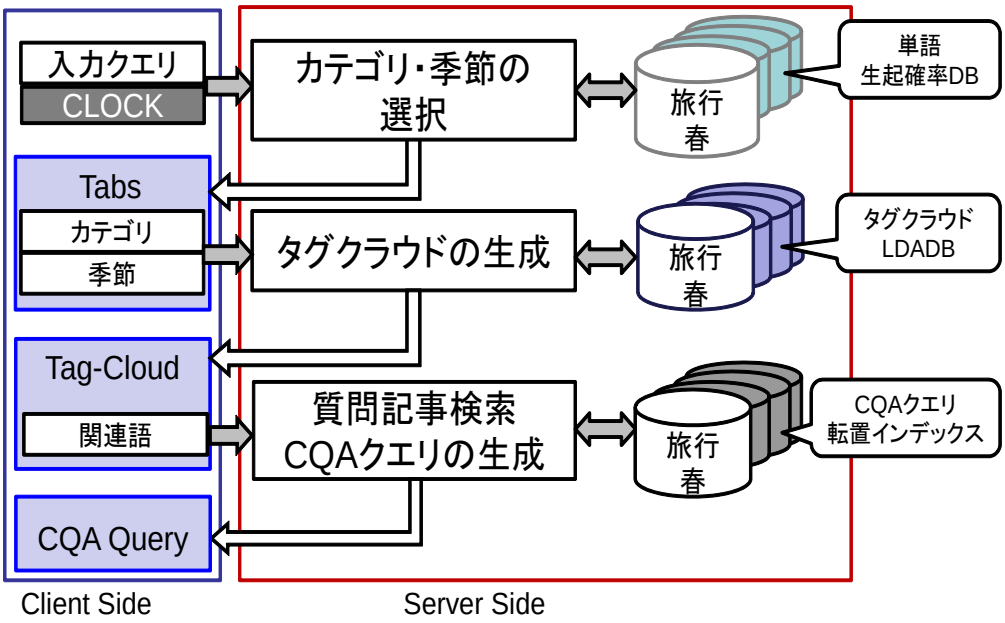


図 4.4　システム構成図

4.2.1　データセット

本システムで使用する CQA コーパスとして，国立情報学研究所の提供する Yahoo!知恵袋[*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．ここでは，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．使用する季節は以下のようになる．

[*1] http://chiebukuro.yahoo.co.jp/

表 4.1　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

- 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 4.1 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ことで，全 80 個のデータセットとなった．

4.2.2　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，最初に検索クエリ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であり、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．

$$P_{C,w} = \frac{N_{C,w}}{N_C} \tag{4.1}$$

$N_{C,w}$ 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．

4.2.3　タグクラウドの実装

ファセット検索ブロックでタブが選択されると，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．タグクラウドは，キーワードを表示順と文字サイズによって差別化できるため，情報検索の可視化で幅広く使用されている [7]．本システムでは，タグクラウドの表示順として，入力キーワードに関連のあるキーワードほど先に表示される．

表示順を決定するための，単語間の関連度を計算する手法を述べる．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を用いたアプローチによる関連度計算を実装する．ここでは，Blei ら [3] 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．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．

$$P(z_i = k|Z_{-i}, W) = \frac{N^d_{k-i} + \alpha}{N^d_{-i} + T\alpha} \frac{N^v_{k-i} + \beta}{N_{k-i} + W\beta} \tag{4.2}$$

ここで，$-i$ 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 $\hat{\theta_d^k}$，トピック k から単語 w が生成される確率 $\hat{\phi_k^w}$ は以下の式で与えられる．

$$\hat{\theta_d^k} = \frac{N_k^d + \alpha}{N^d + T\alpha} \tag{4.3}$$

$$\hat{\phi_k^w} = \frac{N_k^v + \beta}{N_k + W\beta} \tag{4.4}$$

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．

$$sim(w_1, w_2) = cos(P_{w_1}, P_{w_2}) = \frac{P_{w_1} \cdot P_{w_2}}{|P_{w_1}||P_{w_2}|} \tag{4.5}$$

