9 Web空間からの実世界情報の発掘

風間 一洋[1] 栗原 聡[2]
1 日本電信電話（株）NTT未来ねっと研究所 2 大阪大学産業科学研究所

実世界情報とは

インターネットは，当初大学や企業間の電子メールや電子ニュースによる情報交換が中心であったが，World Wide Webの登場により世界的な規模の情報の公開や連携が可能になると同時に，一般ユーザにも急速・広範に普及した．さらにURIで表現されるWeb上のリソースは，プログラムで動的に情報を生成することもできることから，新聞記事の閲覧やデータベース検索，オンラインショッピングなど，多くのサービスが実現され，今まで人間が実世界で行っていたさまざまな行動がWeb上でも可能になった．

Webでは実世界での情報収集やショッピングと比べて利用時に移動する必要がないこと，多くの選択肢から最良のものを簡単に選択できること，広告で利益を得るために無料で利用できるサービスが多いことなどから，現在では日常生活のかなりの割合がネットワークの利用に費やされるようになり，さらに人間の行動において実社会とインターネット社会が互いに密接に関係して分離できないようになってきた．このことから，Webを単なる情報やその利用手段としてでなく，実世界が写像されている仮想世界であると見なすことができるようになった．特に，ブログやTwitterのような情報の更新や伝播に実時間性のあるソーシャルメディアの登場は，相互の連携速度を飛躍的に向上させることとなった．

この結果，人間の行動の履歴や結果が，インターネット上のさまざまな場所に，情報の内容，関係，およびユーザの利用履歴として記録され，これの利用も一種のセンシングと見なせる．実世界のセンシングと異なる点は，多数のセンサを広い範囲に設置しなくても，最初からディジタル化されている膨大なデータにネットワーク経由でアクセスして容易に収集・分析できることである．

そこで，Webから得られたデータを分析して得られる実世界情報に関する研究が盛んになりつつある．この実世界情報とは，たとえば実世界のエンティティ（人，組織，商品など），地理情報（位置，建物など），関係情報（エンティティや地理情報の関係），評判，情報伝播（口コミなど）などである．本稿ではインターネットを単なるセンサデータの転送経路として使うのではなく，実世界で発生したイベントや相互作用の影響を，インターネットから推測・推測し，そしてその結果を利用する研究を紹介し，今後の展開について考察する．

人間関係の抽出と利用

Web情報検索の問題点

Web空間は，その情報量の膨大さゆえに情報の過負荷（information overload）が発生し，目的とする情報が存在するにもかかわらず，うまく探し出せないことが多い．これをハイパー空間の迷子問題と呼ぶ．

Web上の情報を探すための一番有用な方法はサーチエンジンである．しかし，Webが誕生してしばらくは「サーチエンジンは使い物にならない」と言われてい

た．この理由は，新聞記事などと違って用語や文体が統制されておらず，誤字脱字や間違いも多いWeb情報に対しては，従来の純粋な全文検索技術だけでは，目的の情報を見つけ出すのが非常に難しかったからである．

この問題を解決するために，Webの社会性と膨大な履歴が活用された．たとえば，サーチエンジンODINでは検索履歴・閲覧履歴を用いた検索支援語の提示機能（1997年）とアンカーテキストとハイパーリンク構造を用いた検索とランキング，グループ化（1999年）により検索結果の質の向上や情報閲覧性を向上させた．ただし，これらの技術は特定の団体などのWebサイトを探すナビゲーショナルタスクの性能を飛躍的に向上させたが，ある情報を含むWebページを探すようなインフォメーショナルタスクではあまり有効ではなかった．

実世界指向情報探索

インフォメーショナルタスクを支援するための1つのアプローチは，実世界の実体とWeb空間の固有表現を結びつけることでWeb上の情報探索過程を支援する実世界指向情報探索である．

Web空間には，実在の人，団体，物を示す情報が数多く含まれるので，ユーザにとって既知な固有表現を抽出して提示すれば，実世界という視点からWeb空間を見ることができ，Webページの属性の判断やWeb空間内を探索する際のランドマークとして使用したり，目的とする情報に絞り込むことが容易になる．

さらに関連がある固有表現をエッジで繋いでネットワーク構造としてユーザに提示すれば，情報全体の概観を容易に把握できる．固有表現は一般的な検索支援語と異なり，Web空間における出現場所がより局所的であり，意味的に類似する情報群であっても，実社会で区別されている情報は別のクラスタとして表示される．

人名

筆者らは実世界指向情報探索としてNEXAS（Named Entity eXtraction and Association Search）を提案した[1]．これで固有表現として人名を使用した理由は，日本語形態素解析を使えば容易に人名を抽出できること，Web空間の出現確率が比較的高いこと，人間関係としても扱えるからである．

実際には，まず指定された検索語でWebページを全文検索・順位付けし，事前に作成した索引から上位のWebページに出現する人名集合を求める．さらに，出現するサーバ数が多い人名を抜き出して，Webページ内の出現位置が近い人名同士を関係づけて，人間関係を求める．出現サーバ数を用いるのは，特定のWebサーバにだけ大量に表れる人名は著名人とは考えにくいからであり，Webページ内の距離を用いるのは，長い文章中の離れた位置に出現する人名はあまり関係がない可能性が高いと考えられるからである．

