4 CGMマイニングと知識化

山西 健司　NEC 共通基盤ソフトウェア研究所
森永 聡　NEC 共通基盤ソフトウェア研究所
松村 憲和　NEC 共通基盤ソフトウェア研究所

本稿では，テキストマイニング技術を用いて CGM（consumer generated media）情報から知識化を行う枠組みについて解説する．CGM マイニングにおいては，1）トピックのダイナミクスを捉えること，2）トピックの共通文脈を捉えること，3）分散ヘテロな情報を俯瞰すること，といった問題が重要である．これに対して，それぞれ，動的トピック分析，文脈マイニング，分散協調マイニングといった技術によって解決できることを示す．本枠組みの有効性を，BIGLOBE 旬感ランキングにおける事例などを用いて示す．

なぜ CGM の知識化が重要か？

Web2.0 に総称されるユーザ参加型の Web サービスが急激に発展している．その中でブログや SNS を通じてユーザが発信した情報 ── CGM（Consumer Generated Media）──が大きな影響力を持つに至っている．Web1.0 時代では，TV や新聞などのメディアが一方的に情報を発信していたが，Web2.0 時代では，ユーザ側からも情報を発信するようになり，そこで形成されたクチコミや「集合知」は逆にメディアに影響を与える場合も出てきている．このような CGM の傾向を抽出する営みを，ここでは「CGM マイニング」と呼び，得られた傾向を活用できる知識に変えることを「CGM の知識化」と呼ぶことにする．CGM マイニングと知識化はどのような局面で重要なのであろうか？

まず，それは企業や個人の活動を効果測定する手段として重要である．たとえば，企業がある商品を発売し，CM やキャンペーンなどのプロモーションを施したとしよう．その効果を知るには，従来は売上など直接消費行動に結びついた購買データで測定するしかなかった．ところが現在では，消費者が商品評価を CGM として生み出し，購買に影響を与えている．そこに，従来の購買データからは見えてこない，マーケティングに直結した消費者の「本音」が表れている．そのような潜在情報を引き出すことが，将来の商品企画や経営判断の鍵となっていくと考えられる（**図 -1**）．

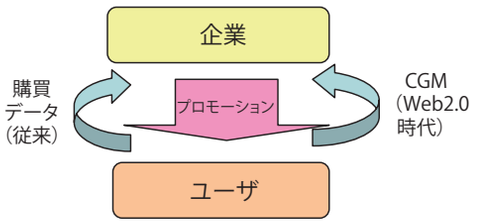

図 -1　プロモーション効果測定の手段としての CGM

さらに，CGM の知識化は企業や個人のリスク管理の手段としても重要である．実際，企業や個人に関する風評は，しばしばメディアより速く CGM として伝播する．そこで，企業に不利益となる風説が立ったときには，経営者は CGM からできるだけ早期にこれに気づき，リスクに対応することが CSR（Corporate Social Responsibility=「企業の社会的責任」）上必要になっているのだ．

上記以外にも，教育や政治などの現場においても教師や政治家の評判を知るのに CGM を活用することが試みられている．

さて，CGM マイニングと知識化は具体的には何を意味するのであろうか？　さまざまなアプローチがあるだろうが，本稿では，3 つの側面からの CGM の知識化に対するアプローチについて述べる．

1）トピックのダイナミクスを捉えること

CGM の世界では，時間とともに新しいトピック（話題のかたまり）の形成や古いトピックの消滅が起こり，知

識のうねりを生み出している．そのようなトピックのダイナミズムを生み出すことは CGM の本質であり，この流れを捉え，可視化することによって，CGM の知識化にアプローチする．

2)トピックの共通文脈を理解すること

CGM のほとんどは口語表現である．そのトピックは，さまざまな表現で同じようなことが語られている．そこで表記のゆれを超えて，共通して語られていることは何か？を理解することは CGM の知識化の重要なポイントである．しかも単語や係り受け関係レベルではなく，文脈レベルで把握することが望まれる．

3)分散ヘテロな情報を俯瞰すること

CGM はさまざまなネットコミュニティの中に分散して存在する．しかも TV や新聞などの報道情報等の異種の情報と互いに影響を及ぼし合っている．これらの分散かつ多様な情報の関係性を俯瞰することで，初めて CGM の真の姿が浮き彫りになる．このような立場から CGM の知識化にアプローチする．

上記 1）～ 3)のアプローチにはデータマイニング，機械学習といった分野において技術的なチャレンジを必要とする．よって，CGM の知識化はそのような技術の発展を促す上でも重要な課題であるといえる．