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．システムでは類似度を単語間の関連度として，関連度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．

4.2.4　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．

$$P(D|Q) = \frac{P(Q|D)P(D)}{P(Q)} \tag{4.6}$$

$P(Q)$ 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．

$$P(Q|D) = \prod_{w \in Q} P(w|\theta_D)^{c(w,Q)} \tag{4.7}$$

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．

$$P(w|\theta_D) = \frac{c(w,D)}{|D|+\mu} + \frac{\mu}{|C|(|D|+\mu)} \sum_{D \in C} \frac{c(w,D)}{|D|} \tag{4.8}$$

$c(w,D)$ 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．

4.3　評価実験

本節では，実装したシステムに対しての評価実験について説明する．評価実験として，ファセットを切り替えた際に提示されたクエリによって検索できる Web ページの多様性に対して評価を行った．

以降は，多様性評価のための実験手法について説明し，次にカテゴリと多様性評価，季節と多様性評価を説明する．

4.3.1　キーフレーズ抽出による Web 検索の多様性評価

本研究で実装したシステムでは，質問記事の違いにより多様な Web 検索を実現している．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．例えば，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 [16] 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．

キーフレーズ抽出による多様性評価実験手法を，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[*2]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[*3]を用いて，スニペット毎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．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，内容をあ表すキーフレーズに相応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．1 回しか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一つの 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，クエリ周辺の話題のみを抽出するためである．

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[*4]，Google[*5]，Yahoo!JAPAN[*6] の拡張クエリについて，評価実験を行う．評価は，“ウィルス” 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 “ウィルス” の拡張クエリを表 4.2 に，“mac” の拡張クエリを表 4.3 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 4.4 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わかる．“ウィルス” 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

[*2] http://developer.yahoo.co.jp/webapi/search/
[*3] http://developer.yahoo.co.jp/webapi/jlp/keyphrase/v1/extract.html
[*4] http://www.bing.com/?cc=jp
[*5] http://www.google.co.jp/
[*6] http://www.yahoo.co.jp/

に関連するクエリであるが，他の 3 個は病気関連，1 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，その他，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 は全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．

表 4.2　入力語 “ウィルス” の拡張クエリ

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 4.3　入力語 “mac” の拡張クエリ

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

表 4.4　各検索エンジンのキーフレーズ抽出数

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

既存手法との比較によるカテゴリとクエリの多様性に関する評価

複数のカテゴリから拡張クエリを作成することで，より多様な Web 検索を実現できることを，既存の拡張クエリ提供サービスとの比較により評価する．実験では，カテゴリごとにキーワードの関連語を出力するプロトタイプシステム[*7]を用いて評価を行なっている．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 カテゴリ全てに存在する 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 4.5 に示す．比較対象として Yahoo!関連検索ワード API[*8] 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** :各カテゴリの第一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 4.6 に示す．作成したは前節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 4.5 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 4.6 と図 4.7 に示す．図 4.6 が “日本語”, 図 4.7 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である．“日本語” では，クエリ数が多くなるに従い，提案法 (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わかる．一方で，“ソフト” は，提案法, 既存手法共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．

