CGM データマイニング技術

森永 聡・山西 健司

要　旨

ユーザ参加型WEBサービスの発達により、ブログや掲示板上での口コミのようなCGMが実社会にも大きな影響力を持ちつつあります。本稿ではこれらCGM情報（サイバー空間）とTVや新聞などの報道情報（リアル空間）を統合分析するためのテキストマイニング技術をいくつか紹介し、それを利用した実際の分析サービスの実例として、NECデータマイニング技術センターが分析結果を提供している「BIGLOBE旬感ランキング」の分析コーナーを紹介します。

キーワード

●CGMデータ ●テキストマイニング ●動的トピック分析 ●分散協調トピック分析 ●文脈マイニング

1. まえがき

ユーザ参加型WEBサービスの発達により、ブログや掲示板上での口コミのようなCGM（Consumer Generated Media）が大きな影響力を持ちつつあります。TVや新聞などの従来メディアがCGMに影響を与えている一方で、逆にCGM発のトピックがメディアで取り上げられる例も出現しています。このような状況のもとで、サイバー空間の情報（CGM等）とリアル空間の情報（メディア情報等）を融合して、その相関性や移り変わりを把握し、世の中のトピックの全体像を俯瞰することが、マーケティング分析や新たなユーザ参加型WEBサービスの提供といった側面で、重要となってきています。

本稿では、まずCGMや報道情報を統合分析するためのテキストマイニング技術をいくつか紹介します。ここでは、1）ダイナミック（動的）でかつ、2）多様（ヘテロ）なデータに対して統合分析ができること、さらに、3）言われている中身を概観できるところが本質となります。これに応えるマイニング手法として、動的トピック分析、分散協調トピック分析、文脈マイニングを紹介します。

次に、上記技術を利用した実際のブログ・TV・検索を融合した統合分析サービスの実例として、NECデータマイニング技術センターが分析結果を提供している「BIGLOBE旬感ランキング」の「分析コーナー」の事例を紹介します。

2. 動的トピック分析

時間とともに定常的に流れ入るテキストストリームからトピック構造やその変化を捉えることは、ブログ分析・報道分析の基本です。ここでは、ダイナミックにテキスト系列をクラスタリングし、新しいトピックの出現を検出する動的トピック分析[1]の枠組みを紹介します。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としています。

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します。

1) トピック構造同定：どんなトピックがどういう割合で存在するかといった構造を発見

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をタイムリーに検出

3) トピック特徴抽出：各トピックに特徴的に出現する言葉を抽出

このようなタスクをデータに関して逐次的に行うために、以下のように問題を定式化しています(**図1**)。

(1) モデル：まず、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、トピックの確率的出現構造を有限混合モデルを用いてモデリングします。ここで、有限混合モデルを構成する各成分（正規分布、またはバイナリ分布）は1つのトピックを表し、混合比はトピックの出現確率分布を表します。

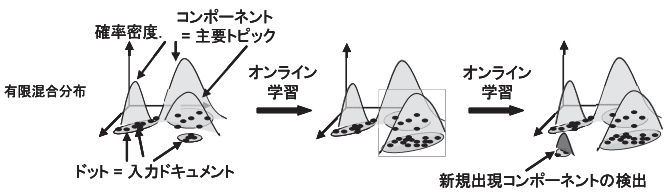

図1　動的トピック分析

(2) 学習：上記トピック構造を、タイムスタンプ付きの忘却型EMアルゴリズムによってオンライン学習し、トピック構造を同定します。ここで、忘却型EM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、非定常データからも学習できるようにしたEMアルゴリズムです（⇒タスク1）。

(3) 最適トピック数選択：時間と共に推移するトピックの最適な数を動的モデル選択によって求めます。ここで動的モデル選択とは、時間とともに変化する有限混合モデルの混合数の最適数をダイナミックに求める機能です。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ます（⇒タスク2）。

(4) トピックの特徴分析：　混合成分に対応するテキストを互いに比較し、情報尺度ESC（Extended Sto-chastic Complexity）に基づくランキングにより各トピックの特徴語を抽出します[2]（⇒タスク3）。

上記枠組みに従って、次々と追加されてくるCGMや報道のデータに対して、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ます。

3. 分散協調トピック分析

複数のリモートサイトに分散蓄積されているテキストデータに含まれる情報を統合して、全体を俯瞰するトピック構造を抽出する問題を考えます。これは情報統合の基本問題です。ここでは特に、1) ヘテロ性を持つデータに対して、2) 生データを1ヵ所に集めることなく（プライバシーの保持）、3) できるだけ少ない通信量で、4) 生データを1ヵ所に集めた場合と同等の精度を実現すること、を目標とします。この問題を解決するための分散協調トピック分析[3]の枠組みを以下に示します(**図2**)。

