Blog 記事のクラスタリングに基づいたカテゴリ別話題変遷パタンの抽出

戸田 智子† 福田 直樹† 石川 博†

†静岡大学情報学部情報科学科 〒432-8011 静岡県浜松市城北 3-5-1

E-mail: †cs3061@s.inf.shizuoka.ac.jp, {fukuta, ishikawa}@inf.shizuoka.ac.jp

あらまし 現在，ブログの爆発的普及と，ブログの特徴である即時性・個人性により，社会の関心事の動向を伺う手段としてブログが注目を集めている．ブログから発信される情報を解析し，話題となっているトピックや，ある特定のトピックの変遷を抽出する試みが行われている．ここでいう話題とは，多数から注目されているトピックである．ブログ中でのさまざまなトピックは，それぞれがある特定の興味分野に属していると考えられる．興味分野に注目して複数のトピックの関連性について分析することが社会の関心の動向を調べる 1 つの手段となりうる．本稿ではこの興味分野をカテゴリと呼び，カテゴリ内での話題変遷パタンを抽出することにより複数のトピックを同時に取り扱えるようにする．複数のトピックに対してカテゴリごとにそれらの変遷を抽出することにより，関連性の高い複数のトピックの変遷パタンの比較も行えるようにする．ブログ記事集合からトピックを抽出し，トピックの階層的クラスタリングによりカテゴリを決定する．そのカテゴリ中で各トピックのタイムスタンプとトピックに含まれる記事数に基づいて，カテゴリの話題変遷の特徴を抽出する．提案手法の有効性を，複数のカテゴリに関連した実際の Blog での関心事の動向の例に適用することで，示す．

キーワード ブログ文書，クラスタリング，トピックの変遷

Extraction of topic transition pattern of each category based on clustering Blog article

Tomoko TODA† Naoki FUKUTA† and Hiroshi ISHIKAWA†

†Department of Computer Science, Faculty of Informatics, Shizuoka University

3-5-1 Jouhoku, Hamamatsu-shi, Shizuoka, 432-8011 Japan

E-mail: †cs3061@s.inf.shizuoka.ac.jp, {fukuta, ishikawa}@inf.shizuoka.ac.jp

Abstract Today, with explosive spread of blogs and their timely publication and individuality of contents, they become a remarkable information source to know the trends of interests of societies. It becomes important to analyze information contained by blogs , to extract topics paid much attention to , and capture transitions of certain specific topics. We assume that a topic in a blog belongs to a specific interest field. In this paper, we call such an interest field “category.” It is important to analyze the relationships between several topics belonging to specific interest fields. One of our research goal is to extract transition patterns of a topics in the category. Another goal is to compare transitions of several strongly-related topics in each category. We propose a method that extracts transition patterns of topics in a category by clustering techniques. Here, topics are extracted from a set of blog articles. The category of the articles is determined by applying a hieratical clustering technique to topics in the category. Transition patterns are extracted based on timestamps of each topic and the number of blog articles in the category. We show that our proposed method can be applied effectively in several recent popular events in blog articles.

Keyword Blog, Clustering, Topic Transition

1. はじめに

新たな情報発信手段として，ブログが注目されている．総務省の調査によると，2005 年末ではブログ利用者は約 335 万人，ブログ閲覧者は約 1,651 万人いるとされ，2007 年末には利用者が約 782 万人，閲覧者が約 3,455 万人にまで達すると予測されている[1].

現在のブログの特徴としては，ウェブ上での個人の日記という側面と，特定のニュースやイベントに対する個人の意見を表現するメディアの 1 つという側面がある．これらはともに，ブログ中に記述される様々な興味分野に対する話題として現れる．その中には，大多数のうちで共有されるような興味分野もあれば，少数で共有されるような興味分野も存在する．話題がさまざまに移り変わっていくような興味分野もあれば，変遷が緩やかであるような興味分野や話題が一定に存在するような興味分野も存在する．ブログ全体の傾向だけでなく，興味分野に注目して傾向を解析することにより，話題変遷の特徴や，その興味分野において最も注目を浴びている話題などが抽出できるのではないかと考えられる． 本稿では，この興味分野をカテゴリと呼ぶことにする．

本論文では，ある特定の興味分野に関する話題の変遷を抽出することにより，話題の変遷が激しいカテゴリ，話題の変遷が緩やかなカテゴリ，などカテゴリごとの話題変遷に関する特徴パタンを抽出する手法を提案する．また，トピックをカテゴリにまとめ，そのカテゴリごとに変遷を抽出することにより，関連性の高い複数のトピックに関する話題の変遷パタンの比較も行えるようにすることを目指す.

