第 2 回ソーシャルコンピューティングシンポジウム(SoC2011)
June 18, 2011
Aoyama Gakuin ASTUDIO, Tokyo

多メディア Web 情報からの社会分析

豊田正史
東京大学　生産技術研究所
Masashi Toyoda
Institute of Industrial Science,
University of Tokyo
mtoyoda@acm.org

Web 情報は，画像・映像等への多メディア化が急速に進むと同時に，ユーザによるリアルタイム発信へのシフトなどメディアとしての性質も変化を続けており，放送映像等の実世界情報と相互に及ぼし合う影響も拡大している．今後の社会活動を分析するためには，これら多メディア Web 情報の大規模な観測・解析が不可欠である．我々は，国立情報学研究所，早稲田大学との共同研究において，多メディア Web 解析基盤の構築及び社会分析ソフトウェアの開発を行っており，本講演では，本プロジェクトにおける我々の取り組みについて紹介する．

1.　はじめに

爆発的に増大を続ける Web 情報は，写真や動画の共有サイトの普及に伴い画像・映像等への多メディア化が急速に進んでいる．同時に，ユーザ発信の形態も，ブログから，ソーシャルネットワーク，マイクロブログ等，リアルタイムユーザ発信へのシフトが見られ，メディアとしての性質も変化を続けている．先の東関東大震災の折には，マイクロブログ上において避難場所等の有用な情報の共有，ボランティア，募金の呼びかけなどの社会活動が行われ，新しいメディアが重要な役割を果たしたことは，記憶に新しい．社会分析，マーケティング，リスク管理等を目的とした調査を可能とするためには，これら多メディア Web 情報の大規模な観測解析が不可欠である．

我々は，国立情報学研究所，早稲田大学との共同研究において，多メディア Web 解析基盤の構築及び社会分析ソフトウェアの開発を行っている．東京大学においては，Web 情報収集・蓄積基盤，多メディア Web 解析の要素技術，及び様々な社会分析ソフトウェアの開発を行っており，本講演ではこれらの取り組みを紹介する．

2.　多メディア Web 収集・蓄積基盤

Web 上のテキスト・画像・動画を含む多メディア情報を、データの更新頻度等に適応して効率よく時系列的に収集し、収集結果を随時検索可能な形式で蓄積する、多メディア対応時系列収集スケジューリング手法、及び高効率データ蓄積手法の構築を行っている．

本手法を用いて，日本語の Web ページ及び画像の大規模収集を継続的に進めており，これまでに収集した多メディア Web 情報は，過去 12 年分，累積約 190 億 URL に上る．本アーカイブは，アジア域において最大級の規模である．こうした大規模収集と同時に，ブログや Twitter に関しては，さらに時間細粒度な収集を行っており，Web 上の様々なメディアに関する分析が行える基盤を整えている．

3.　Web 上の話題構造可視化・探索システム

Web 上の話題の解析においては，話題に関連する文書数のピークの出現と，その時点でのインフルエンサーとを同時に把握することが重要である．これを可能とするために，2 次元，及び 3 次元可視化を用いた可視化・探索システムを提案している（図 1）．