本稿では，上記 1）～ 3）のアプローチを通じて CGM の知識化の実際を示す．次に，それらの方法論を統合した実際のブログ・TV・検索を融合した統合分析サービスの実例として，「BIGLOBE 旬感ランキング」の「分析コーナー」の事例を紹介する．

トピックのダイナミクスを捉える

CGM は Web 空間上でダイナミックにトピックの潮流を生み出している．その傾向を把握し，今どのような変化や流れが起きているのか，という時間的差分情報を発見することが CGM の知識化にとって本質的に重要であると考えられる．ここで，トピックとは特定の事象や活動について述べたテキスト群を意味する．

そこで，本章ではトピックのダイナミクスを捉えるさまざまな研究がある中で [1), 3)]，Morinaga and Yamanishi による**動的トピック分析** [3)] の枠組みを紹介する．

CGM の素材はテキストデータである．この傾向を知るための最も基本的な分析は，類似なテキスト同士をまとめあげる「テキストクラスタリング」である．その際の 1 つのクラスタ（話題の塊り）が 1 つのトピックに相当する．ただし，CGM は時間とともに流れ入るストリームとして捉えなければならない．動的トピック分析では，テキストストリームからダイナミックにテキスト系列をクラスタリングし，トピックの成長や衰退など，トピック構造の時間変化を検出する．

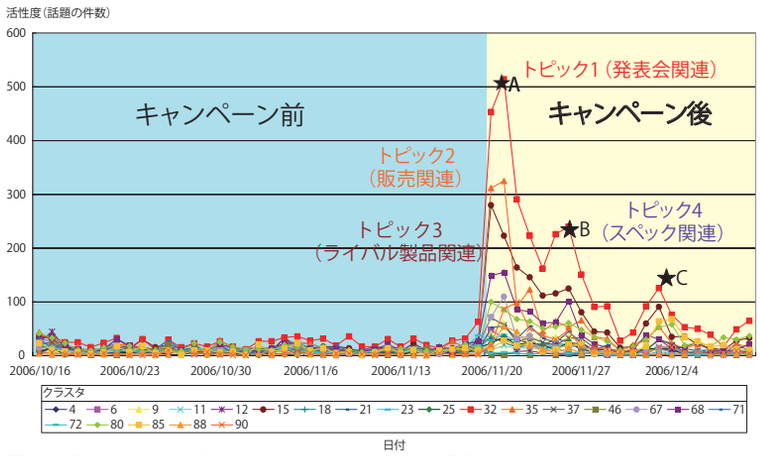

図 -2　動的トピック分析によるキャンペーン効果測定

動的トピック分析のイメージを明らかにするために，実際の CGM データを基に実行した結果を示す（**図 -2**）．

ここで用いたデータは某製品を取り上げ，これが発売された際に，関連話題について述べたテキストを，そのキャンペーン前後 3 カ月に渡ってブログをクローリングすることにより抽出したものである．総数は約 55,000 記事であった（本分析は（株）博報堂の協力を得たものである）．

図 -2 の横軸は日付を，縦軸は活性度（話題の件数）を示している．折れ線はいずれも動的トピック分析により自動的に抽出されたトピックについて，その活性度の時間的推移を計算した結果を示している．

活性度の大きい上位 4 つを取り上げると，「発表会関連」，「販売関連」，「ライバル製品関連」，「スペック関連」といったトピックが抽出されており，図ではそれぞれがキャンペーン後に盛り上がって，その後沈静化する様子を示している．特に，「発表会関連」のトピックに関しては，時間とともに 3 つの局所的なピーク（★で表示）を迎えながら沈静化したことが分かる．それぞれのピークでは

★ A「某俳優が発表会にゲスト出演した」

★ B「某社と人形を景品とする共同 PR 活動を行った」

★ C「月間目標売上を達成，好調な滑り出しと発表」

といった共通文脈が発見された．これは後に述べる文脈マイニング機能によって自動的にマイニングした結果である．また，「ライバル製品関連」のトピックは他のトピックに比べて沈静化の過程が緩やかで，途中さまざまな競合車種との比較話題で盛り上がったことを示している．