2. 関連研究

ブログの話題抽出や話題変遷の抽出として，最近では特に多く書かれている話題について表示し，その頻度をグラフ化しているサービスが提供されている[2]．これは，複数のブログサイトのブログ記事の全体に対して，注目されている話題の抽出，ある特定の話題の変遷の抽出を行っている．これは，ブログ記事全体の中から，ある 1 つのトピックに着目しているという点において，本研究とは異なっている．

トピック抽出手法としては，burst の検出[3]が挙げられる．手法[3]および[4]では，ある語に対し，その語が出現する時間間隔の定常状態を求めておき，その時間間隔よりも短い間隔で語が出現しているとき，その語をトピックに関連する語として抽出する．手法[3]及び[4]では，急激に話題になったようなトピックの抽出を目的としたものであり，ほとんど変化なく取り扱われているトピックについてはうまく抽出できない．本研究では，ゆるやかに話題が変遷，もしくは，ほとんど変遷しないようなものについてもその話題を抽出することを目標としているため，手法[3]及び[4]では目的が達成されない.

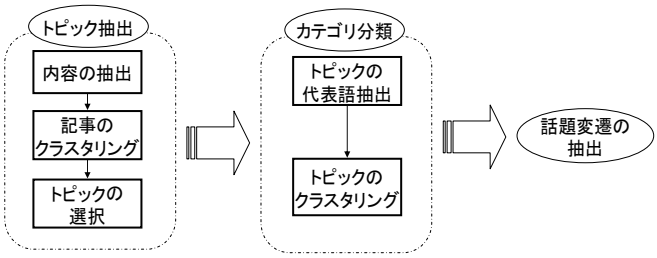

図 1：提案手法の処理の流れ

関口らは，特化した興味分野の話題を抽出する手法を提案している[5]．手法[5]は，関心を同じくする人々の間で話題となっていることを抽出することにより，相対的には少数にしか論じられていないような分野の話題を抽出可能とする．手法[5]では，著者の興味ベクトルを抽出し，それを用いて著者間の関連度を算出，その関連度が閾値以上となる著者間で話題となっている語を抽出することで，話題を抽出している．手法[5]では，著者間の関連度に基づいて話題を抽出しているのに対して，本研究では，著者やその間の関連度を特定できることを仮定せず，ブログ記事から直接，トピックを抽出することを目指すという点において異なっている．

本論文で提案する手法は, 複数のトピックに同時に着目でき，かつ，変遷が非常に緩やかな話題も扱うことができ，blog 記事の著者の同一性の特定が困難な状況でも適用可能な手法である点で，これらの関連研究と大きく異なる.

3. 基本アプローチ

提案手法の処理の流れを図 1 に示す．本手法では，カテゴリ別の話題変遷を抽出するために，最初にブログ記事からトピックの抽出を行う．次に，抽出したトピックにクラスタリングを適用することにより，カテゴリを生成する．続いて，生成したカテゴリごとに話題の変遷を抽出していく.

3.1. トピックの抽出

ブログ記事に形態素解析を行い，名詞を抽出する．抽出した名詞により文書ベクトルを生成することにより，各ブログ記事に記述されている内容を抽出する．生成した文書ベクトルに対してクラスタリングを行うことにより，記事のクラスタリングを行う．生成したクラスタをトピックとし，トピック中から話題として扱えそうなトピックのみを選択することによって，ブログ記事群からトピックを抽出することを行う.

3.1.1. 内容抽出

ブログ記事からそこに記述されている内容を抽出するために，ブログ記事の集合に対して，それぞれタイトルと本文の形態素解析を行う．形態素解析ツールとしては，Sen[6]を用いる．Sen によって形態素解析を行った結果から，抽出された名詞を用いてそれぞれのブログ記事ごとに文書ベクトルを作成する．ここ

表 1：名詞の結合有，結合無の例

結合有	空気 / 清浄 / 機 （一般＋サ変接続＋接尾）
	健康 / 食品 （形容動詞語幹＋一般）
	小泉 / 純一郎 （人名姓＋人名名）
結合無	明日 / テスト （副詞可能＋一般）
	3 / 回 / 目 （数＋接尾＋接尾）