1) 各サイトでのトピック分析：第2章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ります。

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混合モデルを作ります。これを中間生成分布と呼んでいます。

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ません。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します。得られたモデルを全体トピック構造と呼ぶことにします。

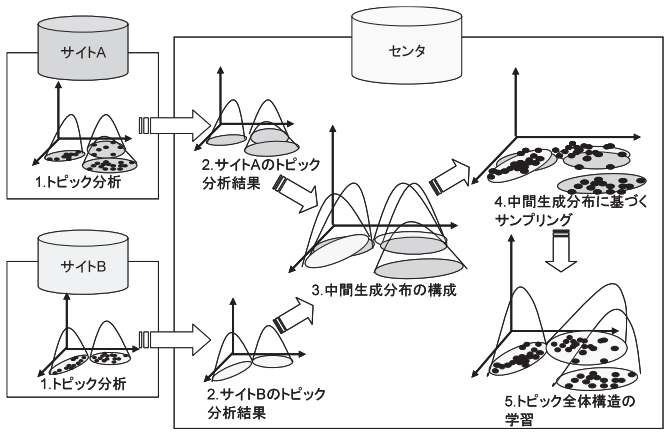

図2　分散協調トピック分析

ひとたび、上記の枠組みで全体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の関係を明らかにすることにより、各サイトに共通なトピックや各サイト特有のトピックが発見できます。これを利用して、CGMにおけるトピックと報道におけるトピックの共通性や差異性を把握するわけです。

4. 文脈マイニング

上記の各技術は、いうなれば「大量のテキストデータを同じトピックに属するものどうしにグルーピング分けする」機能を提供するものです。それに対し、文脈マイニングは、各グループの中で言われている特徴的な内容を抽出・概観する技術です[7]。

文脈マイニングでは、1) テキスト中の文章それぞれに対して、文節間の主述・修飾関係といった構文構造の解析を行い、2) 特定のテキストのグループに偏って多く出現する構文構造を抽出した上で、3) 抽出された構文構造に対応する日本語表現を生成して出力する、といった処理が行われます（**図3**）。

これにより出力される表現は「そこでよく書かれていることは何か」ということを人間が理解できる形で表すものであり、特定のトピックについて書かれているCGMや報道の内容

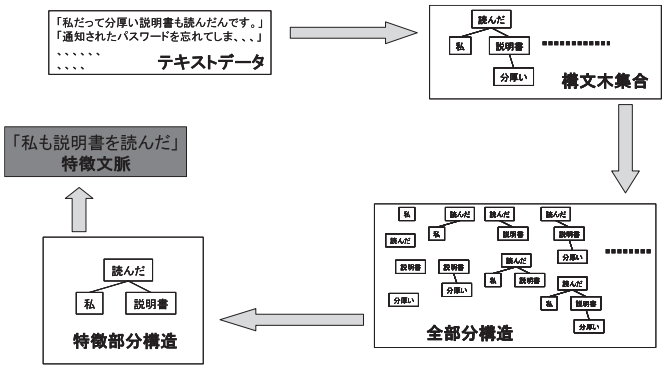

図3　文脈マイニング

を把握することなどに用いられるわけです。

5. 旬感ランキングにおけるCGM報道統合分析

サイバー情報とリアル情報を統合分析する試みとして、NECデータマイニング技術センターが提供する分析結果に基づいてコンテンツ作成している「BIGLOBE旬感ランキング」の「分析コーナー」[4]を紹介します（「分析ブログ」[5]も参照）。ここでは、上記の動的トピック分析や分散協調トピック分析、文脈マイニングの枠組みを利用しながら、ブログ、検索、TV書き起こしデータなど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ます。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）、ドラマ（11月）、次世代ゲーム機（12月）を特集してきました。

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ビッグローブ提供)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[6]、**図4**に示すようなアウトプットを示しました。

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ます。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ます。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ます。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ます。

