RSS配信解析の必要性に関する検討

永江孝規
メディアアート表現学科

A Study on the Necessity for RSS Feed Analysis

NAGAE Takanori
Department of Media Art
(Received November 9, 2007 ; Accepted January 10, 2008)

1. はじめに

筆者らは1996年当時から、今日ではブログと呼ばれているWeb日記の更新監視システムに少なからず関わり合ってきた。この当時のことについては、できるだけ記憶の新しいうちにと1999年に紀要「インターネットにおける自発的コミュニティの形成、特にWeb日記に関して」[1]としてまとめた。発表後、思わぬ反響があり、多くの方のご声援とご批判をいただき，たびたびこの文書をメンテナンスすることになった[2]。2006年3月13日の後記に筆者は次のように書いている：

> 近頃のWeb2.0（wikiやブログやSNSなどを通じて、従来はメディアの受け手の側に沈黙していた一般大衆を取り込む形で巨大にふくれあがりつつあるWebというシステム、大衆を巻き込んで確率論的（創発的）に進化するシステム、誰かが品質保証をするわけではないが（勝手に品質保証宣言されたからといって当てになるわけではないが）ベストエフォートで長期的には役に立つシステム、とでも定義できるだろうか）の登場を見て改めて読み返し、手を加えてみたくなった。Web日記はブログとなり、日記をWebに公開しても誰も「奇異」に感じず、公開すればどうなるかということも120%わかった上で公開するようになっているので、当時の感覚というものが今ではわかりにくくなってしまっている。そのこと自体が今となってはかなり「新鮮な驚き」であり、この文書の「価値」といってもよいだろう。当時Web日記を書くことにあれほどくどくどと言い訳をしなくてはならなかったのがおかしくすらある。
>
> 同時に「日記の本質」を見誤らせてもいた。日記とは他の文章や文芸作品と本質的に何も違わないし、紙媒体であろうと粘土板であろうとWebという巨大で即時性のある電子媒体であろうと、本質的には何も違わない。違わないということが改めて確認されたことに意味がある。

この2006年春という時期は筆者らが「2ちゃんねる可視化計画」[27]および「半自動辞書編纂システムSADE」[31-41]を立ち上げ、筆者らのゼミが「Web開発」に本格的に取り組み始めた時期にあたる。1994年に初めてWebサーバを稼働させ、1996年からWeb日記界の発展と混乱に巻き込まれ、また2006年には、ある意味見切り発車的に、学術的価値があるかどうか見極めるつもりで匿名掲示板の研究を始めた。今日でもそうだが、当時、匿名掲示板を学術的研究対象としようという動きはほとんどなかった。これほど今日の日本社会に大きな影響を与えるに至ったメディアに対して、その学術的取り組みは驚くほど少なく、メディア研究全体がいまだに往年の権威的分野に偏りすぎていることを感じる。そして2007年秋現在、これまでやってきたことをすべて包括する形で、Web開発の取り組むべき課題、確かな方向性が見い出せたように思っている。

Web日記や更新監視システム、または更新情報の流通システムに関してはブログやRSSなどが台頭し、世界標準となっていった。しかしながら我が国が独自にWeb日記やその更新情報蒐集配信システムを発展させたことは今日でも無駄にはなっていないし、非常に有意義なことだった。また、筆者らが今やっている研究も、1996年当時の試行錯誤や議論を脈々と引き継いでいると考えている。「RSSを解析しなくてはならない」という考えに至ったのはそういった長い助走期間における考察から導かれた帰結である。今後もRSS解析を用いたさまざまなプロジェクトを立ち上げていく予定であるが、本稿はそれらに先だって筆者の基本方針を記しておきたい。別の言い方をすれば、本稿はこれから筆者らのWeb開発プロジェクトに参加してくれる人たちにむけて書い

た仕様書を兼ねている。

かつて、1980年代にDTPが登場したとき、それまでの植字工や組版職人などの多くは淘汰され、みながPCで出版を行うようになり、グラフィックデザイナーが組版を兼ね、あるいは執筆者がレイアウトや編集を兼ねるようになった。DTPは単なる先駆けに過ぎず、同様のことが映像編集や音楽制作にも続いて起こってきた。さらにジャーナリズムにも同様の現象が起きてきたが、これはWeb 2.0がもたらした成果の一部であるといえる。