表 4.5　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 4.6　入力語 “ソフト” の拡張クエリ

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

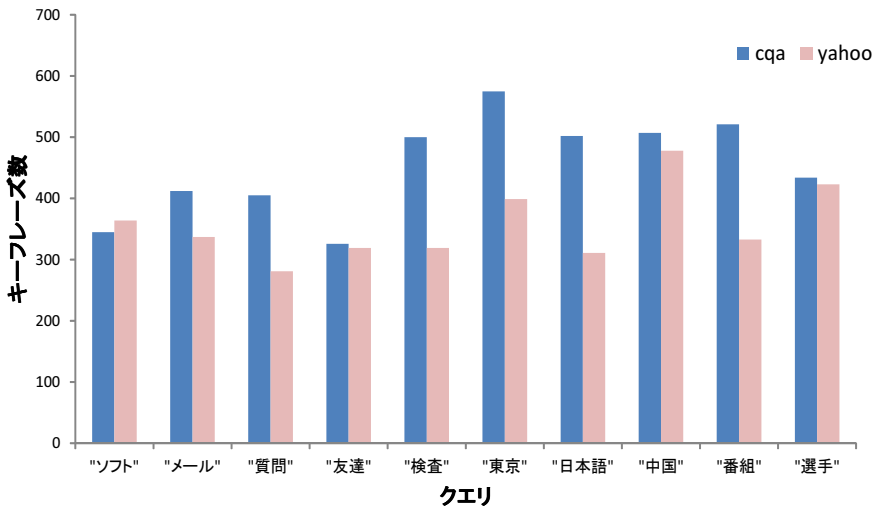


図 4.5　提案法 (cqa) と既存手法 (yahoo) のキーフレーズ抽出数

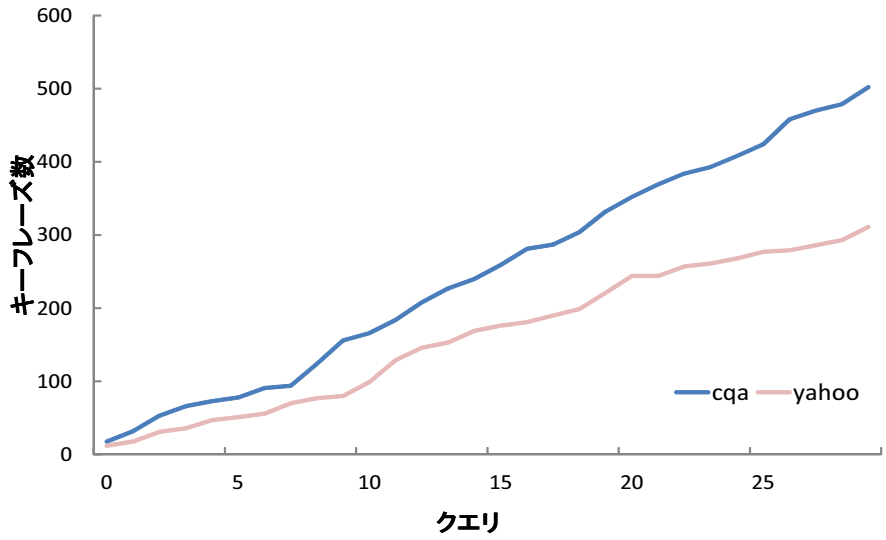


図 4.6　“日本語”でのキーフレーズ抽出数の推移

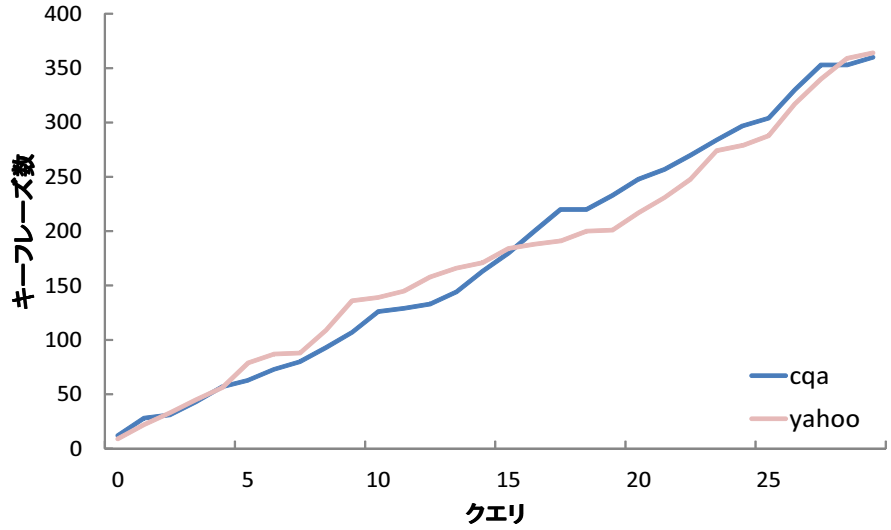


図 4.7　“ソフト”でのキーフレーズ抽出数の推移

カテゴリ数と多様性の評価

表 4.5 に示した実験用の入力クエリは，提案法において，用意した 10 カテゴリ全てで拡張クエリが推薦されている．しかし，実際には，10 カテゴリ全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に対して，拡張クエリを作成し，キーフレーズ抽出数について評価を行う．入力クエリには 2004 年と 2005 年の

*7 プロトタイプシステムでのデータセットは 2004 年 4 月から 2005 年 9 月までのデータを用いている

*8 http://developer.yahoo.co.jp/webapi/search/assistsearch/v1/webunitsearch.html

表 4.7　全クエリに対する提案法と既存手法の比較

	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

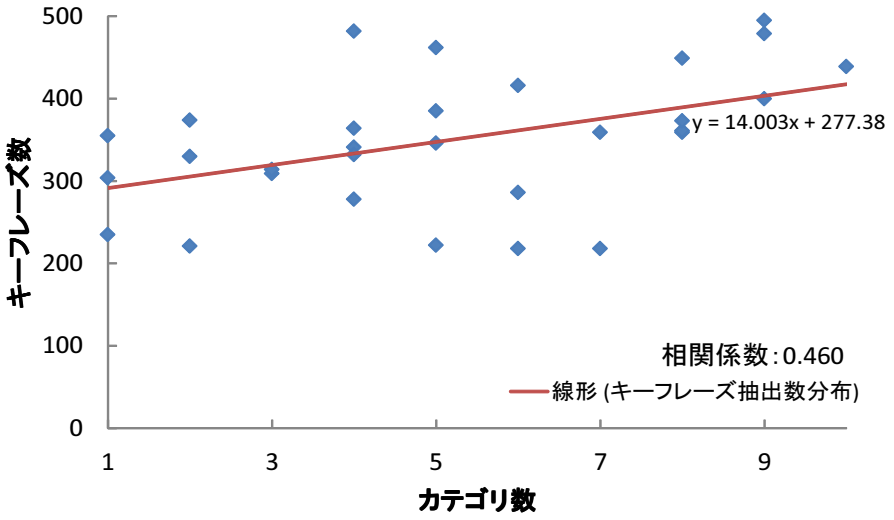


図 4.8　抽出カテゴリ数とキーフレーズ数の相関