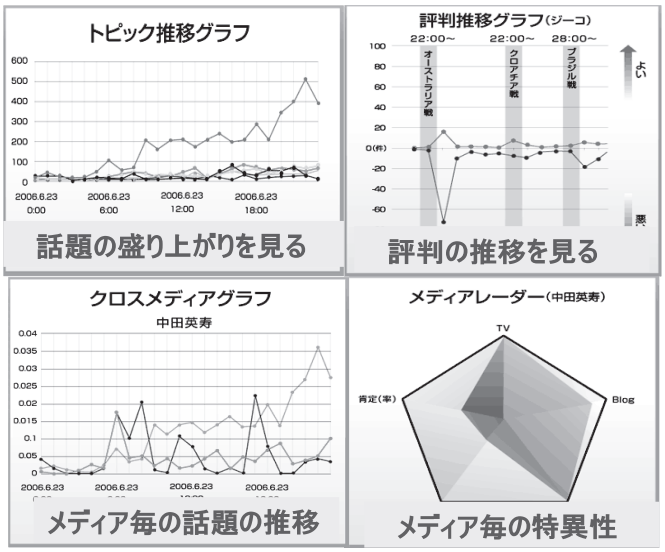

図4　旬感ランキング分析例

また、夏休み映画特集では、ブログへの書き込みにおける好意的意見数（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）をグラフ化し、公開開始から日が経つにつれて、どのようにグラフが変化していくかをアニメーション化しました(**図5**)。こうすると「ブログに好意的意見が多く書き込まれていった映画は、バブルが上方向へ飛んでいく」「右方向へ飛んでいくものは、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」ということになります。

ドラマ特集では、特定のドラマに対して各放送回に対応す

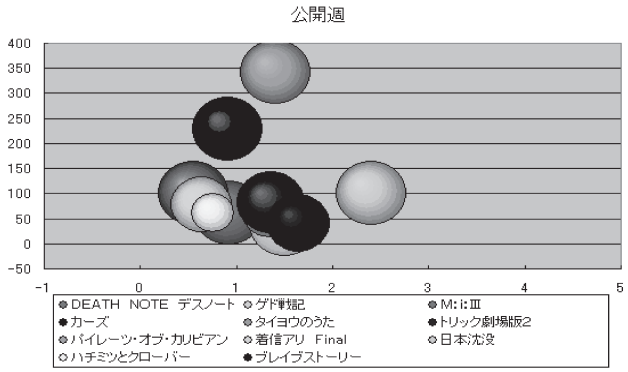

図5　バブルチャートの例

第1回 2006/10/16放送
変態ピアニストvs俺様指揮者のラプソディ

1位	原作のイメージどおり
2位	今日から始まる
3位	上野樹里
4位	竹中直人
5位	CG
6位	レクイエム
7位	胴体着陸
8位	玉木宏
9位	悲愴
10位	たまごっち

第2回 2006/10/23放送
落ちこぼれ変態オケ！波乱のスタート！？

1位	試験
2位	春
3位	ピアノの伴奏
4位	泡を吹いて倒れる
5位	いたずらが愉快
6位	バイオリンソナタ
7位	転科する
8位	変態オケ
9位	峰くん
10位	へたくそ

図6　特徴的なブログ書き込みの例

るブログ上での特徴的な書き込み抽出していきました（**図6**）。テレビでの放送内容のなかで、どの点がブロガーたちに注目されて書き込みにつながっているのかについて見て取ることができます。

ここで示したのは分析結果のほんの一例ですので、詳しくは当該ページおよびブログをご覧いただければと思います。

6. おわりに

本稿では、CGMの分析および報道との融合分析のテキストマイニング技術による実現例を紹介しました。動的トピック分析や分散協調トピック分析、文脈マイニングといった技術を組み合わせることで、サイバー空間での話題の盛り上がりや、リアル空間との関連性の分析といった俯瞰的な視点を手に入れることができました。

今後、ますますユーザ参加型WEBサービスは発展していくと考えられ、その状況を把握することができるこれらの技術は、重要になっていくものと考えられます。

*本稿に記載している会社名、製品名は、各社の商標または登録商標です。

参考文献

1) S. Morinaga and K. Yamanishi: “Tracking Dynamics of Topic Trends Using a Finite Mixture Model,” in Proc. of KDD2004, ACM Press, 2004.
2) Yamanishi and H. Li: “Mining Open Answers in Quessionare Da - ta,” IEEE Intelligent Systems September/October, 2002.
3) 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).
4) http://search.biglobe.ne.jp/ranking/
5) http://mining.at.webry.info/
6) S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima: “Mining Product Reputations on the Web,” in Proc. of KDD2002, ACM Press, 2002.
7) S. Morinaga, H. Arimura, T. Ikeda, Y. Sakao, and S. Akamine: “Kes Semantics Extraction by Dependency Tree Mining,” in Proc. of KDD2005, ACM Press, 2005.

執筆者プロフィール

森永 聡
共通基盤ソフトウェア研究所
主任研究員

山西 健司
共通基盤ソフトウェア研究所
兼 データマイニング技術センター
主席研究員