で，ベクトル作成に用いる名詞には，非自立語，接尾語，数，代名詞を除くものとする．

連続して出現している名詞はもともと複合語である名詞が分割されたものと考えられるため，これらを結合し 1 つの名詞とみなした語もベクトル中に登録する．ブログ中においては，口語体のような記述がなされている記事や，句読点があまりつけられていないような記事も多く存在するため，複合語以外にも名詞が連続して出現する場合がある．名詞が連続する全ての場合に結合を行うと，本来ベクトル中に登録したい，記事中に現れる特徴的な複合語以外にも，不必要に多くの語数が登録されることとなってしまう．結合を行う場合は，続いて現れる名詞が副詞可能，非自立語，数，代名詞でない場合のみに行うこととする．結合を行う語のうち，接尾語に関しては，直前に出現している名詞に結合する場合のみ登録することとし，単独での登録は行わないようにする．Sen では人名は姓名詞と名名詞に分割されるが，姓名詞と名名詞が連続して出現するような場合には，一人の人物の姓および名であると考えられるため，姓と名を結合した人名名詞も，姓名詞・名名詞とあわせて，文書ベクトル中に登録することとする．名詞が連続して出現した際における，結合を行う場合と行わない場合の例を表 1 に挙げる．

ブログ記事ごとの文書ベクトルの各要素には一般的な TF･IDF を用いる．また，固有名詞はトピックを強く特徴付ける語が多く含まれると考えられるため，固有名詞に関しては特に重み付けを行って算出する値とする．あるブログ記事 E における単語 t の重み w_E^t は式 1 によって算出する．

$$w_E^t = w(t) * \frac{\log(tf(t,E)+1)}{\log(M)} * \log(\frac{N}{df(t)})$$

(式 1)

ここで，$w(t)$ は固有名詞に対する重み，$tf(t,E)$ はブログ記事 E 中に単語 t が出現する頻度，$df(t)$ は全ブログ記事中において単語 t が出現しているブログ記事数，N は形態素解析を行ったブログ記事の総数，M はブログ記事 E より抽出された単語の種類数を示す．

3.1.2 記事のクラスタリング

作成した文書ベクトル群に対して，凝集型の階層的クラスタリングを行う[7]．凝集型の階層的クラスタリングでは，初期段階としてそれぞれを 1 つのクラスタとみなし，それらを併合していくことによってクラスタを生成していく．最終的にはクラスタ数が 1 になるまで併合されていくが，クラスタリング終了の際に任意の閾値を設けることにより，任意の大きさのクラスタを生成する．ここでは，関連する内容を記述している記事をまとめることにより，トピックとそのトピックに関連する語を抽出することを目的としているため，生成されるクラスタがあまり大きくなりすぎないように閾値を設定する．階層的クラスタリングを終了する際の閾値は別途実験により決定する．また，計算量の軽減のため，類似度の算出にはベクトル中のすべての語を使用するのではなく，その単語の TF･IDF 値がある閾値以上のもののみを使用することとする．この閾値は 4.1.2 にて決定する．

3.1.2.1 ブログ記事間の類似度算出

文書ベクトルの一般的な類似度算出式では，それぞれのベクトルの大きさによる正規化を行っている(式 2)．ブログの特徴のひとつとして，ひとつの記事中に二つ以上のトピックについて言及しているような場合が多く存在することが挙げられる．これは，長い記事であってもそのトピックに関することが必ずしも多く記述されているわけではないということを意味する．

今回は，ブログ記事間の類似度を算出することにより，同一のトピックについて記述している記事を検出することを目的としている．1 つの記事で，複数のトピックについて記述されているような場合においても，記事の長さに影響されず抽出が行えるほうが本研究では望ましい．したがって，ブログを対象とする場合には，ベクトルの大きさによって正規化しないほうが良い結果が得られる可能性がある．

$$\mathrm{sim}(E_\mathrm{i}, E_\mathrm{j}) = \frac{w_\mathrm{i}^1 w_\mathrm{j}^1 + ... + w_\mathrm{i}^\mathrm{m} w_\mathrm{j}^\mathrm{m}}{\sqrt{(w_\mathrm{i}^1)^2 + ... + (w_\mathrm{i}^\mathrm{m})^2} * \sqrt{(w_\mathrm{j}^1)^2 + ... + (w_\mathrm{j}^\mathrm{m})^2}}$$

(式 2)

$$\mathrm{sim}(E_\mathrm{i}, E_\mathrm{j}) = w_\mathrm{i}^1 w_\mathrm{j}^1 + ... + w_\mathrm{i}^\mathrm{m} w_\mathrm{j}^\mathrm{m}$$

(式 3)

予備実験として，ブログ記事 E_i，E_j における類似度を式 2 と式 3 において比較を行った．検証データとしては，同一トピックに関する記述が含まれているブログ記事 5 件と無作為に抽出したブログ記事 995 件，計 1000 件のブログ記事を用いた．検証実験はそれらのブログ記事間での類似度算出により行った．関連するブログ記事として選択したブログ記事には，さまざまな長さの記事を含ませている．加えて，トピックの関連語が出

現しているが，異なるトピックについて記述されているもの，複数のトピックについて記述されているものについても含ませている．これらのような特徴の記事を含ませることにより，ブログ記事の長さによる影響，複数トピックの含有による影響などを計ることができると考えられる．