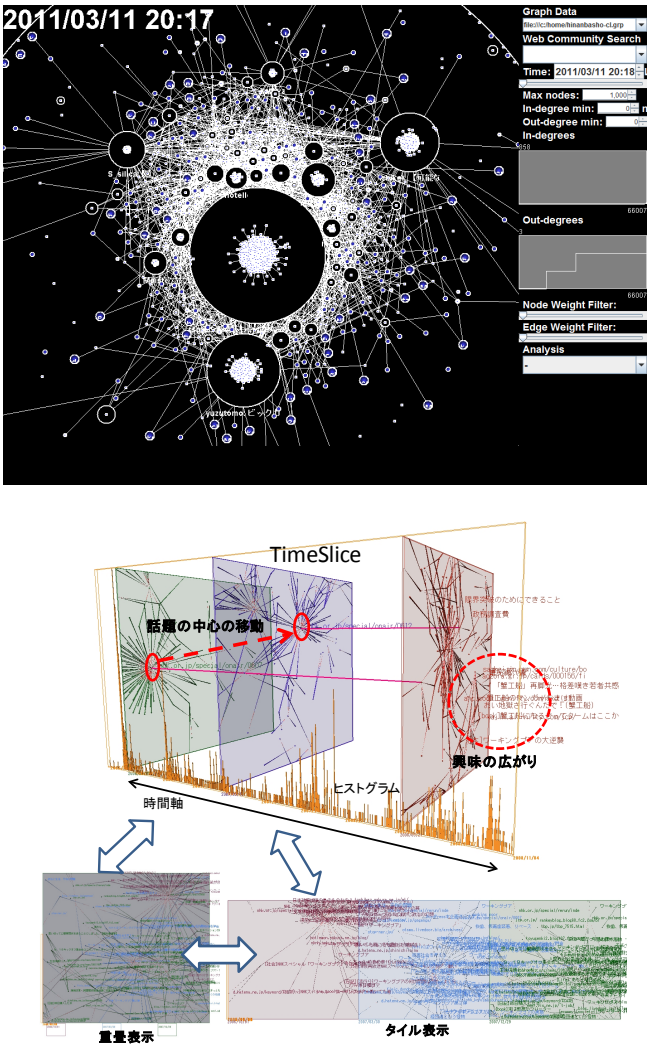

図 1. Web 上の話題構造の変遷可視化・探索システム
Figure 2. A visualization system for navigating the evolution of topics on the Web.

本システムは Web アーカイブから抽出された特定の話題に関するリンク構造を表す Web グラフを可視化する．2 次元可視化（図 1 上）では，大規模な話題に関して話題のクラスタの時間変遷をアニメーション表示できる．3 次元可視化（図 1 下）においては，ある時間の Web グラフのスナップショットを表示するパネルを TimeSlice[1] と呼び，ユーザは TimeSlice をマウスでドラッグすることでグラフの変化をアニメーションさせながら任意の時間におけるグラフを閲覧することができる．TimeSlice の側面には文書数を表すヒストグラムが表示されており，急激な増加など特徴的な変化が起きた時点でのグラフを容易に表示することが可能であ

る．また，異なる時間の Web グラフを比較するため，新たな TimeSlice を自由に追加することが出来る．これにより，Web 上の話題に関する話題の中心、興味の広がり等の時系列変化を把握可能とした．さらに、より詳細な比較を可能にするタイル表示および重畳表示手法を 3 次元空間で統合し，これらをシームレスに切り替え可能にした．ユーザは変化の全体像を俯瞰しながら，より局所的な変化の詳細を観測することが可能である．

4. ユーザの行動・興味に関する時間推移の可視化・探索システム

ブログ等の CGM の普及にともない，ユーザは自身の興味，行動，主観的意見を即座にかつ簡単にウェブ上に反映することが可能になってきている．これら，時間・社会状況とともに変化するユーザの生の声は，製品，人物，政策等の評判調査など，社会分析の観点から重要なデータとなってきている．

Web 上におけるブログユーザの行動・興味に関する記述をイベント（例: 新型インフルエンザが流行する）として抽出し、その時間変化を把握可能にする 3 次元可視化・探索システムを提案した[2].