Yahoo!JAPAN の検索ワードランキング[*9] の総合ランキングから “ANA”，“dell”，“hotmail” などの企業名やサービス名,“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 $n_{q,c} = [\frac{30}{N_q}]$ 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 4.7 に示す．全 30 クエリの内 22 クエリにおいて，提案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 4.8 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

[*9] http://picks.dir.yahoo.co.jp/new/review2004/
http://picks.dir.yahoo.co.jp/new/review2005/

カテゴリを抽出できていた．1 カテゴリでも 300 程度のキーフレーズを抽出できているが，回帰分析の結果, 抽出カテゴリをが多いくなるにつれ，キーフレーズ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，これは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，多くのキーフレーズが抽出できていた．

4.3.2　季節の違いによるクエリ拡張の評価

3 章の実験で，CQA の質問記事の内容の話題変動は，季節性を持つことがわかっている．本節では，実際のシステムにおいて，データセットを季節ごとに分割した際に，ことなる拡張クエリが推薦されるのかを評価する．3 節の実験で家電カテゴリでは “年賀状”，旅行カテゴリは “桜”，恋愛カテゴリでは “チョコ” が季節性を持つ語であることを明らかにしている．そこで，これらの語を入力クエリとしてシステムに投稿した際，システムに出力された拡張クエリを季節ごとに比較する．