予備実験の結果，式3によって求めた類似度が式2に比べ，トピックの抽出という目的に対してより精度がよいことが確認された．よって，本研究では，ブログ記事E_i，E_jにおける類似度は式3に基づいて算出することとする．

3.1.3 トピックの選択

ブログ記事のクラスタリングにより抽出したトピックのうち話題として扱えそうなトピックを選択する．話題として扱うためには，ある程度の記事間で共有されているようなトピックでなくてはならない．1 つのトピックのクラスタに含まれる記事がごく少数(特に 1 つしか含まれない場合)であるようなトピックは，話題としては適さないと考えられる．この基準により話題として適さないと判断されたトピックは，次のフェーズであるカテゴリ分類の対象とはしないこととした．

3.2. カテゴリの分類

選択したトピックから代表語を抽出し，その代表語により各トピックの文書ベクトルを生成する．生成した文書ベクトルにクラスタリングを適用することにより，トピックの集合であるカテゴリを生成する．

3.2.1 トピックの代表語の抽出

選択したトピックから代表語を抜き出し，その代表語によって文書ベクトルを作成する．代表語はトピックを特徴付ける語のうち，そのトピックと関連した別のトピックにも共通に現れると考えられる語を抽出する．代表語は，トピックの中に出現した上位の語のうち，人名・地域名を除いたものとする．人名・地域名はトピックに特化したものであることが多いため，これらを用いるとトピック間の関連度がうまく算出されず正しくカテゴリが生成できない可能性がある．固有名詞のうち組織・一般に分類される語(例：国連，衆院など)については，同一カテゴリにまとめられる複数のトピック中に比較的共通して現れやすいと考えられるため，代表語として抽出することとする．このように代表語を単に重みの高い語ではなく，人名・地名を除去することにより，トピックは実際の出来事に応じて細分化され，カテゴリはそれらを全て含有するものとして生成可能であると考えられる．

また，代表語によって作成する，トピックTの単語tにおける文書ベクトルの要素の重みC_T^tは，トピック中に含まれる全記事においての単語tの重みの合計をそのトピック中に含まれる記事数で割ったものとする．式4に基づいて算出する．

$$C_T^t = \frac{\sum_i^n w_{Ei}^t}{n} \qquad \text{(式 4)}$$

ここで，E_iはトピックT中に含まれるブログ記事，nはトピックT中に含まれるブログ記事の総数，w_{Ei}^tは式1で求めたブログ記事E_iにおける文書ベクトル中に登録されている単語tの重みである．

3.2.2 トピックのクラスタリング

各トピックから抽出した代表語に基づいて生成した文書ベクトルを用いて，トピックの抽出時と同様に凝集型階層的クラスタリングを適用する．階層的クラスタリングを終了する際の閾値を変化させることによって，カテゴリの大きさを任意に変更可能とし，含まれるトピックの関連度を任意に変更することが可能である．また，トピック抽出の際のクラスタリングの閾値とあわせて変更することにより，関連性の小さなトピックを同一カテゴリに含ませる際にも，含まれるトピックの，数が多くなりすぎないようにすることも可能である．

トピックの抽出においては，実際の出来事に基づいた，より細分化されたピックを抽出することを目的としており，カテゴリ生成では，それらのトピックをまとめあげることを目的としている．この理由から，トピックの抽出の際には，重み付けをした固有名詞を用いているが，カテゴリ生成の際には人名・地名を除くことにより，類似度算出に対する基準を緩くし，関連したトピックが複数のカテゴリに細分化されてしまうことを避けることが出来ると考えられる．

3.3. 話題の変遷の抽出

トピックのクラスタリングによって生成したそれぞれのカテゴリごとに，話題の変遷を抽出する．変遷の抽出はトピックのタイムスタンプに基づいて行う．トピックに含まれている記事についてタイムスタンプごとの分布を求め，同一カテゴリ中のものを重ね合わせていく．このようにすることにより，複数のトピックが同時に出現している場合に，それぞれの時期の重なり合いなどが抽出できると考えられる．

カテゴリ中に含まれるそれぞれのトピックに対して，そのトピック中に含まれる記事のタイムスタンプを抽出する．抽出するタイムスタンプは年月日及び時刻とする．トピック中の記事を抽出したタイムスタンプに基づいて分類する．1 日，1 週間，など時間を任意の単位に区切り，その単位ごとにトピック中の記事数を数え上げる．区切った時間の単位と，その単位ごとの記事数をグラフ化することにより，カテゴリ中の各トピックの注目の度合いを可視化していく．可視化することによって，同一カテゴリ中に含まれる，関連した複数のトピックの変遷パタンの比較を行うことができると考えられる．関連した複数トピ