ジャーナリズムは情報蒐集、執筆、編集、発行、流通などを新聞社や出版社などの営利企業が行ってきた。それらは大手から零細までさまざまあったが、それら細分化され分業化された職種のすべてを大手メディアは雇用し、寡占巨大企業が支配するジャーナリズムの構図を作り上げた。中小のメディアや、媒体を所有しない一般市民たちは、大手メディアに果敢に戦いを挑むことがあっても、その努力は必ずしも報われなかった。しかし今はどうだろうか。沈黙を強いられてきた一般大衆は、あるいは匿名掲示板やwiki、ブログなどの「新メディア」を用い、既存ジャーナリズムに劣らない世論形成力を持ちつつある。ニュースソースの流通経路もWebのために多様化しつつある。ニュースの蒐集から流通までを一社が牛耳るよりは、ニュースの蒐集と販売流通を分離した方がコストもずっと安く、速報性も高くなりつつある[46-47]。

このような状況においては旧来のメディアに固執するよりは、いち早く新しいパラダイムに乗り換えて先発の利益を得たいと考えるようになるだろう。新聞や出版、電波などのジャーナリズム全般は今まさにその方向に向かいつつある。そのような中でニュースソースの最小単位は何かと言えばRSSである。RSSさえ配信できれば良いという状況になれば新聞社や出版社など維持コストが高い組織は不要になるだろう。記者も編集者も、ブロガーがそうであるように、いったん個人レベルにまで還元され再編成を余儀なくされるかもしれない。情報の送り手と受け手というロールプレイ変更のコストはブログの進化とともにまったくなくなった。ユーザ一人ひとりがときにはニュースの送り手となり受け手となり得る。あとは実力次第でどこまで自分のRSSを流通させ得るか、いかに読者を獲得し得るか、それだけが問題となる。

しかしRSSが情報流通の主役になるにはまだまだ多くの関門が残っている。そもそもRSSなるものにそれほどの価値があるとは、ほとんど誰も気づいていないし、現在も価値ある使われ方がなされていない。本稿では以後、今後のRSSのあり得べき姿について考察していく。

2. 現今のRSS配信の問題

本章は現在のRSSにどのような問題があるかを知っていただき、次章につなげることを目的として記述している。RSSの一般的な解説ではなく、筆者の意図や独自解釈が相当含まれていることをあらかじめお断りしておく。

2.1 RSS形式の概要

以後の説明の都合上、ここでRSSについてごく簡単にまとめておく。

RSSは最初Netscape Communicationsから、RDF（Resource Description Framework）に基づく「サイト要約」（RDF Site Summary）として提唱された。この最初のRSSは現在RSS 1.0として広く利用されている。これとは別にRDFを用いず、簡略化されたRSS 2.0と呼ばれる形式や、RSSと非常に良く似た目的に使われているがRSSとは異なる標準化がなされたAtomなどがある。

RSS 1.0、RSS 2.0、およびAtomはいずれもXML文書フォーマットを採用しており、またその内部記述の差異については、本稿で扱う範疇では特に区別する必要はない。したがって以後は必要ない限り単にRSSと総称することにする。またRSSやAtomのことをフィードと呼ぶことがある。RSSはニュースサイトやブログなどから「配信」されるからであるが、フィード、RSSフィード、RSS配信などはほぼ同義であって、本稿でも特に区別しない。

RSSはXMLで記述されたサイト要約である。XMLはHTMLと良く似たマークアップ言語だが、人手で記述することも不可能ではないHTMLと違い、XMLをテキストエディタなどで人間が直接書いたり編集するのは現実的ではない。また、上述のようにRSSは内部ではいくつかのバージョンや種類に分かれているので、今日ではXMLパーザもしくはRSSパーザ（パーザとはここではスクリプトやプログラムなどを読み込んで解釈するソフトウェアのこと。parser。構文解析ソフトとも。4.3.1で再述する）などと呼ばれる特殊なソフトウェアによって読み書きされる。

XMLは非常に強力でかつ複雑な汎用文書フォーマットだが、ほとんどのRSSはその一部の機能しか使っていない。すなわち通常RSSにはひとつだけchannelが含まれていて、channelには通常複数のitemが含まれており、ひとつのitemにはtitle、link、descriptionがRSS 1.0必須要素として含まれている。itemは個別の記事に相当する。titleは記事のタイトルであり、linkはその元記事へのリンク、つまりURLである。descriptionは記事の要約であるが、傾向としていえることは、多くの大手ニュースサイトはdescriptionをほとんど書かないか、ダミーのdescriptionを入れているだけである。商業ニュースサイトとしては、RSSを読んだだけで記事の概要がわかってしまっては商業的な価値が半減してしまうし、読者を自分のサイトに誘導することができない。したがってRSSで概要を流通させたいという動機付けに欠けているといえる。あるニュースサイトはRSS配信とその著作権保護に非常に慎重である上に、記事とは直接関係ない広告を挿入しているものさえある。