このように処理することにより，検索結果に出現する人名集合，そしてその人間関係を求めて，検索結果と一緒に表示することができる．

人間関係を利用した情報探索

人間関係を単に表示するだけでなく，それを探索経路として情報を巡回できるようになれば，効率的な情報探索が可能になると考えられる．そこで，人名集合と人間関係の表示に加えて，それらを用いた情報探索が可能になるコミュニティナビゲータを作成した[2]．

最上部の入力フィールドに「XML」という検索語を入力して検索すると，図-1の検索GUI部では，人名を含むWebページ群，それらのWebサーバ群，検索結果に含まれる人名，指定した人名と共起する人名群を表示し，図-2の人間関係可視化部では，得られた人間関係をばねモデルで可視化する．これらは互いに連動しており，たとえば，「浅海智晴」をクリックすると，検索GUI部では，彼の名前が出現するWebサイト，Webページ，そして関係がある人名集合をリストの上部にハイライト表示し，人間関係表示部では，彼の名前が赤で，関係のある人名がピンク色で表示される．サーバ単位の表示機能を提供しているのは，Webサーバが現実の何らかの団体に対応して

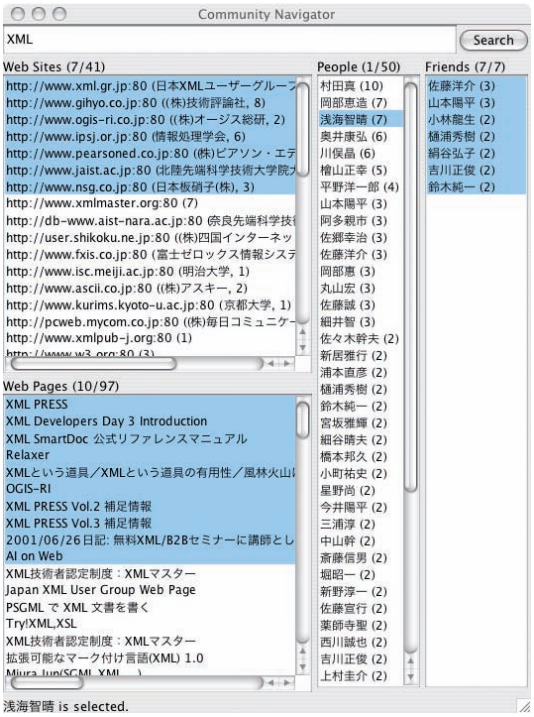


図 -1　人間関係の検索例

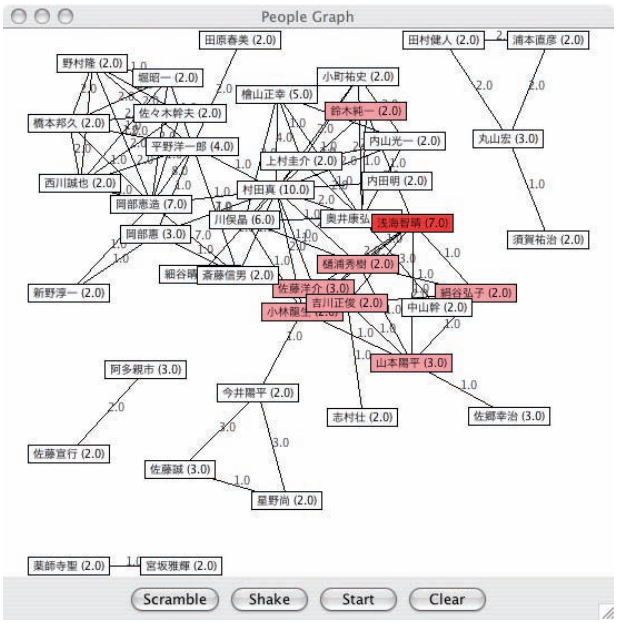


図 2　人間関係の可視化例

いることが多いからである．

このような人名と人間関係を用いた情報探索には，いくつかの利点がある．まず，人間という理解しやすい存在を Web 空間のランドマークとして使用できるので，ハイパー空間の迷子問題を回避することができる．たとえば，ある人間が開発者であるとか，国際化を専門としているというような事前知識を，人名選択や経路選択に役立てることができる．次に，人名リストの上位に表示される人名は多くの人間や情報に関係するキーパーソンであり，そこから開始すれば効率的に情報を探すことができる．さらに，人間関係ネットワークにはノードが密に関連しているコミュニティ構造が複数存在するが，これは現実の会議や団体などの集団に関係していて，ある特定の集団を詳しく調べたり，関連する集団を渡り歩くことができる．最後に，人間関係はスモールワールドと呼ばれるように，疎であってもノード間平均距離が小さいネットワーク構造を持つので，広大な情報空間を効率的に情報探索するための経路として有用である．

■ 重なりを許すコミュニティ抽出

人間関係で特に密な部分は，その人たちが一緒に活動していることを示し，それはコミュニティとして抽出できるはずである．しかし，図-2 では W3C の XML ワーキンググループで活躍した村田真は一番のキーパーソンとして複数のコミュニティを互いに結び付け，他に何人も複数のコミュニティに参加している．結局現実のコミュニティはこのような重なりを持つので，従来使われてきた弱い部分を切断するタイプのクラスタリングでは正しく分離できない．これを適切に処理するためには重なりを許容して密なコア部を抽出する必要がある．