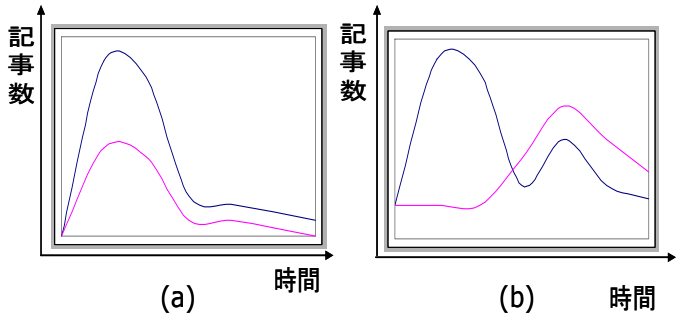

図 2(a),(b)：同時性可視化例，影響可視化例

ックの変遷の可視化イメージを図 2(a)(b)に示す．図 2(a)では，関連したトピック同士が，同時期にある種の盛り上がりを見せていることを示しており，関連したトピックの同時性が抽出可能であると考えられる．図 2(b)では一度収束したトピックが，関連した別のトピックに刺激されて再び活発になる様子を示しており，関連したトピック間での影響がどのように及ぼされているかが抽出可能であると考えられる．

4. 実験

本実験では，データセットとして，クローラで収集した livedoor ブログの記事 42,439 件(2006 年 7 月 3 日～2007 年 1 月 31 日)を使用する．これらのうち，「携帯」の語を含む記事 8,787 件に対してトピックの抽出，カテゴリの生成，変遷の抽出を行う．

本手法の有効性の評価は生成されたカテゴリの妥当性，および抽出された変遷パタンの比較によって行うこととする．生成カテゴリの妥当性の評価については，同一カテゴリ中に含まれるトピック間の関連度を評価する．トピック間の関連度に基づいて，関連性の高いものが含まれているかどうかを判断し，カテゴリの妥当性を評価する．変遷パタンの比較については，カテゴリ内での変遷の同時性やトピック間での影響などをどのように抽出できるかによって評価する．

4.1 各種パラメータの設定

実験に際して，使用する各種パラメータの設定を行う．本実験において，使用するパラメータを次のように設定する．

トピック抽出段階では，固有名詞に対する重み付けを式 5 のように設定する．

$$w(t)=\begin{cases}2 & (t=\text{固有名詞})\\ 1 & (t\neq\text{固有名詞})\end{cases}$$

(式 5)

トピックの選択において，本実験ではトピックの変遷についても抽出するため，トピックとして抽出する際の，クラスタ内に含む最低の記事数は 5 エントリとした．

トピックの変遷の可視化において，変遷を抽出する粒度としては，今回は対象としている期間が短いことから，1 日を基準とすることとした．

トピック抽出段階およびカテゴリ生成段階においての類似度算出に使用する語句の閾値の決定，およびトピック抽出段階およびカテゴリ生成段階における凝集型階層的クラスタリング終了閾値の決定の方針，および設定した閾値に関して，以下に記述する．

4.1.1 類似度算出に用いる語の TF･IDF 閾値の決定

計算量の軽減のため，類似度算出にはベクトル中の全ての語を使用するのではなく，その単語の TF･IDF 値がある閾値以上のもののみを使用することとする．予備実験として，無作為に抽出した 18,699 件のブログ記事に対して，形態素解析の結果抽出された名詞に関して，重み付き TF･IDF 値の分布を求めた．18,699 件のブログ記事を形態素解析して得られた語句，586,321 語の名詞について式 1 に基づいて TF･IDF 値の分布を算出した．固有名詞に対する重み付けは，4.1 パラメータの設定に記載してある条件と同様に，固有名詞に対しては通常の TF･IDF 値を 2 倍することによって重み付けすることとした．

予備実験の結果，TF･IDF 値の最大値は 27.429，最小値は 0.120 であった．また，最も分布が高いような TF･IDF 値は 0.8～0.9 の間であった．よって，この範囲よりも TF･IDF 値の重みが小さいような語に関しては，そのブログ記事を強く特徴づける語ではないと推測される．無作為に抽出したブログ記事 100 件に対して，それぞれのブログ記事の抽出語数のうち，その語の TF･IDF 値が 0.8 以上となる語句の割合を調べた．調査の結果，各ブログ記事中の約 72.1%の語において TF･IDF 値が 0.8 以上となることがわかった．抽出語句の半数以上が含まれているため，ブログ記事を特徴付ける語はこの閾値を用いた場合でも十分に含まれていると考えられる．本実験における類似度算出に用いる文書ベクトル中の語は，トピック抽出段階においては 0.8 以上，カテゴリ生成段階においては 0.4 以上の語のみとすることとした．