Googleなどは積極的にRSSに比較的詳細なdescriptionをつけているし、もっと親切なサイトでは画像をitemごとに持たせているものもある。EngadgetやGizmodo、ASCIIなどがこれに該当する。このようなサイトのRSSであれば、linkをたどって本家サイトを閲覧しなくても、RSSリーダ（RSSを巡回蒐集して効率よく読めるように作られたソフトウェア）を読むだけである程度の記事の内容がわかってしまう。

RSSは今日ほとんどすべてのブログに採用されている。ブログは比較的頻繁に更新される日記、もしくはそれに類するニュースサイトであって、RSSも比較的最近更新された記事の要約として配信される。ブログは通常descriptionに十分な要約を含み、場合によっては全文をdescriptionに記述している。上に書いたEngadgetやGizmodoはブログサイトの最大手であるから、その由来からしてRSSが充実しているのは当たり前とも言える。すなわち、自分のサイトに誘導するより、自分が書いたことをできるだけ流通させることが目的であれば、RSSのdescriptionを充実させたほうがよいということになり、ブログでは以前からそのような手法が発達した。いずれの戦略（ビジネスモデル）が商業サイトにとって最終的に有利なのか、いまだ結論は出ていない。

itemには、記事を作成した時刻、つまりdateという要素がほとんどの場合含まれる（ただしRSS 1.0では副次的な項目。RSS 2.0ではpubDateとも呼ばれる）。このdate要素によってほとんどのRSSリーダはitemを新着順にソートして表示するので、dateは事実上の必須要素といえる。itemには他にもauthor、categoryなどの、有益と思われる要素が用意されているが、それらを表示させるRSSリーダはほぼ皆無である。RSSの規格には用意されていて、有益だと思われてもRSSリーダに実装されていなければほとんど無意味である。また、channelなど必ずひとつあるものもいちいち気にする必要がない。そこで、以後本稿では、RSSには複数のitemが含まれ、ひとつのitemにはおよそtitle, link, description, dateが含まれるという前提で話を進めていく。

図1はRSSとitemの関係を図示したものであり、本来itemは複数個が個別にRSSに属することを意味している。表1は便宜のためにitemに含まれる要素を列挙したものである。

2.2 現在のRSSの利用形態

RSSはブログではほとんど標準仕様であるが、このようなRSS自動生成機能がついていない通常のWebサイトでもそのサイトの要約としてRSSを利用することは可能である。図書館のOPACが書籍の要約を管理しているように、また論文のアブストラクトがデータベース化されているように、言葉通りにRSSを「サイト要約」として利用することは可能だが、RSSの用法は単なるサイト要約にとどまってはいない。もしRSSが所

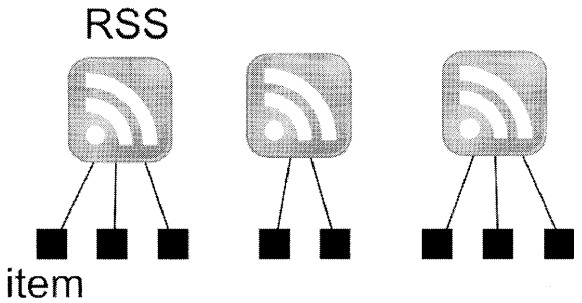

図1　RSSとitem

表1　itemの主要要素

Title	記事のタイトル。必須。
Link	オリジナル記事のURL。必須。
Description	記事の概要。必須だが、ニュースサイトではダミーの場合も多い。
Date	記事が作成された時刻。事実上の必須要素。
Author	記事の著者。ほとんど使われない。
Category	記事のカテゴリ。ブログでは多用。ニュースサイトではまれ。

期の用途どおり「サイト要約」ならばサイト全体の変更箇所を動的に反映できるような「サイトマップ」のように使われるだろう。実際にはそのような利用形態はほとんどなく、RSSには最新の更新情報だけが記述される。つまり「サイト要約」というより「更新差分」や「新着情報」のように使われることがほとんどなのである。