表 4.8 にシステムの出力結果をまとめたものを示す．入力クエリ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．家電カテゴリでの “年賀状” が冬，旅行カテゴリでの “桜” が春，恋愛カテゴリでの “チョコ” が冬に最も多くのユニーク語を持つことがわかる．ユニークキーワード数の推移をグラフにまとめたものを図 4.9 に示す．横軸は季節，縦軸はタグクラウドのキーワードを示しており，棒グラフがタグラウクドに出力される総出力語数，折れ線グラフがユニーク語数である．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．

表 4.9 に，実システムでの CQA クエリの例を示す．例では，入力クエリとして “桜” をシステムに入力し，旅行カテゴリ，春のタブを選択，出力されたタグクラウドから “花見” を選択したときの例である．タグクラウドから “花見” を選択した時点で，CQA の旅行カテゴリで春に投稿された質問記事から “桜” と “花見” が含まれる質問記事が検索される．一つ目の CQA クエリの例では，アメリカ，ワシントン DC の桜についての質問記事であり，もう一方の CQA のクエリは花見をしながら食事ができるレストランを探す質問記事となっている．同じキーワードが含まれる質問記事であっても，質問の主題は異っていることがわかる．また，質問記事から作成される Web 検索用のキーワード組（クエリ）も異なるものとなっている．

表 4.8　タグクラウドの出力例とユニークキーワード数

入力クエリ [カテゴリ]		出力語	ユニーク語	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

4.4　考察

4.4.1　クエリ拡張におけるカテゴリと多様性

4.3.1 節の実験では，検索された Web ページのスニペットからキーフレーズ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，多くのキーフレーズを抽出できていたことから，キーフレーズが多様性の評価に有効に機能していると考えられる．提案法と既存のクエリ拡張の比較では，ほんどの場合において提案法の方がより多くのキーフレーズを抽出できていたことから，提案法は多様性を持ったクエリ拡張が行われていると考えれることができる．

図 4.8 より，展開されるカテゴリが多ければ多いほど，より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，やや強い正の相関となったが，実験で使用した 30 クエリに表 4.5 の全カテゴリで出現する代表的な語を含めると，相関係数は 0.589 となりより強い相関となる．そのため，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 4.8 では、単一のカテゴリのみで拡張