4.1.2 クラスタリング終了閾値の決定

本手法では，階層的クラスタリングの際に最長距離法を用いるため，階層的クラスタリングを終了する際の閾値は，各ブログ記事から抽出される名詞数とそれらの重みに基づいて算出することとする．

類似度算出に使用する TF･IDF 値の閾値の決定と同様に，無作為に抽出したブログ記事 18,699 件に対して形態素解析を行い，それぞれのブログ記事から名詞を抽出した．ここで,抽出された名詞には 3.1.1 に示した複合語も含む．

予備実験の結果，1 つの記事からの最大の抽出名詞数は 1,751 語，最小の抽出名詞数は 0 語であった．また，1 つの記事中から抽出される名詞数は 5～15 個である記事が最も多かった．抽出名詞のうち，約 3 分の 1 程度が共通であれば同じトピックについて記述しているのではないかという仮定に基づき，今回は 5 語程度の名詞が共通であるような場合に同一トピックとして抽出可能であるように閾値を設定する．4.1.1 類似度算出に用い

表 2：生成されたカテゴリ

生成カテゴリ数		26	
カテゴリ	トピック数	ラベル	特徴語
カテゴリ 1	6	ソフトバンク, MNP	ソフトバンク, ドコモ, DDI, NTT ドコモ, 通話
カテゴリ 2	4	ネットで稼ぐ方法	起業, 商, 特典, 方法, ノウハウ
カテゴリ 3	3	携帯ストラップ	ラップ, スト, ストラップ, サンプルショップ, ペア
カテゴリ 4	3	金融広告	創, 英, ユニー, スルガ銀行, ローン
カテゴリ 5	3	求人情報	求人, ハローワーク, アルバイト, クローズアップページ, 吉宗

表 3：番号ポータビリティに関連したカテゴリ内のトピック

カテゴリ特徴語		ソフトバンク, ドコモ, DDI, NTT ドコモ, 通話	
カテゴリ内トピック数		6	
トピック	記事数	ラベル	特徴語
トピック A	42	ソフトバンクの CM について	ソフトバンク, ディアス, ブラッド, 孫, 予想外
トピック B	39	各会社の料金プランについて	ソフトバンク, 通話, プラン, ドコモ, 円
トピック C	21	MNP 全般 電波状況など	ドコモ, ソフトバンク, 電波, ビックカメラ, 端末
トピック D	13	MNP ソフトバンク受付停止	ソフトバンク, 番号, 停止, 申し込み, 変更
トピック E	12	MNP au 一人勝ち	DDI, ソフトバンク, NTTドコモ, 川井, 川井徹
トピック F	11	携帯機種変更について	ソフトバンク, 機種変更, 機種, プラン, 変更

表 4：Web 上での商業に関連したカテゴリ内のトピック

カテゴリ特徴語		起業, 商, 特典, 方法, ノウハウ	
カテゴリ内トピック数		4	
トピック	記事数	ラベル	特徴語
トピック G	39	メールマガジン	円, 不定期, 週間, 毎日, マガジン
トピック H	35	ネットで起業	菅野, ムリ, 菅野一, 宮川, 一
トピック I	24	アフィリエイト, ブログ広告	商, 商材, 岡田, 無料, 田渕
トピック J	10	ネットで収入(兼業)	片山, 羽根田, 趣味起業, 対談, 趣味

る語の TF･IDF 閾値の決定においての予備実験より，TF･IDF 値の分布のピークは 0.8～0.9 の範囲においてであることがわかる．よって，本実験でのクラスタリング終了閾値はこれらを用いて，(0.9*0.9)*5 = 4.05 とした．本実験においては，トピック抽出段階及びカテゴリ生成段階ともに同様の閾値を用いて実験を行うこととした．

4.2 トピック抽出及びカテゴリ生成

「携帯」の語を含む記事 8,787 件に対して，トピックの抽出及びカテゴリの生成を行った結果，生成されたカテゴリについてまとめたものを表 2，生成されたカテゴリのうち，カテゴリ内に含まれているトピック数の上位 2 つのカテゴリに関して，カテゴリ内の各トピックについて示したものをそれぞれ表 3，表 4，に示す．これらの表中における特徴語とは，それぞれそのカテゴリ・トピック中においてその語の重みが高いもの上位 5 語のことを示している．語の重みとは，それぞれ 3.2.1 式 4 にて算出した値のことである．各カテゴリ・トピックのラベルとは，クラスタ内に含まれる記事やトピックの内容に基づいて，人手によってつけたものである．

表 2 では，トピックのクラスタリングによって生成されたカテゴリのうち，カテゴリ中に含むトピック数が多いものの上位 5 つのカテゴリについて示している．表 2 中の生成カテゴリ数とは，トピックのクラスタリングによって生成されたクラスタのうち，2 つ以上のトピックを含んでいるようなクラスタの数を表している．