従来から、あるサイトが更新されたかどうかはHTTPレスポンスのLast-Modifiedを見るか、またはHTTPリクエストのIf-Modified-Sinceを使えば知ることができた。ただし、どの部分がどの程度更新されたかということはリクエストを送る側が差分を解析するなどしなくてはならない。

このようなWebサービスは「べんりくん」から「朝比奈アンテナ」などを経て今日では「はてなアンテナ」などが提供しており、一朝一夕に始まったものではない。このような1990年代から行われきたWeb日記の更新監視の自動化を推し進めたものがRSSであるといえなくもない（朝比奈アンテナ等ではRSSではなくDI（document information）という独自のメタデータを用いて更新情報をやりとりする[48]）。またRSSが特にブログと相性が良く、ブログとともに発達してきたこともまた事実である。

RSSはまた、Webブラウザのブックマークやお気に入りと呼ばれる機能の拡張ととらえることもできる。実際今日のRSSリーダというものは、巡回機能が付いたブックマークという性格のものである。最近では大手ニュースサイトやポータルサイトもRSSを配信している。天気予報や交通情報などの専門情報サイトや、通常のニュース、さらにブログなどを同じRSSリーダで仕分けして読むことができるところに利点があるといえる。さらにはRSS配信を直接行っていないサイトからも、その更新情報を抜き出してRSSを代理配信するサービスも生まれ始めている。

今日ではサーチエンジンの能力が向上して、数日前に更新されたサイトの情報もGoogleなどのサイトで検索できるようになっている。したがって優秀なサーチエンジンがあればRSSなど必要ないという見方もあり得る。確かにRSSを配信するサイトがごく一部の例外であれば、Webの全検索という荒っぽいやり方に頼らざるを得ないかもしれないが、私たちはGoogleの検索結果が不十分であり、単に検索の上位に表示されないというだけでなく、検索結果自体に多くの漏れがあることをしばしば経験している。ブログやwikiの中の検索はそれぞれのサイトにそなわっている検索機能を用いた方が完全確実であり、Googleはその代用にはなり得ない。もし将来ほとんどすべてのサイトがRSSを配信するようになり、それぞれのサイトが検索品質を保証するようになれば、Web検索というおおざっぱで不確実な方法は廃れ、ブログなどのコンパクトで効率的なやり方の方が、情報伝達の主流になる可能性もあるだろう。

2.3 RSSの問題点

今日のRSSの根本的問題を一言で言えば「RSSリーダ」が単なる「巡回ブックマーク」的利用形態にとどまっているということに尽きる。自分のお気に入りのサイトをブックマークに登録し、定期的に巡回し、その要約を読み、必要ならばオリジナルのサイトを確認に行くという利用形態は、確かに便利ではあるが、RSSが潜在的にもっている機能の一部しか利用していないと筆者は考えている。RSSが担うべき機能は、RSSの配信元であるニュースサイトやブログと、受け手であるRSSリーダだけでは実現不可能であり、RSSを解析し、再構築する中間的なサイトが不可欠である。あるいは現在のRSSリーダよりもはるかに強力な処理能力を備えた「メタRSSリーダ」が必要になるだろう。

RSSはWebのサブクラスであり、したがってWebクローラがRSSの蒐集や分析も兼ね得る、とは考えるべきではない。むしろ、Web全体のサーチエンジンはより低レベルの、生データの「採集と蓄積」であり、RSSはそれよりも高次の、よりモジュール化された情報の「能動的流通」だと考えたほうがよい。

今のRSSリーダでは、登録サイトが増すにつれてあっというまに数千数万の記事が集まってしまう。当然その全部に目を通すことはできないので、あるときRSSを読むこと自体をやめてしまうか、あるいは実際にはその一部のRSSだけを定期購読することになるだろう。キーワード検索にしても、おすすめ機能にしても、読みたい記事を絞り込むにはあまりにも貧弱な機能しかもっていない。

個人が蒐集し得るRSSの量には限りがあるので、非常に近視眼的な情報蒐集になりがちである。当人はRSSで一生懸命毎日情報蒐集しているつもりでも、実際には偏った情報だけを読むことになりかねない。つまり特定少数もしくは特定多数の情報源からニュースを蒐集することはできるが、不特定多数の、予期せぬニュースや主

張を目にすることにはならない。

RSS リーダには技術的な問題もある。驚くべきことに今日の RSS リーダは数万件の item を一度にさばくことができない。XML パーザは PC に相当な負荷をかける。あるいは現在の XML パーザは最適化が不十分であるか、もしくは、大量の item を瞬時に裁くことを想定して作られていない。数百件の item ならばなんとか快適に動く RSS リーダでも、数万件の item を一度に表示させようとするとおそらく RSS リーダはフリーズしてしまうだろう。また数万件の item を常時巡回しようとすると PC や OS に過大な負荷をかけてしまい、最悪の場合 HDD がクラッシュしてしまうことになるだろう。