この問題を解決するために，Palla らの k-クリークコミュニティ法では，サイズ k の完全部分グラフ（k-クリーク）で構成されるコミュニティを抽出して，重なりを持つコミュニティを抽出する[3]．それに対しての SR 法と SR-2 法はスペクトラルグラフ分析の一手法で，ネットワークの隣接行列の固有ベクトルに基づいてノードをランキングしてノード集合を求めてから既抽出のコア部のエッジの削除を行う処理を再帰的に繰り返し，複数のノード集合を抽出する[4]．SR 法

とSR-2法の違いは固有ベクトルを量子化する際の解の探索時に前者は最大値を，後者は最初のピークを用いる点である．

図-2の人間関係ネットワークに対して，国内の最大のXML開発者会議であるXML開発者の日，JISなどのXML関連の標準化，XML関係の業界団体であるXML Publishing Forum，情報処理学会，XMLの資格制度であるXMLマスターという5つの集団と，抽出されたコミュニティの対応を調べると，SR法では1つのコミュニティとして，SR-2法では別々のコミュニティとして抽出された．k-クリークコミュニティ法では，SR-2法よりさらにコア部しか抽出できず，リンク数が極端に多いノードが存在する巨大なネットワークは処理できなかった．この結果はWeb空間から複雑に関係したコミュニティを抽出・分離できることを示すだけでなく，SR法とSR-2法のように解の探索手法を変えることで互いに協調している専門家の集団や，さらに細分化した現実の活動をしている集団など，目的に応じたコミュニティ抽出の可能性を示す．

地理情報の抽出と利用

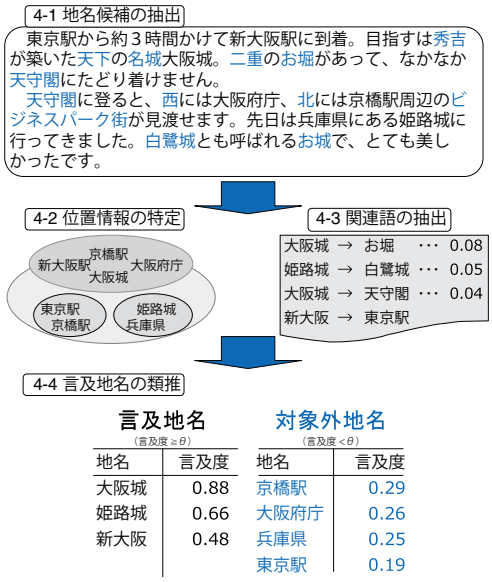
ブログやSNS，Twitterなどは，その瞬間に見たことや感じたことを気軽にWeb上に公開して友だちと共有するなど，まさに世の中の関心をリアルタイムに反映する新たなメディアであることから，これらのメディアからの有用な実世界情報の自動抽出を目的とするさまざまな研究も盛んである．ただし，これらのメディアにおいて，人々が書く文章が常に正しい文法に基づいているとは言えず，口語調での記述や，お互いにとって暗黙知的な地名であれば明記されないこともめずらしくはない．現在の検索システムにおいても地理情報は効果的に活用できているとは言えない．たとえば『清水寺』に関する情報を検索したい場合，ユーザは『清水寺』をクエリとして検索を行う．しかし，出力されるヒット件数は膨大であり，『清水寺』が文中に出現しているものの，実際には清水寺に関して言及していないブログ記事も多く抽出される．結局，ユーザが膨大な検索結果の中から個々に目を通して内容を確認する方法でしか欲しい情報を得ることはできず，実際は一部に目を通すのが実情で偏った見解に陥ってしまう可能性もある．

この状況に対して，Webから実世界情報の自動抽出への試みが行われているが，あらかじめ登録されている地理情報を利用する方法や，文章に付加されたメタ情報を用いる手法が多い．日本語文章に対して利用できるサービスであるLocoSticker1[☆1]もその1つで，こちらは位置情報（緯度経度情報）のみを用いて言及地名を類推する．こちらもメタ情報を用いることで正確な地名と文章のマッチングが可能となるが，人手によりメタ情報を付与することはコストが高く，適切なメタ情報が付与されたブログ記事は少ないのが現状である．

地名の曖昧性の除去のためには，地名階層の上下関係（東京都と世田谷区の関係など）を用いる方法や．文章内の地名と地名間の関連性，そしてリンク関係やユーザログなどを用いる手法も提案されているが，地名が詳細に記述されていない短い文章では利用できない．これについては，すべての文章は多かれ少なかれ地名と関係があるという仮定に基づき，あらかじめ地名に関連する単語に地名としてのメタ情報を付加する方法などが提案されているが，そのためにはメタ情報を付加するためのコストが発生してしまう[5]．