以上，動的トピック分析の枠組みを通じて得られた分析結果が，商品の発表後の CGM の動きを知識化していることを見ることができる．

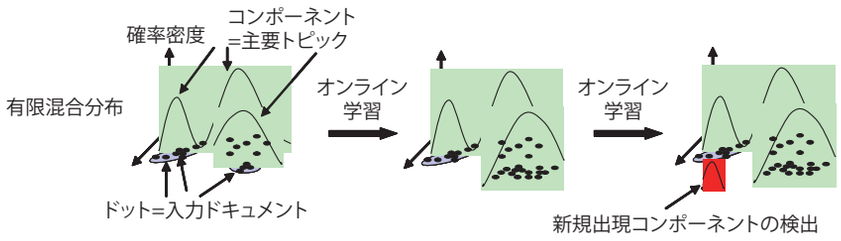

図 -3　動的トピック分析の流れ

動的トピック分析の原理

動的トピック分析の枠組みで，CGM の動的な側面を知識化できることを見た．ここでは，その原理を簡単に述べる．動的トピック分析では，テキストの時系列データを入力として，以下のタスクを実行する．

1）**トピック構造同定**：テキストのクラスタリングをダイナミックに行う．各時刻でどんなトピック（クラスタ）がどういう割合で存在するかといった構造をオンラインで学習する．

2）**トピック出現検出**：新しいトピックの出現や既存トピックの消滅など，クラスタの数の増減を伴うトピックの構造的な変化を検出する．

これらの基本タスクを実行するのに，以下のように問題を定式化する．

A）**モデル**：テキストを単語の出現頻度あるいは tf-idf 値を要素とする多次元ベクトルとして表現し，テキストの出現に関する確率分布を**有限混合モデル**を用いてモデリングする．ここで，有限混合モデルとは，いくつかの確率分布を線形結合して得られる確率分布である．**図 -3** に示すように，いくつかの山の重ね合わせとして表される．有限混合モデルを構成する各成分は 1 つのトピックを表す．また，混合比はトピックの出現確率分布を表す．ここで，各成分に対応する確率分布は正規分布ないしはバイナリ分布で表す．

B）**学習**：上記確率分布を，**忘却型学習アルゴリズム**によってオンライン学習する（図 -3）．ここで，忘却型学習アルゴリズムとは，過去のデータの影響を徐々に忘却していくことによって時変な構造を適応的に学習するアルゴリズムである．これによって，有限混合モデルの各成分のパラメータの値や，その混合比を各時点で求める．（⇒タスク 1）

C）**最適トピック数選択**：時間とともに推移するトピックの最適な数を**動的モデル選択** [6] によって求める．ここで動的モデル選択とは，有限混合モデルの混合数の最適値をダイナミックに求める統計的モデル選択の新しい機能である．混合数の増加を検出することで新しいトピックの出現を検出することができる．（⇒タスク 2）

上記枠組みに従って，トピック構造およびその変化をダイナミックに捉えることができる．

トピックの文脈を理解する

CGM の 1 つの特質は，口語で記述されるため，1 つのトピックがさまざまな表現で語られる，ということである．このような表記のゆれによらず，トピックの中で共通して語られている文脈を発見することは CGM 理解の本質につながる．その技術的枠組みとして**文脈マイニング** [4] の枠組みを紹介する．

文脈マイニングのアウトプットイメージは，先のキャンペーン分析の例で示した通りである（「発表関連」トピックの 3 カ所のピーク★の文脈）．

文脈マイニングが行うことは，各トピックに偏って多く出現する，共通の構文表現を抽出することである．その流れを**図 -4** に示す．つまり以下を行う．

1）同一のトピックに含まれるテキストに対して，文節間の主述，修飾関係などの構文木構造を解析し，

2）構文表現は表記のゆれを伴ったものでも同一視し，かつそのような構文表現の可能性を網羅する大規模な部分構文木空間を構成し，

3）その中から最もそのトピックに関して偏って現れている部分構文木を情報量規準に従って効率的に抽出し，

4）最後に日本語表現に再生して出力する

ここで 3）の偏りの大小を測る情報量規準としては**情報尺度 ESC（Extended Stochastic Com-plexity）** [5] を採用している．このような文脈マイニングは表現の多様性を持つ CGM の中から「いったい何が言われているか」を大雑把に把握することを可能にする．

分散ヘテロ情報を俯瞰する

複数のリモートサイトに分散している多様なテキスト情報を統合して，全体のトピック構造（これを**グローバルトピック構造**と呼ぶ）を俯瞰する問題を考える．たとえば，複数の異なるコミュニティの CGM を統合して，全体としての共通性や個別性を把握し，個々のローカルなトピックを全体の中で位置づけることを考える．この