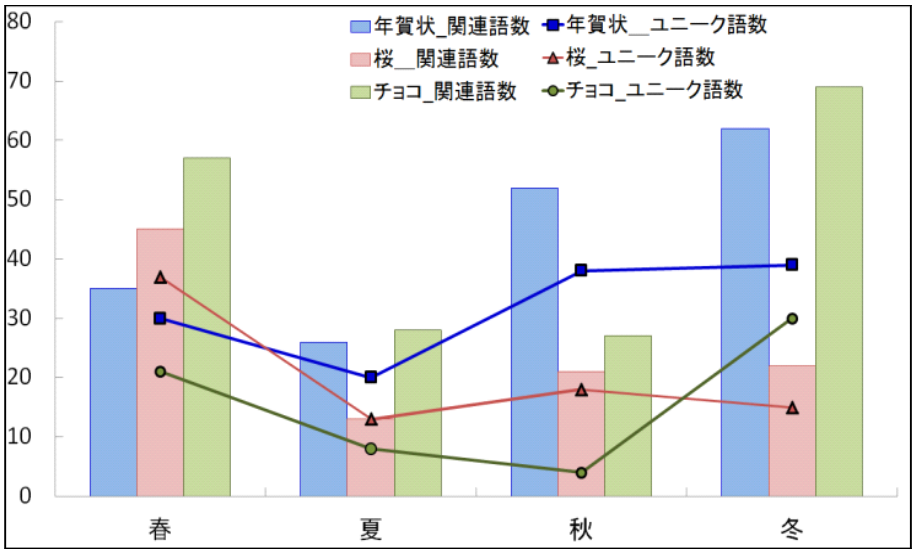


図 4.9　ユニークキーワード数の季節ごとの推移

表 4.9　CQA クエリの例

タグクラウド			CQA クエリ	
選択クエリ	季節	カテゴリ	質問記事本文	キーワード組
桜 花見	春	旅行	アメリカ、ワシントン D.C. のポトマック川周辺に日本から輸入された桜があるそうですが、花見行ったことある人いますか？	桜 花見 アメリカ
			室内でお花見（桜）出来るレストラン等、ご存知の方教えてください。	桜 花見 レストラン

クエリを作成してもある程度のキーフレーズが抽出できた．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．膨大な数に及ぶ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．

4.4.2　CQA の季節性とクエリ拡張への影響

システムの出力結果では，3 章でバーストが発生していた時期に，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．バーストが発生した季節以外にも，多くのユニーク語が出力されており，恋愛カテゴリでの “チョコ” は 2 月にバーストする冬のキーワードであるが，春にも多くのユニーク語が出力されている．“お返し”，“ブランド” というキーワードからこれは，3 月のホワイトデーに関するキーワードであると考えられる．また，家電カテゴリの “年賀状” にかんしては，“dpi”，“スキャン” などのユニーク語が出力されているが，これは，年賀状作成のための準備としてプリンタを秋に購入することが多いからではないかと考えられる．このように，バースト以外の季節によっても，システムでは季節に応じた独特な拡張クエリを提示できることを示しており，拡張のコンテキストとして季節を導入したことは妥当であると考えることができる．

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．

4.5　まとめ

本章では，Web 検索における情報要求言語化支援のための，クエリ拡張手法を提案した．コミュニティ QA の質問記事を外部リソースとして，拡張クエリを実装しキーワード抽出を行った質問記事とセットで提示する “質問記事付き拡張クエリ” により，キーワード組の背後にある疑問の根拠を提示することにより，ユーザ自身では思いつくことのできない情報要求を言語化している．

コミュニティ QA の質問記事のメタ情報である投稿カテゴリと投稿時期を検索のコンテキストとして，タブとタグクラウドを組み合わせたユーザインターフェースにより多様なクエリ拡張を提供するシステムの実装を行った．実験により，異なるカテゴリから拡張クエリを作成することで，既存のクエリログを用いた拡張クエリよりも多様な Web ページを検索できることを明らかにした．また，季節ごとにデータセットを分割し，異なる拡張クエリを提供することで季節特有の情報要求に対するクエリが推薦されることを明らかにした．

第 5 章

結論

本論文では，コミュニティ QA（CQA）質問記事に着目し，ユーザが言語化できない情報要求を言語化するための研究を行った．以下で本論文の内容を簡潔にまとめ，今後の展望について述べる．

3 章では，CQA に投稿される質問記事の内容を分析し，質問記事の中には，周期的に似た内容の質問記事が投稿される話題があることを明らかにした．自然言語で記述されている質問記事の周期性を明らかにするために，時系列トピックモデルにより，質問記事集合を “話題” の集合であるトピック集合に変換した．時系列トピックモデルでは，トピックを表現する単語の出現確率分布が時間経過と共に変化するモデルである．確率分布の変化を JS ダイバージェンスにより，定量化することで，話題変化の時系列データを作成した．話題変化のパターンを離散フーリエ変換により周波解析した結果，CQA の質問記事の話題変動パターンは，時間とともに話題が徐々に変化するパターン（単調変化型），ある時期に急激に話題が変化するパターン（バースト型），そしてある周期性を持って話題が変化するパターン（周期型）の 3 パターンに分類できることが明らかとなった．周期型トピックでは，単一のトピックであっても，季節によって投稿される質問の内容に違いがあることが明らかとなった．例えば，旅行カテゴリにおいて，京都に関するトピックでは，春には “桜” に関する質問が多く投稿され，秋には “紅葉” に関する質問記事が多く投稿されている．以上のように，自然言語で記述された質問記事を分析することにより，Web ユーザの情報要求の変化を話題単位で追跡可能になった．