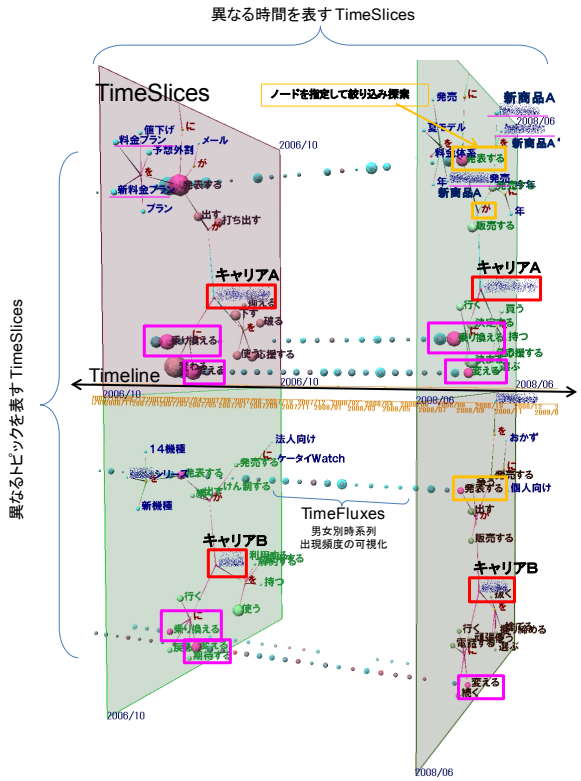

図 3. ユーザの行動・興味に関する時間推移の可視化・探索システム

Figure 3.A visualization system for navigating the evolution of users' activities and interests.

本システムは，入力キーワードを中心として，それに関する特定年月のイベント群（対象と行動の関係）をツリー表現で可視化する(図 3)．複数ツリーの並列可視化によりキーワードの比較が行え，アニメーションおよび異なる年月を表す複数ツリーの同時可視化によるイベント群の時系列構造変化も調査可能である．さらに，各イベントの時系列頻度変化の可視化が可能である．

5. CGM 画像の組織化手法

社会事象を解析する際には、話題がどのメディアから始まったかを同定し，どのようにその話題が多メディアの間に広がっていったかを分析する必要がある．例えば、Web 上で尖閣諸島に関する動画が公開された結果，それがテレビのニュースに現れるといった伝搬が考えられる．Web 上の話題がどのメディアから始まったかを同定し，その話題が多メディアの間にどのように拡散したかを解析することを目的とし，ブログ等の CGM 上における多数の画像を詳細な話題に分類しラベル付けすることにより組織化する手法を実現した．本手法は，まずユーザから与えられた検索語を入力としてブログ記事アーカイブを検索し，検索語を含むブログの記事集合を取得する．次に，記事中に貼られている画像を取得する．抽出した画像間の類似度を，画像特徴量，周辺テキスト，時間差の 3 種類の特徴を用いて算出し，階層クラスタリング手法を用いて分類を行う．画像のクラスタは，含まれる画像の重要度及びクラスタ内画像の類似度を用いてランキングされ，図に示すようにクラスタごとに時系列上に可視化される．本手法により以下のような分析が可能となる．

・ニュースがメディアにおいてどのような画像で扱われていたかを調査する
・イベントや集会などの模様や賑わいを視覚的に把握する
・商品画像の変化から、人気の度合いやデザインの変化を把握する

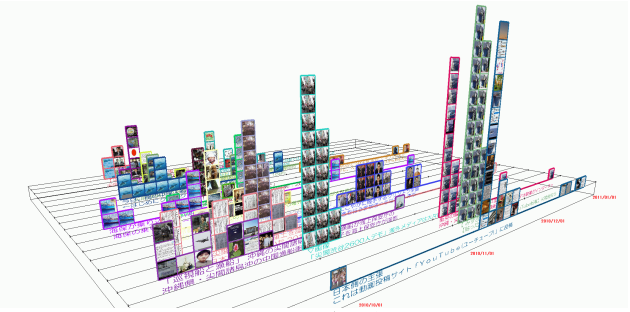

図 4. CGM 画像の組織化

Figure 4. A snapshot of organized CGM images

[文献]

[1] Masahiko Itoh, Masashi Toyoda, and Masaru Kitsuregawa: "An Interactive Visualization Framework for Time-series of Web graphs in a 3D Environment," The 14th International Conference on Information Visualization (IV2010), 2010.

[2] 伊藤正彦、吉永直樹、豊田正史、喜連川優,ブログユーザの行動・興味に関する時系列推移 3 次元可視化システム，第 3 回データ工学と情報マネジメントに関するフォーラム(DEIM2011)，2011.

豊田 正史 Masashi TOYODA

東京大学生産技術研究所戦略情報融合国際研究センター准教授．1999 東京工業大学情報理工学研究科博士後期課程修了，博士（理学）．ウェブマイニング，ユーザインタフェース，ビジュアルプログラミングの研究に従事.