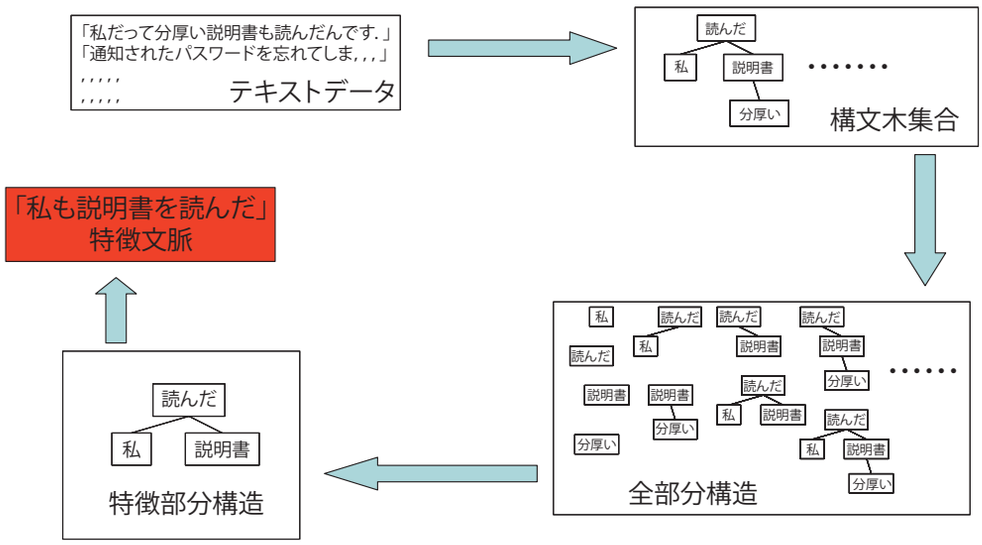

図 -4　文脈マイニングの流れ

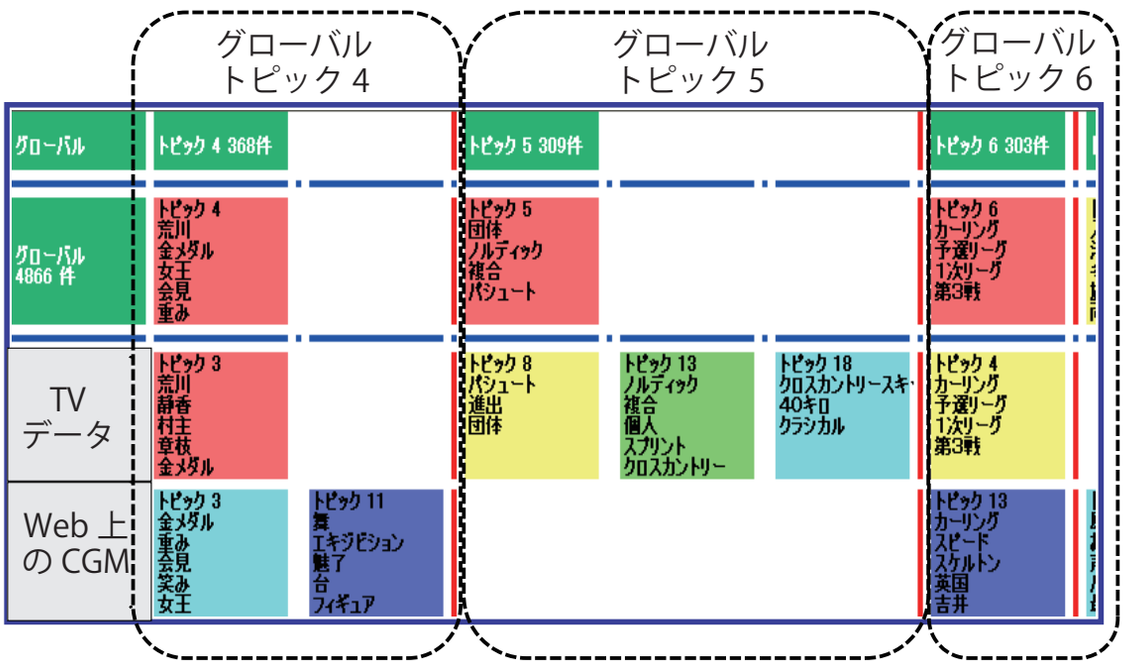

図 -5　トピック関係マップ

ような分析の枠組みとして**分散協調トピック分析** [5] の枠組みを紹介する．ここで前章同様，トピックとは似たテキストの塊であり，テキストクラスタリングにおけるクラスタに相当するものとする．