4 章では，CQA の質問記事を用いた拡張クエリによる Web 検索システムを提案した．CQA の質問記事を Web ユーザの情報要求とみなし，質問記事から作成したキーワード組と質問記事本文をセットで提示する “質問記事付き拡張クエリ（CQA クエリ）” を提案した．CQA クエリは，自然言語で記述された情報要求が付与されていることになるので，ユーザがキーワード組を意味を理解できない場合，付与されている自然言語の質問記事を参照することで，キーワード組を背後にある情報要求を明確に言語化できるようになる．3 章で得た CQA の周期性を活用し，カテゴリと季節を切り替えるコンテキストを切り替え型クエリ拡張インターフェー

スを実装した．Web 検索ユーザの情報要求の背後にある，コンテキストをカテゴリと季節に反映させ，コンテキストを切り替えることにより多様なクエリを推薦する手法を提案した．実験の結果，一つのカテゴリで多くの拡張クエリを作成するよりも，複数のカテゴリから少数の拡張クエリを作成し，統合したほうが，より多様な Web 検索を実現する拡張クエリを推薦できることを明らかにし，また，季節を切り替えることで，同じカテゴリであっても異なる拡張クエリが推薦されることを明らかにした．

今後の展望として，より季節性の強い拡張クエリの推薦が考えられる．4 章で実装したクエリ拡張システムでは，季節ごとに異なるクエリを作成するために CQA のデータセットを投稿時期ごとに分割している．しかし，3 章で示した通り，カテゴリに投稿される質問記事の中には，周期的な話題，単調に変化している話題が混在している．そのため，単にデータセットを分割しただけでは，周期性に関連する質問記事も周期性に関係ない質問記事も等しく扱われクエリが作成されている．3 章では，周波数解析を用いて，周期的な話題に関するトッピックを抽出できている．そこで，周期的なトピックに関連する質問記事のみを抽出し，そこから拡張クエリを作成することで，より季節性の強い拡張クエリを推薦できるようになると考えている．

謝辞

本論文は，筆者が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中の研究成果をまとめたものである．同研究科の佐藤哲司教授には主指導教員として，卒業研究から 3 年間にわたりご指導をいただいたこと，謹んで感謝申し上げます．また，共著者として多くの論文執筆にご協力頂いた，関洋平助教にも感謝の意を表し，お礼申し上げます．手塚太郎准教授にも，副指導教員として，丁寧にご指導いただき大変感謝しております．

研究を進めるにあたり，実験環境や研究室運営について尽力し，研究についても常に適切な助言をして頂いた，研究室 OB である，法政大学マイクロナノテクノロジー研究センターの島田愉さんにも大変感謝しております．さらに国立情報学研究所の神門典子先生には，研究データである Yahoo ！知恵袋データ提供のほか，研究方針の助言や英語プレゼンテーションの指導を頂き大変お世話になりました．心より感謝いたします．

研究室の同期として共に 2 年間助けあいながら，研究を進めてきた香川雄一くん，渡邊直人くん，そして関研究室所属の宮嶋清人くんにも大変感謝しております．ありがとうございました．本研究を最後までやり遂げることができたのは，先生方，先輩方，後輩たち，同期の皆様のおかげです．本当にありがとうございました．

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．

参考文献

[1] A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. Knowledge Sharing and Yahoo Answers: Everyone Knows Something. *Proceedings of the 17th International Conference on World Wide Web(WWW'08)*, pp. 665–674, 2008.

[2] 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.

[3] 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.

[4] 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 Data. *Proceedings of the 14th ACM international conference on Knowledge discovery and data mining(KDD'08)*, pp. 857–883, 2008.

[5] Miles Efron. Linear Time Series Models for Term Weighting in Information Retrieval. *Journal of the American Society for Information Science and Technology*, Vol. 61, No. 7, pp. 1299–1312, 2010.