この状況を踏まえ，筆者らが現在提案している「文字情報解析と位置情報解析の両方を効果的に利用する手法」について簡単に説明する[6]．ここでは「人がある特定の地名に関する文章を作成する際には，その文章内で引き合いに出される他の地名よりも，その特定の地名に関する文の割合が高くなるはずである」，という仮定を前提とする．たとえば，京都での出来事に関して記述する際，南禅寺に行った場合は湯豆腐を食べたことや，琵琶湖疎水を見たこと，嵐山ならトロッコ電車を使ったことや渡月橋を渡ったことが記述されている可能性が高い，といったことが挙げられる．以下抽出の過程を簡単に説明する．

☆1 LocoSticker, http://locosticker.jp/



言及地名（言及度 ≥ θ）

地名	言及度
大阪城	0.88
姫路城	0.66
新大阪	0.48

対象外地名（言及度 < θ）

地名	言及度
京橋駅	0.29
大阪府庁	0.26
兵庫県	0.25
東京駅	0.19

図 -3　システム構成

■ 地名の抽出

まず，一般に利用できる地名辞書を用いて文章中から地名を抽出するだけでは不十分である．たとえば，「○○コンビニは混んでいる」と言うときの○○コンビニは単なる名詞であるが，「○○コンビニで待ち合わせる」の○○コンビニは明らかに地名である．そこで，自然言語処理の固有表現抽出技術を用いて地名候補の抽出を行う必要がある．そのため，人手により地名のラベル付けを施した文書（ブログ記事など）を用いて，地名が書かれる文章のパタンの学習を行い，地名抽出の精度を高める作業を行った．具体的にはCRF (Conditional Random Field)☆2による識別モデルを作成して文脈からの地名抽出を行った（図-3に処理の流れの概略を示す）．

■ 位置情報から主題地名の特定

上述するように，通常みかける実世界情報を含む文書においては，ある地名を主題とする文書中にも，その地名に関連する地名など複数の地名が書かれていたり，日本橋などの地名においては日本全国に同名の地名が多数存在する場合もある．よって，地名を

☆2 CRFは，系列ラベリング問題のために設計された識別モデルであり，正しい系列ラベリングを他の全ラベリング候補と弁別するような学習を行う．品詞付与，固有表現抽出，HTMLからの情報抽出，といった系列ラベリング問題に適用されている．

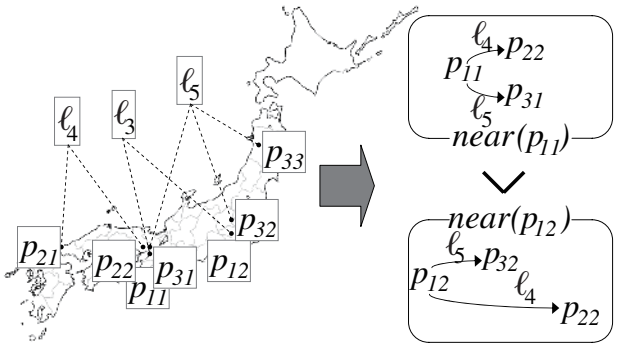


図 -4　クラスタリングの概念図

複数含む文書からこの文章における主題となる地名を同定する必要がある．この課題において，複数の地名を含む文章においては，主題となる地名に関連する地名が，他の地名に比べて多く出現することが想定されることから，文章に登場する複数の地名のそれぞれの緯度経度情報から，地理的な近接関係を用いて主題となる地名を推定できる可能性がある．

そこで，筆者らは，前処理にて抽出したそれぞれの地名liに対して位置情報p_{ij}（緯度経度）を求め，地名間の距離に基づくクラスタリングを行い，各クラスタに含まれる地名の数やクラスタの密度などから，各クラスタの中心となる地名に対して，その文書における主題である可能性を示す度合いを与える（図-4にクラスタリングのイメージを示す：p_{11}とp_{12}では，p_{11}の方が高い度合いとなる）．

■ 文字情報からの主題地名の特定

次に，地名と関連の高い単語は，多くの文章においてその地名が出現した場合に，他の語と比べて近い位置に出現する傾向があると考えられることから，文字情報からの各地名の主題度を，次のように定義した．

$$rel(l_i, w) = \frac{P(l_i|w)}{count(l_i)} \times \sum_{l_i,\ w\text{を含む文章}} \frac{1}{\min(dist(l_i, w))}$$

候補地名をl_i，単語をwそして，その関連度を，最も近い単語との距離の逆数の合計を抽出対象の全ブログ記事に対するl_iの出現数$count(l_i, w)$で割ったものとした．$dist(w_1, w_2)$はw_1, w_2の文単位での距離であり，同一文に出現すれば1，前後の文なら2, 3...