分散協調動的トピック分析のイメージを明らかにするために，Web 上の CGM と TV 報道の書き起こしデータを 2 つのソースとして，これらを統合し，全体を俯瞰する分散協調トピック分析の事例を示そう．

ここで用いたデータは，2005 年のトリノオリンピックに関連した，TV 報道の書き起こしデータ（プロジェクト社提供）と Web 上の CGM である．

図 -5 は TV データと Web データ（CGM）のそれぞれのトピック構造と全体を統合したときのグローバルトピック構造との関係を示している．各トピックの中の単語はトピックの特徴語を表す．たとえば，グローバルトピック 4 には TV データのトピック 3 と Web データのトピック 3 とトピック 11 が対応していることが分かる．また，グローバルトピック 5 は，TV データのトピックのみからなり，それは 3 つのトピック（8,13,18）から構成されている．

図 -6 では，共通トピックの 1 つである「フィギュアスケート荒川静香」について，これを構成する TV データのトピックと Web データのトピックの活性度合いの時間的変化を示している．この場合，Web で話題が活性してから TV が追随している様子が分かる．これは Web から TV への影響があったことを示唆する．吹き出しの

文章は，前出の文脈マイニングによって求めたものである．

以上，分散協調トピック分析の枠組みを通じて，異種の情報（TV データ）と掛け合わせた全体像の中から CGM と TV 報道との影響関係をマイニングできることが分かる．

分散協調トピック分析の原理

上述の分散協調トピック分析の枠組みで，CGM と他の情報を掛け合わせることで，全体を俯瞰し，CGM の相対的な位置づけを知ることができた．ここでは，この枠組みの原理を簡単に述べる．

分散協調トピック分析は，各サイトの個別性と全体性をプライバシーを保護した分散学習の方法に基づいている．このような研究がさまざまある中で[2), 7)]，ここでは松村らの研究[7)]を紹介する．そこでは，1）ヘテロ性を持つ複数の分散したデータに対して，2）生データを 1 カ所に集めることなく（プライバシーの保護），3）できるだけ少ない通信量で，4）生データを 1 カ所に集めた場合と同等の精度でグローバルトピック構造を推定する，ことを目標とする．これを達成するために以下のようなプロセスに従って，情報を統合分析する（**図 -7**）．

1）各サイトでのトピック分析：本稿の動的トピック分析の方法に従って，各サイトのトピック構造を同定し，そのパラメータのみをセンターに送る．

2）センターでの情報集約：ヘテロな情報を統合するための辞書知識を用いながら，各サイトから送られたパラメータのみに基づいて，センターでは各サイトのトピックを単純に重ね合わせた有限混合モデルを作る．これを**中間生成分布**と呼ぶ．

3）再学習による全体構成：中間生成分布は似たトピックがまとめられていないので全体構造はまだ見えない．そこで，これに基づくサンプリングにより新たにデータを取り直し，これからトピック構造を再学習する．得られたモデルをグローバルトピック構造とする．

ひとたび，上記の枠組みでグローバルトピック構造ができあがると，その各トピックと，各サイトのトピックがどう結びついているのかを，トピック間の距離を計算することにより，各サイトに共通なトピックや各サイト