ほうれん草やキャベツなどの生野菜もそのままでは決して大量に食べることはできない。私たちは馬や羊ではないので一日中草を食べているわけにはいかない。フードプロセッサにかけて熱を加えて調理して味付けして初めて十分な栄養を短い時間に（しかもおいしく）摂取することができる。RSS もまた同じである。RSS を解析し、類似記事ごとに分類し、整理整頓してその分析結果を読めばよい。そのためには数万数十万の item を常時解析するエンジンが必要であり、現在のアーキテクチャやパラダイムでは対応できていないのである。

3. 今日のメディアの問題と RSS の意義

3.1 メディアの相対化と多極化

あえて断っておくが、本稿で「メディア」というのは世論形成力を持つ電波・出版・Web などの報道媒体、あるいはそれに準じるもののこととする。従来、このメディアと呼ばれるものは、特定少数の、資本力となんらかの既得権益に基づいて成立しているメディアのこと、つまりマスメディアやマスコミュニケーションと呼ばれるものしか事実上なかった。なお社会主義国などの国家権力に基づくマスメディアやプロパガンダもありえるが、本稿の考察とは直接関係ないので除外する。

メディアといえば、かつてはこれら従来型の大手商業メディアをさし、またそれ以外の選択肢もあり得なかったから、受け手である一般大衆は唯々諾々としてその報道を受け止めてきたし、また送り手側もその特権的立場を十二分に利用してきた。言わば、大衆に対して特権的であること、一方的であることがメディアの属性であるかのように見なされてきた。このようなマスメディアに対して、私たちにはわずかに「メディアリテラシー」と呼ばれる批判的な学問体系を対抗手段として持ち、あるいは市民活動レベルの草の根運動や示威行為が行使できる程度だった。

しかし今日では大手資本によらずに世論形成を行い得るメディアが育ってきた。そのひとつは不特定多数の人々がベストエフォートで作り上げていくメディアやコンテンツがこれに相当する。具体的に例を挙げれば Wikipedia と 2 ちゃんねるが相当する。すべての wiki や匿名掲示板が Wikipedia や 2 ちゃんねるのような有用なサイトに成長するわけではない。その成立条件を考察することも学術的に非常に有意義なテーマだが、本稿では省略する。

もうひとつは特定個人によるメディアであるブログである。従来から評論家や作家、大学教員などの研究者は個人的なジャーナリズム活動を行ってきたが、彼らがブログを使うとしても最初は「本業の補完」的なものだった。しかし、今ではブログが個人レベルのジャーナリズムでは最有力となり、格段に便利になると同時に社会的にも認知されるようになってきた。今では匿名・実名・個人・団体・営利・非営利のブログが入り乱れてさかんに情報発信を行なっている。

3.2 異種メディア間の相互比較と補完

RSS がなぜ異種メディアの相対化を促進するかといえば、RSS はもっとも流通コストがかからず、従ってリスクが小さく、かつオンライン上ですべてのメディアが利用し得る共通プラットフォームだからである。RSS を利用すれば、人件費も印刷費もかからず安い費用で流通できるので、大手メディアにとっても大きなメリットだが、同時にブロガーと同じ土俵に上がることを強いられる。外に出したくない情報、出さなければ売れたかもしれない情報を外に出さなくてはならなくなる。そしてブログと大手メディアが標準化されたプラットフォーム上で比較され、自由競争にさらされることになる。独自の情報源を持たず、ただ大手メディアを追随し、共同通信や時事通信やロイターなどからニュースを買って成り立っているような地方メディアや末端メディアはその存在意義をおびやかされるだろう。

また、独自の取材能力を持つメディアも、他のメディアとの重複や差異をいちいち比較されることになる。事実、偏向報道を抑止するもっとも効果的な方法は異種メディア間の相互監視と、その差異の客観的クライテリアによる定量化にある。同種メディアどうしの監視は馴れ合いに終始するかもしれないし、比較の範囲も狭い。言論に言論で対抗しても、それは単に主義主張の違いであ