[6] 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)*, pp. 619–628, 2010.

[7] Yusef Hassan-montero and Victor Herrero-solana. Improving Tag-Clouds as Visual Information Retrieval Interfaces. *International Conference on Multidisciplinary Information Sciences and Technologies(InScit2006)*, pp. 25–28, 2006.

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

[9] Jonathan Koren, Yi Zhang, and Xue Liu. Personalized Interactive Faceted Search. *Proceedings of the 17th International Conference on World Wide Web(WWW'08)*, pp. 447–486, 2008.

[10] Yuan Lin, Song Jin, Hongfei Lin, Yunlong Ma, and Kan Xu. Social Annotation in Query Expansion a Machine Learning Approach. *Proceedings of the 34nd international ACM SIGIR conference on Research and development in information retrieval(SIGIR'11)*, pp. 405–414, 2011.

[11] Qiaoling Liu, Eugene Agichtein, Gideon Dror, Yoelle Maarek, and Idan Szpektor. When Web Search Fails, Searchers Become Askers: Understanding the Transition. *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'12)*, pp. 801–810, 2012.

[12] Yajie Miao, Chunping Li, Jie Tang, and Lili Zhao. Identifying New Categories in Community Question Answering Archives: A Topic Modeling Approach. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management(CIKM'10)*, pp. 1673–1676, 2010.

[13] 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.

[14] 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.

[15] 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.

[16] Markus Strohmaier, Mark Kroell, and Christian. Koerner. Intentional Query Suggestion: Making user goals more explicit during search. *Proceedings of the 2009 workshop on Web Search Click Data(WSCD'09)*, pp. 68–74, 2009.

[17] 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.

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

[19] Soungwoong 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(WISE'09)*, pp. 145–158, 2009.

[20] 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.

[21] 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(WebDB Forum 2010), 2A-3, 2010.

[22] 山本岳洋, 中村聡史, 田中克己. Q&A コンテンツからの観点抽出に基づくウェブ検索支援. 情報処理学会論文誌 データベース, Vol. 4, No. 2, pp. 74–87, 2011.

[23] 岩田具治, 渡部晋治, 山田武士, 上田修功. 購買行動解析のためのトピック追跡モデル. 電子情報通信学会論文誌. D, 情報・システム, Vol. 93, No. 6, pp. 978–987, 2010-06.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 芹澤翠, 小林一郎. 潜在トピックの類似度に基づくトピック追跡への取り組み. 第 25 回人工知能学会全国大会論文集 (JSAI2012), 2012. 3F3-2.

[29] 山家雄介, 中村聡史, アダムヤトフト, 田中克己. ソーシャルブックマーキングの周期性発見と時期連動型検索ランキングへの適用. 情報処理学会論文誌 データベース, Vol. 2, No. 3, pp. 130–140, 2009.

[30] 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.

[31] 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.

発表論文

学術雑誌論文

- 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.
- 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 検索システムに関する一検討．情報処理学会論文誌データベース，Vol.4, No.3, pp.1-11. 2011.

国際会議論文

- Atsushi Otsuka, Yohei Seki, Noriko Kando, Testuji Satoh．QAque: Faceted Query Expansion techniques for Exploratory Search using Community QA Resources. Workshop in Community Question Answering on the Web(CQA2012)（Proceedings of WWW '12 Companion), pp.799-806. 2012.

国内会議論文

- 大塚淳史, 関洋平, 佐藤哲司．時系列トピックモデルを用いたコミュニティ QA からの話題変動の抽出．第 5 回 Web とデータベースに関するフォーラム (WebDBF2012). A3-3.2012.
- 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 (DEIM2012), F8-4.2012.
- 大塚淳史, 関洋平, 神門典子, 佐藤哲司．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．情報処理学会, マルチメディア，分散，協調とモバイルシンポジウム (DICOMO2011) 論文集,1B-2．2011.