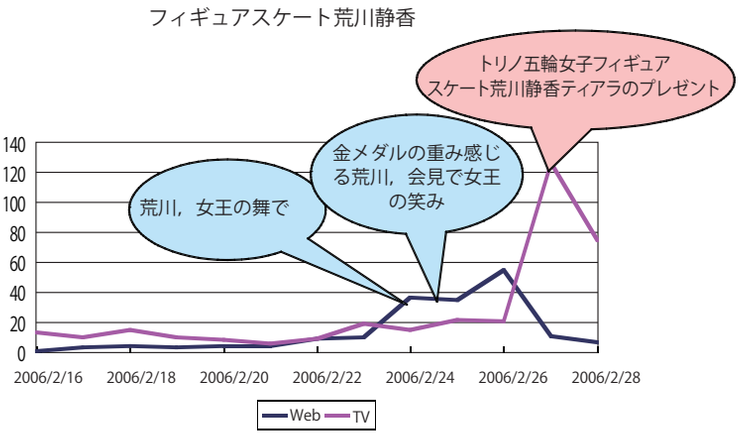

図 -6　TV データと CGM の時系列相関

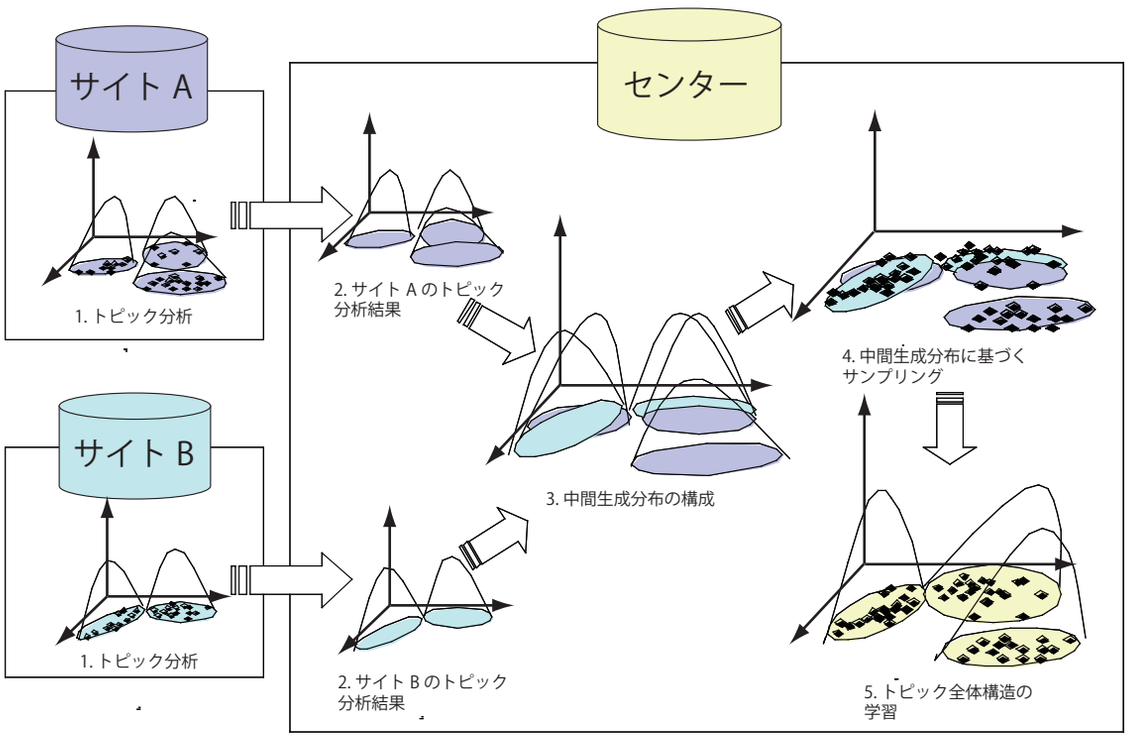

図 -7　分散協調トピック分析の流れ

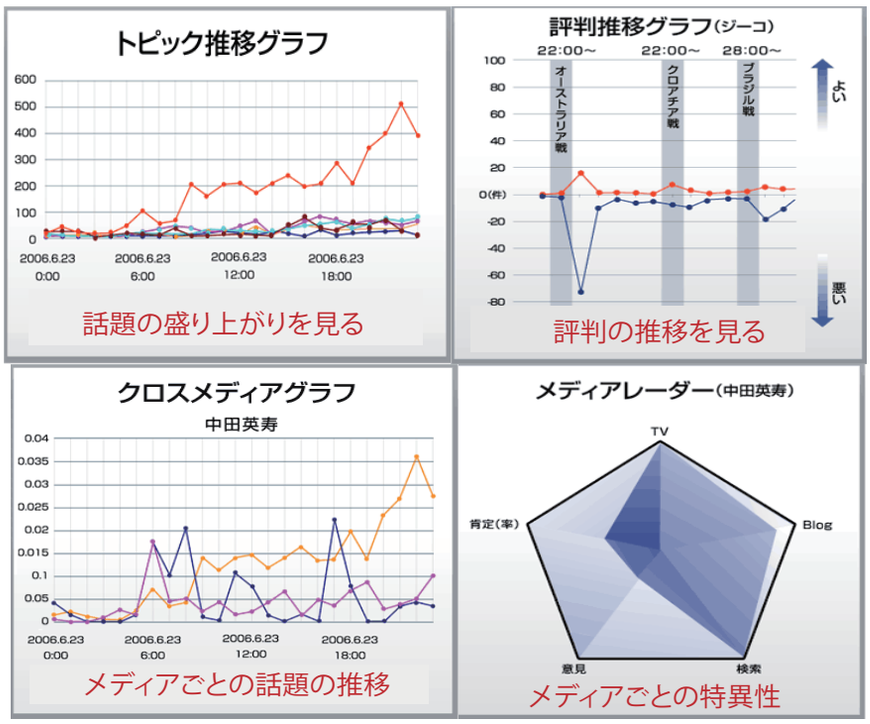

図 -8　2006 年サッカーワールドカップの CGM マイニング

特有のトピックを発見できる．

旬感ランキングにおける CGM 報道統合分析

CGM 情報とリアル情報を統合分析する試みとして，**BIGLOBE 旬感ランキングの「分析コーナー」**[8] を紹介する．ここでは，上記の動的トピック分析，文脈マイニング，分散協調トピック分析の枠組みを総合的に利用しながら，ブログ，検索，TV 書き起こしデータ等を掛け合わせて得られる世界を多面的に俯瞰することを目的としている．これまで，サッカーワールドカップ（2006 年 7 月），夏休み映画（8 月），ゲーム（9 月），温泉（10 月），秋ドラマ（11 月），次世代ゲーム機（12 月）を特集してきた．

サッカーワールドカップ特集の場合では，ブログ（データセクション（株），NEC BIGLOBE 提供）と，TV 書き起こしデータ（（株）プロジェクト提供）から評判情報抽出を行いつつ，**図 -8** に示すようなアウトプットを示した．**トピック推移グラフ**では，ブログと TV 書き起こしデータに共通のトピックとして，どんなトピックが活性化しているかを時系列的に示している．