るというだけでうやむやに終わってしまう。主観に対して主観で反論しても無意味である。Aというメディアと Bというメディアでは同じニュースに対してこの程度取り上げ方、露出頻度、編集方法などに違いがあるということをできるだけ幅広い、多くのソースから蒐集・分析・比較して、受け手の側ではそれらの判断材料をもとに主体的に判断するという場を提供すればよい。

4. RSS解析によるメディア比較

筆者らがRSS解析によって行なおうしている「メディア比較」は、おそらくすでにメディアリテラシーの分野でこれまで何度も議論されてきた方法論を実践するものになるだろう。また解析結果は有効な手段によって可視化されるべきである。したがって筆者らがこれから必要とする研究分野は

- メディアリテラシー理論のサーベイ
- RSS解析のための適切なクライテリア
- 解析結果可視化のためのデザイン的手法

これらを含むと考えられる。今回、メディアリテラシーと、可視化デザイン的なアプローチについて本稿に盛り込むには間に合わなかったので、本稿では特にRSS解析の企画的部分と技術的部分に限って論述することになる。

4.1 前処理としてのメタサマリー化

RSSはサイトのタイトルや要約などのことであるから、サイトのメタデータであるといえる。しかしそのメタ化あるいはメタ処理というものは、配信元の恣意に任されていて、単純にRSSどうしを比較することはできない。したがってRSSの比較を行う前処理として、ニュースサイトや匿名掲示板などから蒐集してきたRSSを標準化する必要がある。どのようなクライテリアに基づいて標準化やメタ化を行うかで結果は大きく変わってくるが、これは統計学全般と共通する性格のものである。また、多くのRSSは著者やカテゴリ、キーワード、メディアの種別、出典などのメタデータを欠いているので、これらを補完する必要がある。すなわちメタデータのメタ化という、二度手間をかけることになるが、この処理は絶対に必要であると筆者は感じている。RSSのメタ化で得られるメタデータを仮にメタサマリーと呼ぶことにする。

4.2 メタサマリー間の相関分析

メタサマリーが得られたとすると、そのRSSには類似のitem、すなわち同一ともみなされる記事が含まれると考えられる。ブログの記事は個人個人がたまたま自分の関心のあることを書いているので、それほど相関はないと考えられるが、ニュースサイトの今日のトップ記事などのRSSを比較した場合には、おそらくはかなりの頻度で重複記事を見出すことができるだろう。その場合にtitleやdescriptionに含まれる単語の出現頻度によっておおよそどれとどれが類似の記事かを判別することができる。

特に「事件」「問題」などの一般名詞や「日本」「東京」などのありふれた固有名詞よりは、人名や組織名などの固有名詞を用いた方が適切な相関を見出しやすい。また当然ではあるが、「東京」「工芸」「大学」などの一般名詞のORで検索するよりは、「東京工芸大学」のような連語の固有名詞の方が検索精度がはるかに向上する。

そこで、筆者らはRSSどうしの類似度を計算し、図2に示すように、同一とみなされるitemと、そのitemを共有するRSSを抽出する。この過程において、人気のあるitem、すなわちいろいろなニュースサイトで同時に取り上げられている記事を抽出することができる。

同一のitemを共有するRSSを抽出できたら、次には共有item数を指標として、RSS間の距離を計測することができる。RSS間距離が近いということはそれらのニュースサイトが同じニュースに関心が高いことを意味している。RSS間距離が得られれば図3のようにRSSの