となる．また，関連語対象とする語wは，その出現数が全対象記事中にある一定回数以上出現しており，$dist\,(l_i, w) <= 3$であるものだけを対象とした．これは，語の出現数が極端に少ない場合，地名l_iとの同時発生数も少なくなってしまい，関連度に偏りが生じることが想定されるためである．

さらに，対象の地名とすべての語との距離の逆数の合計をしただけでは，「こと」,「人」,「お店」といったその地名を特徴付ける語でない一般的な語が高い関連度として抽出されてしまう可能性があることから，地名に対する対象単語の条件付き確率を導入することで，多くの地名と共起している一般的なキーワードの重みを下げ，相対的にその地名を特徴付ける語の関連度が向上するような定義となっている．

文字情報解析結果と地理情報解析結果の統合

最終的にブログ筆者が文章中で地名l_iを言及している度合いを示す言及度を，位置情報のスコア$(cl(l_i))$と関連度$(rel_{sum\,(li)})$を用いて以下のように定義する．

$$score\,(l_i) = \frac{(1+\beta^2 \times cl(l_i)) \times rel_{sum(l_i)}}{\beta^2 \times cl(l_i) + rel_{sum(l_i)}}$$

各l_iに対して$cl(l_i) = max(c(p_{ij}))$であり，文章内の各地名同士の近接度を表す$rel_{sum\,(li)}$は，文章内での各地名l_iとその関連語との関連度の合計で，以下のように定義した．

$$rel_{sum(l_i)} = \sum_w rel(l_i,\ w) \times \frac{1}{dist\,(l_i, w)}$$

ここで重要なパラメータがβ(0以上の正の整数)であり，近接度に対して関連度を何倍重視するかを表すものである．そして，$score(l_i)$が閾値θ以上であれば文章中で言及されている地名と判断した．

そして，関連語を用いた言及地名類推の有効性について，被験者実験を通した検証を行ったところβ=5あたりにおいて最も的確に主題となる地名を推定できることが分かり，統合することにメリットがあることが検証された（図-5に抽出例を示す）．また，上述したLocoStickerのような位置情報に基づいて文章と関連性が低い地名を除去するサービスとの比較も行い，本手法がより正しく主題となる地名を出力できることも確認された．

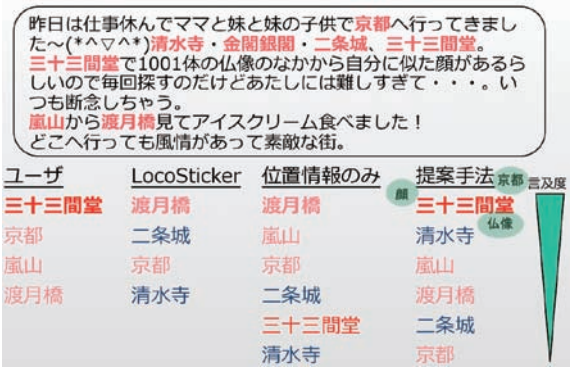


図-5　抽出例

情報伝播経路の抽出と利用

ソーシャルメディア

ブログ，Twitterなどのソーシャルメディアは，誰もが参加でき，社会的な相互作用を通じて発言をインターネット上に素早く広めることができることから急速に普及した．しかし，発言数が膨大でも，個々の発言の情報量が少なかったり，内容が自己完結的でなかったり，誤りやスパム，重複があることも多く，ソーシャルメディアを単なる情報源として考えた場合の有用な情報の割合は非常に少ない．

ソーシャルメディア全体を有効に活用する方法として，評判分析技術や属性推定技術がある．前者は発言に含まれる単語の出現頻度や発言のP/N判定☆3の結果を利用して，人気度や評判の良し悪しを解析する技術であり，後者は著者の年齢，性別，職業，居住地などを推定する技術である．これらは従来の自然言語処理技術の延長線上であり，マーケティング的な利点も大きいことから，盛んに研究や開発が行われている．

新しいアプローチとして，ソーシャルメディアを，実世界の出来事に対して即座に反応する動的なシステムと見なす手法がある．実社会で，たとえば，ワールドカップにおける日本代表の決勝進出のような注目度が高い出来事が起こった場合には，ソーシャルメディア上でそれに呼応するように大量の発言や，そ

☆3　文章が肯定的か否定的かを自動的に判定する技術．

の連鎖が発生する．そこで，発言状況から注目する出来事の発生を推定したり，その反応の速度や広がりの様子から，その出来事が世の中に与えたインパクトの度合いを調べることができる．従来の評判分析技術との違いは，個々の発言を独立に扱うのではなく，時間的変化や発言間の相互作用を考慮することである．つまり，評判分析や属性推定は発言自身を分析するのに対して，時間的変化や発言間の相互作用を考慮することが特徴である．

ブログの情報伝播ネットワーク

本節では筆者らの調べたブログ空間の情報伝播ネットワークの特性分析について述べる[7]．ブロガーは，実社会の何らかの出来事に対して他のブログ内部の発言や新聞やニュース，オフィシャルサイト，まとめサイトなどの情報源を引用してブログエントリを書く．その出来事の影響力が大きければ大きいほど，より広い範囲で意見が伝播し，複数の異なる意見の影響を同時に受けることで新しい意見が生まれることもある．また，情報の内容によって，伝播の様子にも違いが生じる．

ブログ空間における情報伝播経路のネットワーク構造を抽出する手法は，文章の引用やテキスト類似性に基づく方法とハイパーリンクに基づく方法に大きく分けられる．前者は明示的でない伝播も発見できるが精度が低くなり，後者は精度は高いが網羅性が低い特徴を持つ．たとえば，Kim らは引用の有無を高速に判定する方法を提案した[8]が，筆者らは抽出精度を重視してハイパーリンクに基づく方法を選択した．

まず，観測したい出来事を示す単語で検索・収集したブログエントリの本文からハイパーリンクを抽出し，ノイズを除去してから，異なるサーバからの被リンク数がある閾値以上のURLを情報源として，ハイパーリンクを逆方向にたどって情報伝播ネットワークを抽出した．