表 3 は，生成されたカテゴリのうち，ソフトバンク・番号ポータビリティに関連する話題についてのカテゴリに含まれるトピックについて示している．このカテゴリは，携帯に関連するトピックのうち，ソフトバンクや番号ポータビリティに関連するトピックによって構成されているカテゴリを示している．これらのトピックは，番号ポータビリティという，ある現実の 1 つのイベントに対する，さまざまな側面に関するトピック

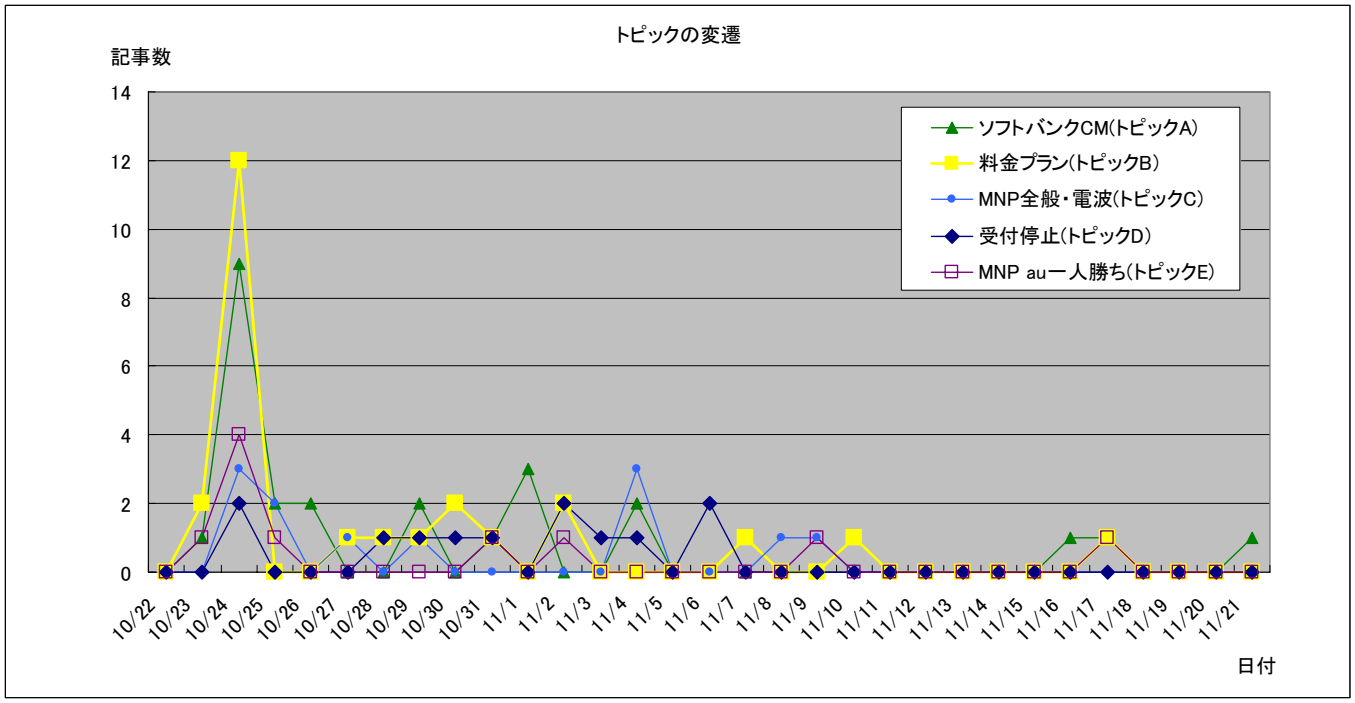

図 3：番号ポータビリティに関連するカテゴリ内のトピック変遷パタン

であるということが出来る．

表 4 にあるインターネット上での起業やアフィリエイトなどについてのトピックが含まれている．これらは総じて，Web 上での商業に関するトピックであることがわかる．これらのトピックは，Web でのメールマガジンやアフィリエイトの普及によって Web 上での商業がより身近になり，それらの情報の需要が出てきたため，出現してきたものであると考えられる．表 2 のそのほかのカテゴリに関しても，広告や，求人情報をまとめたブログサイトなどといった，いわゆるブログスパムの類によって構成されているカテゴリであるが，比較的同様の話題に関するトピックが含まれていると考えられる．これらのカテゴリ中に含まれているトピックの関連度は高いと考えられるため，生成されたカテゴリは妥当なものであると判断できる．