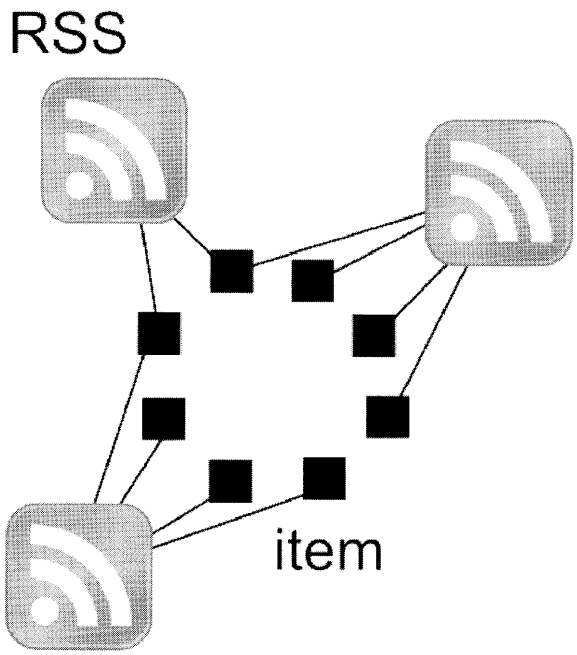

図2 類似itemを共有するRSSを抽出。

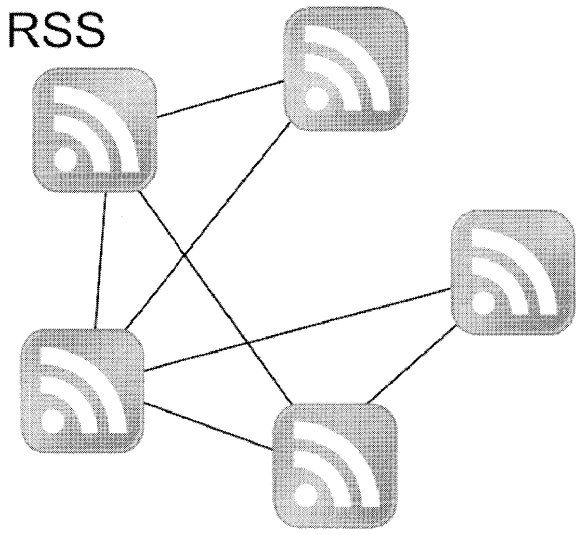

図3　RSS間距離の計測

相関を図示することが可能になる。

4.3　解析に必要な技術

RSS解析に必要な技術はWebのサーチエンジンに必要とされるものとほぼ共通する。それらを以下に列挙する。

4.3.1　XMLパーザ、もしくはRSSパーザ

今日ではRSSだけでなく多くのオンライン文書がXMLで提供されるようになり、またこれまでプレインテキスト形式やHTML形式などで作成されてきた文書もXML化されつつある。たとえばWikipediaなどのデータベースもXMLで公開されているし、ワープロ文書もXMLによる標準化が進んでいる。またInkscapeのようなグラフィックソフトウェアの保存形式もSVG（Scalable Vector Graphics画像フォーマット）のようなXML文書を用いる場合がある。あるいは、Wikipedia APIやGoogleが提供する各種のWeb APIなども、ほとんどの場合データをXML形式でやりとりする。このようにXMLパーザの重要性はますます増しつつある。

XMLのパーザは通常PHP、Perl、Rubyなどのスクリプト系言語に用意されている。筆者らはPHPのsimple_xml関数もしくはRubyのRSSライブラリなどを主に利用している。

4.3.2　正規表現・文字コード変換など

文字列の切り出しには、XML文書の場合には上記のようなXMLパーザを用いることができ、また後述する形態素解析ソフトウェアによって品詞分解することもできる。しかしそれらのソフトに頼らずに、ある特定の文字列切り出しを行うには正規表現を使って、自分でプログラムを組む必要がある。正規表現とはあるパターンと文字列を照合させて、パターンに適合した文字列を検索したり置換したりすることである。正規表現はPerl、Ruby、PHPなどのスクリプト系言語が得意とする領域であり、筆者らはRubyもしくはPHPを用いている。

4.3.3　形態素解析ソフトウェア

形態素解析（morphological analysis）とは日本語のように分かち書きされていない文章を形態素（morpheme；文章を構成する最小単位）に分解し、その品詞を決定することであり、最も需要があるのはワープロの仮名漢字変換であるが、そのほかにも構文解析を必要とするサーチエンジン、自動読み上げソフトウェアなどに使われている。形態素解析は商業的価値もあり実用性も高いものであるが、国立研究所や大学などでもさかんに研究されており、無償で利用できるものがいくつか公開されている。そのソフトウェアにはKakasi、ChaSen、Mecab、Jumanなどいくつかの種類がある[38-39]。筆者らはMeCabをRubyから使うMecab/RubyもしくはMeCabをPHPから使うPHP-Mecabなどを利用している[28]。

4.3.4　形態素解析用辞書

形態素解析ソフトウェアは形態素解析および品詞の推定のために辞書を使う。この辞書の語彙数や精度は形態素解析の正確さに直接的な影響を与える。形態素解析に用いられる辞書は、単語の意味ではなくて品詞や前後の文脈での出現確率、音読したときの読み方などを保持している。特に単なる品詞情報だけでなく、固有名詞であればそれは地域名称なのか人名なのか組織名なのかという情報を含んでいることが、分かち書きや検索には重要な意味を持つ。