ただし，得られた情報伝播ネットワークの構造は均一ではなく，各情報源ごとに，情報源の被リンク数，情報同士の結合状況，情報源からの経路の長さなどの周辺の構造が異なる．そこで，情報源から到達可能な複数の部分ネットワークに分割してから，各部分ネットワークの情報伝播特性を分析する．

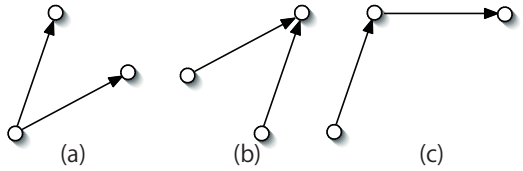


図 -6　情報伝播の基本構造

このようなネットワーク特性の分析には，行為者の関係性に着目して社会現象を捉える社会ネットワーク分析でも，さまざまな手法が提案されている．ただし，このような分野で扱うスケールフリーネットワークとは異なりクラスタ性が小さく，時間の経過につれて成長するために，時間的方向性を考慮した新たな手法が必要である．このネットワークは図 -6 に示す情報拡散 (a)，情報集約 (b)，情報転送 (c) を表す 2 本の有向エッジからなる 3 つの基本構造で構成される点に着目して，部分ネットワークに含まれる各構造の数を部分ネットワークのノード数で正規化して，情報拡散度，情報集約度および情報転送度を定義する．これらは部分ネットワークの情報伝播特性であるが，その部分ネットワークを生成するきっかけを作った情報源の特性としても用いることができる．

図 -7 と図 -8 に，それぞれ「毎日新聞」と「iPad」という単語で検索・収集したブログエントリから抽出した情報伝播経路の可視化結果を示す．前者は毎日新聞英語版サイトの低俗記事問題に関して虚偽の情報を発表していたという新しい事実が発覚した時期に，後者はアップルが iPad を発表した時期に収集した．Web ページとブログエントリをノードとして表示しているが，出次数に応じて直径を大きくしているので，有益な情報源ほど大きい．各ノードは，古いほど青く，新しいほど赤く着色している．情報源ノードが青い傾向があるのは，まず同時多発的に情報源が出現し，それを多くのブログで言及するという傾向があるからである．なお情報伝播のエッジの向きはハイパーリンクの向きと逆である．この図からは分からない

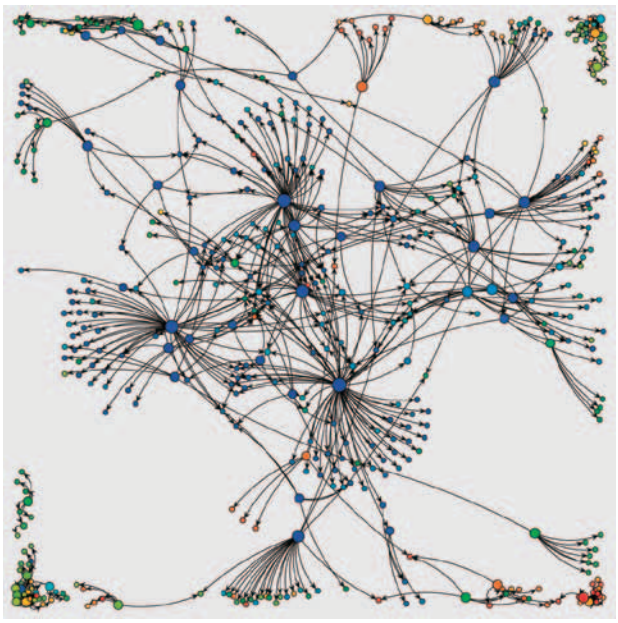


図-7 「毎日新聞」の情報伝播ネットワーク

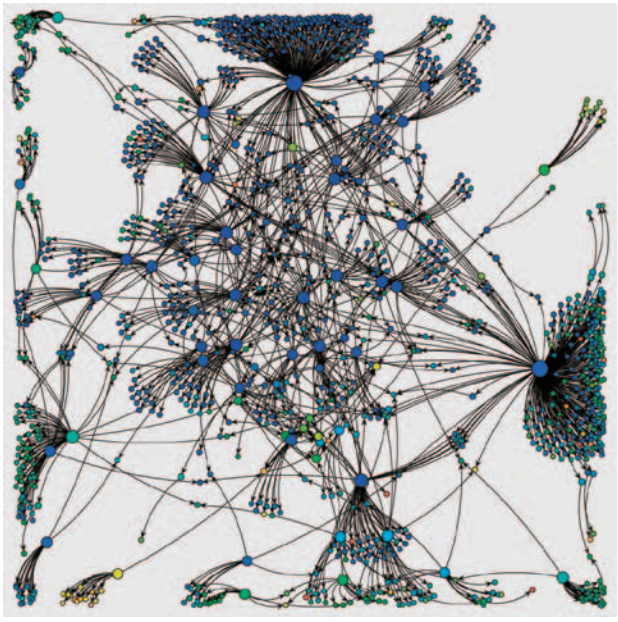


図-8 「iPad」の情報伝播ネットワーク

が，各ノードをクリックすると，そのノードから到達可能な部分ネットワークがハイライト表示される．

図-8で特に多くのエッジを持つ情報源は，日本とアメリカのiPadオフィシャルサイトであるが，図-7にはそのような顕著な情報源が見られない．この理由は，マスコミは同業の不祥事を報道しなかったが，ブログ・WikiなどのCGMはそれを補うかのように盛んに新事実の報道や事実のまとめ作業を行って，マスコミの代わりに情報を広める役を担ったからと推測され，ブログ空間における情報伝播がCGM主導型とマスコミ主導型などに分類できることが分かる．