4.3 変遷パタンの抽出

表 3 で得られたカテゴリ内のトピックに関して，話題の変遷を図示したものを図 3 に示す．カテゴリ内のトピックのうち，含有記事数の多い上位 5 つのトピックについて，変遷を抽出した．変遷を抽出した期間としては， 2006 年 10 月 22 日～2006 年 11 月 21 日の間の 1 ヶ月を対象とした．

図 3 の結果によると，10 月 24 日に急激に各トピックの記事数が増えていることから，このカテゴリに関連する話題がこの時期に一斉に注目を浴びていることがわかる．これは，番号ポータビリティが 10 月 24 日に開始されたことに影響をうけていると考えられる．加えて，トピック B がカテゴリ中のその他のトピックよりも先に記事数が増加していることがわかる．これは，番号ポータビリティ開始日の前日の 10 月 23 日にソフトバンクが新料金プランを発表したことに起因していると考えられる．

10 月 26 日～10 月 27 日の付近で一度カテゴリ内のトピック全体的に話題が収束しているが，10 月 28 日以降，またしばらくの間，特定のトピックに関しては議論が行われるようになっている．これは，10 月 28 日から 3 日間にかけて，ソフトバンクの番号ポータビリティの受付が停止したことによるものと考えられる．受付を停止したことを記述しているトピック D に加え，ソフトバンクの CM に関するトピック A が発展しているのに対し，番号ポータビリティで au が一人勝ちであるということに関するトピック E はあまり発展していないことなどから，互い同士が関連しているトピックであっても，現実のある出来事に対して常に同じ反応を示すとは限らないということがわかると考えられる．

全般的にトピック B の各会社の料金プランに関するトピックがよく現れていることがわかる．このことより，番号ポータビリティに関する話題の中でも，料金プランについて一般的に多く関心が寄せられている話題であるということが考えられる．このことより，同一の出来事に関するトピックであっても，その中でどのような側面に対して最も関心があるかなどが抽出可能といえるのではないかと考えられる．

このように，同一カテゴリでもすべてのトピックが現実の出来事に反応したわけではなく，特定のトピックのみが変化していることがわかる．このことは，カテゴリを細分化して複数のトピックにすることに意味があること，つまりトピックの粒度がある程度適切であることを示唆していると考えられる．

本実験においては，同一トピック内に含有されている記事数

が少数であったため，各トピックの変遷を抽出した際に，少数の記事の変化に対しても大きく影響が出てしまっていた．これは，今後，大量データに対して本手法を適用した際には改善されると予想される．

5. おわりに

ブログ記事からトピックを抽出し，それらのトピックをカテゴリに分類することにより，カテゴリごとの話題を抽出する手法を提案し，予備的実験を行った．今後の展望としては，大量データへの適用行い，さまざまな分野のカテゴリに関して変遷を抽出し，関連するトピックの同時性だけでなく，トピック間で影響を及ぼしている事例などの抽出も行うことを目指す．またその他の課題としてはトピックの抽出方法，および，話題の変遷の扱い方についての 2 つが挙げられる．

トピックの抽出方法に関しては，現段階では，TF･IDF で重み付けした文書ベクトルの類似度でのクラスタリングにより行っているが，それに加えて、タイムスタンプを考慮したクラスタリングを行った場合に結果にどう影響が出るか，比較・検討を行っていきたい．

また，話題の変遷の扱い方については，現段階では話題の変遷を，そのトピックの記事数を用い，注目されている度合いのみに着目して行っている．しかし，単純に記事数のみで計るのではなく，そのほかに，よく共起している単語に関しても同時に調べることなどが考えられる．このことにより，カテゴリ中において，そのトピックがどれだけ関心を寄せられているかという注目の度合いだけでなく，そのトピックに関してどのような議論が展開されていっているかなどの，そのトピックに関する意見の変遷なども抽出可能となる手法も検討していきたい．

参考文献

[1] 総務省，“ブログ・SNS の現状分析及び将来予測” http://www.soumu.go.jp/s-news/2005/pdf/050517_3_1.pdf ，2005.

[2] kizashi.jp http://kizasi.jp/

[3] Jon Kleinberg，“Bursty and Hierarchical Structure in Streams”In Proc. the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.

[4] 藤木稔明､南野朋之､鈴木泰裕､奥村学，“document stream における burst の発見”，情報処理学会研究報告，2004-NL-160, pp.85-92.，2004.

[5] 関口裕一郎、川島晴美、奥田英範、奥雅博，“コミュニティ構造を利用した話題ナビゲーション手法の検討” DEWS2006 1B-oi1，2006.

[6] 形態素解析システム Sen http://ultimania.org/sen/

[7] 石川博，“次世代データベースとデータマイニング” 第 6 章 クラスタリング，CQ 出版社，2005.