**評判推移グラフ**では，特定の選手の良い評判，悪い評判の多さの時間的推移を示している．クロスメディアグラフでは，特定の選手のブログ，TV，検索での露出度合いの時系列変化を示している．メディアレーダーでは，特定の選手の TV，ブログ，意見率，検索率，意見の多さを軸としたレーダーチャートを示している．

また，夏休み映画特集では，ブログに書き込みの肯定意見数（縦軸），TV・CM 放映時間（横軸），興行収入ランキングポイント（バブルの大きさ）を表にした（**図 -9**）．ここで，「バブルが上方向へ飛んでいくほど，ブログに肯定意見が多く書き込まれた映画」「右方向へ飛んでいくほど，宣伝などの目的で TV/CM で放映された時間が長かった映画」ということが分かる．

さらに秋ドラマ特集では各テレビドラマで話題になったことを CGM から文脈マイニングで抽出したところ，**図 -10** のような結果が得られた．これによって各ドラマで何がポイントになって受け入れられたのかが理解できる．

CGM 分析のこれから

いまや CGM を分析するサービスは巷にあふれている．その 1 つ 1 つは，本稿で紹介したような方法論を持ち出すまでもなく，頻出単語や，評判の良い悪いといった情報を取り出すだけで成立しているのがほとんどである．それだけでも貴重な情報が得られるのが実情である．しかし，データ量が膨大になって，CGM に潜むより深い世界を覗いてみたい，と思ったときに，本稿のような，「CGM マイニングと知識化」の手法の数々が近い将来，必ずや必要とされると考えられる．