このような情報伝播ネットワークに対して，オフィシャルサイト，ニュースサイト，CGMの3種類に分類した情報源の情報伝播特性を調べると，オフィシャルサイトは情報拡散度が，ニュースサイトは情報集約度が，CGMは情報転送度が高くなる傾向があった．

図-8のデータについて各情報伝播特性値でランキングした結果を調べると，情報拡散度の場合は，上位3位はiPadの日本語・英語とアップルの公式サイトであり，上位10件中5件がオフィシャルサイトであった．情報集約度の場合は，上位10件中6件がアップルがiPad発売をアナウンスしたというニュースであった．これらは比較的客観的な事実で短い記事である．情報転送度の場合は，1位がアップルがiPadを発表したイベントに関するレポートであり，上位10件のうち4件がアップルのイベントとiPadの使用レポート，2件がiPadの商標が他社に所有されていた事実に関する推測記事，2件が次期iPhoneとiPhone OSに関する推測記事であった．これらの内容は比較的長く，主観や推測に基づいている．

つまり情報伝播特性を分析すれば，注目されている権威がある情報，客観的で資料性の高い情報，主観的で議論を引き起こすような口コミで伝わりやすい情報などを発見することができると考えられる．この性質を情報源のランキングに応用すると，ユーザの情報探索の目的に応じて情報を提示することが可能になる．

Twitterの情報伝播ネットワーク

Twitterは発言の文字数が短いために，目的のトピックを示す単語で検索しても検索漏れが多く，ブログと同じ抽出方法は難しいと思われるので，あるユーザを中心に3ホップ先までのユーザの15日間のツイートを収集して，返信，非公式リツイート，ハイパーリンクを手がかりに情報伝播ネットワークを抽出すると，図-9に示すようなネットワークが多数得られる[9]．楕円で示すノードは各ユーザのツイートで

あり，中にユーザ名を表示している．実線の矢印は返信であり，破線の矢印は非公式リツイートである．四角のノードはURLであり，点線の矢印はハイパーリンクを表す．

このように得られた多数のネットワークを分析した．まずTwitterでは返信は27%と比較的多いが，非公式リツイートは8%と比較的少なく，ほとんど返信で構成されている．また，図-9のような大きなネットワークは少なく，大部分がサイズが小さいネットワークであり，しかも二者間の発言がかなり多い．

図-6に示した情報伝播の基本構造については，まずTwitterには複数の発言を同時に参照できない機能的な制約があるので，URLの参照を除けば情報集約構造は存在せず，ほとんど情報拡散構造と情報転送構造で構成される．また，ブログと比較すると情報転送構造が多い特徴がある．これはブログにおける二者間のやりとりのほとんどがコメント機能を用いるのでブログエントリを結びつけることがないのに対し，Twitterにおける同様の二者間のやりとりはツイートを結びつけるからである．これらの特徴は，ソーシャルメディアの設計と使われ方は，情報伝播特性の違いとして分析できることを示している．さらに，図-9のようなサイズが大きなネットワークは複数のユーザによる議論を示すが，特に有用な議論には，複数の人たちの興味を引いたことを示す多くの情報拡散構造が含まれるはずであり，自分が知らない有益な議論の発見の手がかりになると思われる．

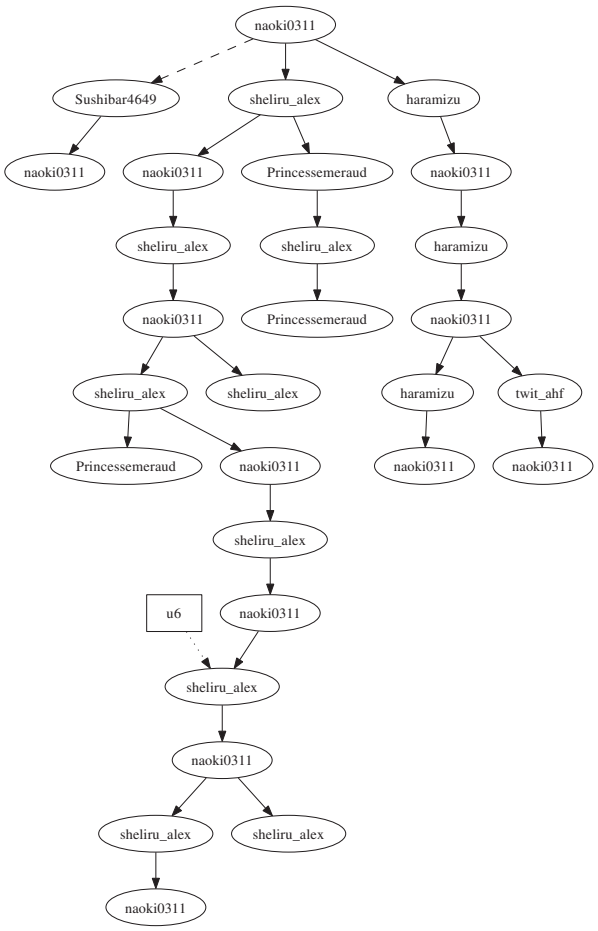


図-9　Twitterの情報伝播ネットワーク

■ 今後の展開

自然言語処理技術の発展により大規模データのタグ付けが容易になるとともに，センサを多数搭載した携帯端末が普及し，膨大な実世界情報が得られるようになるが，それゆえにさまざまな問題を抱えることになる．