本稿では，ネット内のコミュニティに関するリンク解析は対象にしなかった．しかしながら，ネット上のリン

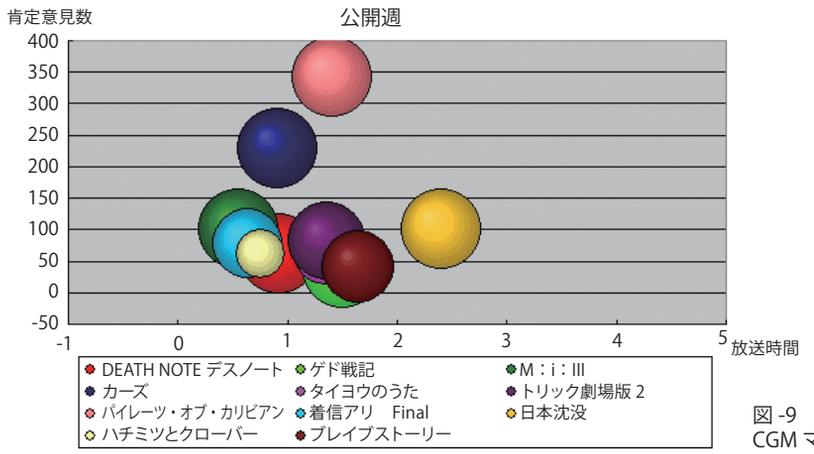

図 -9　2006 年夏休み映画の CGM マイニング

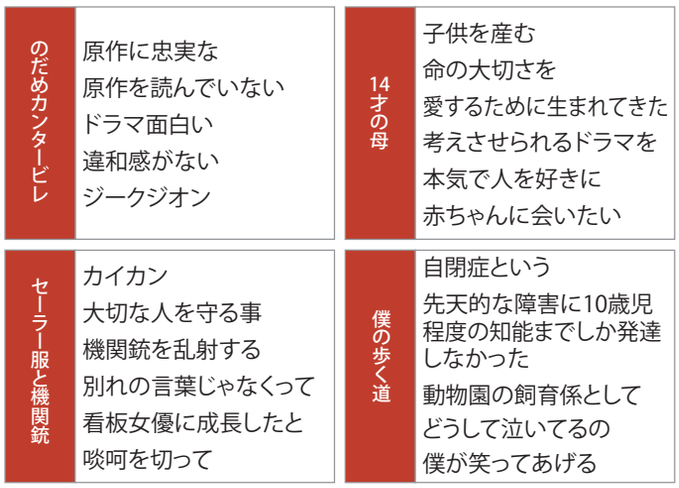

図 -10　2006 年秋のドラマの CGM マイニング

ク構造は CGM を把握する上で重要な情報である．たとえば，ブログのリンクやトラックバックの数はそのブログの信憑性にかかわってくるであろうし，情報の流れや信頼性を加味した分析をする上で将来活用されるものと期待できる．このように CGM 分析は，コンテンツのテキスト解析だけでなく，リンク解析とも融合しながら，大きく成長していくものと考えられる．

参考文献

1）Mei, Q. and Zhai, CX. : Discovering Evolutionary Theme Patterns from Text-An Exploration of Temporal Text Mining, Proceedings of the Eleventh ACM SIGKDD International Conference on Data Mining and Knowledge Discovery, ACM Press, pp.198-207 (2005).
2) Mergu, S. and Ghosh, J. : A Distributed Learning Frame-work for Heterogeneous Data Sources, Proceedings of the Eleventh ACM SIGKDD International Conference on Data Mining and Knowledge Discovery, ACM Press, pp.208-217 (2005).
3) Morinaga, S. and Yamanishi, K. : Tracking Dynamics of Topic Trends Using a Finite Mixture Model, Proceedings of the Tenth ACM SIGKDD International Conference on Data Mining and Knowledge Discovery, ACM Press, pp.811-816 (2004).
4) Morinaga, S., Arimura, H., Ikeda, T., Sakao, Y. and Akamine, S. : Key Semantics Extraction by Dependency Tree Mining, Proceedings of the Eleventh ACM SIGKDD International Conference on Data Mining and Knowledge Discovery, ACM Press, pp.666-671 (2005).
5) Yamanishi, K. and Li, H. : Mining Open Answers in Quessionare Data, IEEE Intelligent Systems, pp.58-63 (Sep./Oct. 2002).
6) Yamanishi, K. and Maruyama, Y. : Dynamic Model Selection with Its Applications to Novelty Detection, IEEE Transaction on Information Theory, Vol.53, Issue 6, pp.2180-2189 (June 2007).
7) 松村，森永，山西 : 分散・ヘテロなデータからのトピック全体構造の学習 , FIT2005.
8) http://search.biglobe.ne.jp/ranking/

（平成 19 年 7 月 12 日受付）

山西　健司　k-yamanishi@cw.jp.nec.com

1987 年東京大学大学院工学系研究科計数工学専攻修士課程修了．同年 NEC 入社．1992 年東京大学より博士号（工学）取得．1992 ～ 95 年 NEC リサーチインスティチュートに Visiting Scientist として出向．現在，NEC 中央研究所共通基盤ソフトウェア研究所兼ビジネスイノベーションセンター，主席研究員．情報論的学習理論，データマイニングの研究に従事．

森永　　聡　morinaga@cw.jp.nec.com

1994 年東京大学大学院工学系研究科計数工学専攻修士課程修了．同年 NEC 入社．1999 年東京大学より博士号（工学）取得．現在，NEC 中央研究所共通基盤ソフトウェア研究所主任研究員．テキストマイニングの研究に従事．

松村　憲和　n-matsumura@jz.jp.nec.com

2004 年奈良先端科学技術大学院大学情報科学研究科情報生命科学専攻修了．現在，NEC 中央研究所共通基盤ソフトウェア研究所在籍．専門はテキストマイニング，分散マイニング．