通常は膨大な情報をそのまま処理できないので，多段の処理を行って複雑な解析が可能になる規模までデータ量を減らす．この場合には，Hadoopなどの大規模分散データ処理フレームワークが有用である．従来の並列分散システムと異なるのは，数値計算ではなくデータ処理指向で設計されていることであり，現時点では超大規模データを所持する組織が限られることと，維持管理コストが高いために研究者は手を出しにくい状況にあるが，ターンアラウンドタイムの短縮により試行が容易になることから，技術の成熟とノウハウの蓄積とともに将来的に普及すると思われる．

また，時間軸を含む超多次元データをうまく扱うためのデータマイニングなどの手法の研究も必要となる．

これらの実世界情報を利用するサービスに対するユーザの理解を得るためには，プライバシーの侵害が起こらないように注意しなければいけない．より詳細な分析ができるからと，ありとあらゆるデータを収集するような行為は，いざ深刻な問題が発生したときの社会の過剰な反発を招くので，この分野全体を

潰しかねない．名前のような識別子や，ユーザIDのような準識別子に対して匿名化を施した後でも，複数のデータの照合により一意に個人を特定できる場合がある．このような危険性を回避するために，プライバシー保護データマイニングの研究が行われている．たとえばk-匿名性などの手法が提案されているが，過度な処理が行われたり，計算コストが高いために，今後のさらなる進展が期待される．

また，本稿ではインターネットを単にセンサデータの転送経路として使う研究を除外したが，センサデータを共有するためのインフラストラクチャの開発や，公開するためのデータへのアクセス制御やプライバシーを考慮した画像などのデータの匿名化などの研究も進んでいる．

参考文献

1) Harada, M., Sato, S. and Kazama, K. : Finding Authoritative People from the Web, Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2004), Tucson, Arizona, USA, pp.306-313 (2004).
2) 風間一洋，佐藤進也，福田健介，村上健一郎，川上浩司，片井　修：Web空間における人間関係を用いた情報探索の一手法，情報処理学会論文誌：データベース，Vol.46, No.SIG 13 (TOD27), pp.26-39 (2005).
3) Palla, G., Derényi, I., Farkas, I. and Vicsek, T. : Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society, Nature, Vol.435, No.7043, pp.814-818 (2005).
4) 風間一洋，佐藤進也，斉藤和巳，木村昌弘：人間関係からの重なりを持つコミュニティ構造の抽出，コンピュータソフトウェア，Vol. 24, No. 1, pp. 81-90 (2007).
5) Wang, C., Wang, J., Xie, X. and Ma, W. : Mining Geographic Knowledge Using Location Aware Topic Model, Proceedings of the 4th ACM Workshop on Geographical Information Retrieval (GIR '07), New York, USA, pp.65-70 (2007).
6) 松本光弘，二宮亜佐美，長岡　諒，沼尾正行，栗原　聡：WWW上の文章に含まれるイベント情報からの地名の同定，第88回知識ベースシステム研究会 (SIG-KBS) (2010).
7) Kazama, K., Imada, M. and Kashiwagi, K. : Characteristics Estimation of Information Sources by Information Diffusion Analysis, Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence, Toronto, Canada (2010).
8) Kim, J. W., Candan, K. S. and Tatemura, J. : Efficient Overlap and Content Reuse Detection in Blogs and Online News Articles, Proceedings of the 18th International Conference on World Wide Web (WWW '09), pp.81-90 (2009).
9) 風間一洋，今田美幸，柏木啓一郎：Twitterの情報伝播ネットワークの分析，第24回人工知能学会全国大会 1F2-OS8-4 (2010).

(平成22年7月2日受付)

風間 一洋（正会員） kazama@ingrid.org
1988 年京都大学大学院工学研究科修士課程修了．同年 NTT 入社．2005 年京都大学大学院情報学研究科博士課程修了．博士（情報学）．現在，NTT 未来ねっと研究所主任研究員．

栗原 聡（正会員） kurihara@ist.osaka-u.ac.jp
大阪大学産業科学研究所知能システム科学研究部門准教授．慶應義塾大学大学院理工学研究科計算機科学専攻修士課程修了．NTT 基礎研究所を経て 2004 年より現職．マルチエージェント，ネットワーク科学等の研究に従事．翻訳『スモールワールド』(東京電機大学出版局) など．博士（工学）．JSAI, JSSST, IEICE, ESHIA 各会員．
http://www.ai.sanken.osaka-u.ac.jp/~kurihara/