このような目的に利用される辞書にはIPA辞書、UniDic、Juman付属辞書などがあるが、筆者らは現在MeCab用に手が加えられたMeCab-IPA辞書2.7.0を用いている[28]。なおIPAは独立行政法人情報処理推進機構の略称であり、IPA辞書はICOT（（財）新世代コンピュータ開発機構）フリーソフトウェアに由来するものである。

IPA辞書の特徴としては、地名や組織名などの固有名詞は比較的充実しているが、2ちゃんねるなどで多用されるカタカナ語などの新造語はほとんど収録されていないということが挙げられる。形態素解析ソフトウェアは、辞書に登録されていない単語を、登録されている単語に

分解して解釈する傾向がある。たとえば「ガンダム」という未知の単語（辞書に登録されていない単語という意味）は「ガン」と「ダム」という二つの既知の単語に分けてしまう。「ガンダム」を認識させるには、「ガンダム」という固有名詞を新たに辞書に登録しなくてはならない。つまり新造語を形態素解析で正確に認識するためにはその単語を次々に辞書登録しなくてはならず、新造語を切り出すために形態素解析を使いたくても役に立たないということになる。鶏が先か卵が先かという問題になる。筆者らはこれを「ガンダム問題」と呼んでいる。

この「ガンダム問題」を解決するために筆者らはこれまでにいくつかの取り組みを行ってきた。一つは、2ちゃんねるのスレッドタイトルや本文に連続して出現するカタカナ部分や英数字部分を正規表現で切り出して、これらを新造語としてIPA辞書に登録するという方法である。ほかにはWikipediaのデータベースの中の見出し語を名詞としてIPA辞書に登録するというものである。これらの方法で非常に多くの新造語を追加できる[31-32]。WikipediaデータベースはGPLライセンスで公開されているが、そのほかにも比較的利用制限がゆるやかな辞書データベースがいくつか公開されている[33-37]ので、これらをIPA辞書と統合することによって未知語を減らしていくことが可能であると考えている。

形態素解析にはほかにもさまざまな問題がある。我々人間であれば「外国人参政権」を「外国人」と「参政権」にすぐに分けて解釈できるが、MeCab-IPA辞書では「外国」「人参」「政権」と解釈してしまう。この解決には「外国人」「参政権」などの三字熟語の出現確率を「人参」という二字単語よりも高くするというチューニングを個別に行わなくてはならず、実現は非常に困難である。

4.3.5 データベースサーバ

大量のデータ（数百万の語彙を持つ辞書など）を効率よく、コンピュータに負担をかけず、または実時間で処理するためにはデータベースサーバを利用したほうがよく、筆者らはMySQLをもちいている。筆者らはかつてPukiWikiなどの、データベースサーバを用いずにローカルにプレインテキストファイルを保存するたぐいのソフトウェアを用いていたが[41]、登録件数が数万程度で限界に達した。そのため今ではMySQLを用いて一からオリジナルのシステムを構築している。

4.3.6 APIとの連携

Googleなどから提供されている種々のAPIをうまく利用し、自らのシステムに組み込むことによって、比較的簡単に他のデータベースとの連携を図ることができる。

5. RSSのメタサマリー化と可視化の例

本章では、筆者らが実際にこれまで取り組んできた、あるいはこれから取り組もうとしているWebサービスの事例を列挙してみる。古いものと新しいもの、進捗状況等によって、記述内容の分量や詳細さにかなりばらつきがあるが、その点はお許し願いたい。

5.1 匿名掲示板のメディアとしての意義

巨大匿名掲示板群2ちゃんねるは自然発生的に生まれては消えていったあまたの電子掲示板と同様、その濫觴期においてはアンダーグラウンドな存在であったことは、否定しようのない事実であって、故に良識ある一般市民が近づくべからざるものと考えられていた。先行する掲示板「あめぞう」の避難所として1999年に生まれた2ちゃんねるは、2000年の西鉄バスジャック事件犯人の犯行予告と見られる書き込みに代表される、匿名掲示板につきものの犯罪性、またそれに付随する訴訟や荒らし行為などで、その悪評が世間に知れ渡ることとなった。

しかし一般人の利用が年々増え続けるとともに古くからのマニアやオタクの占める割合が相対的に小さくなり、一般人にとっても読み応えのある有益な話題が増えてきた。「ニュース速報＋」や「地震速報」といった特定の掲示板は、匿名性を保ちつつも、もはや単なる無責任な発言の集まりではなく、ある種社会の公器として育ちつつあると言って良い。テレビや新聞など既存のマスコミで取り上げられるような社会性のある話題はほとんどすべて2ちゃんねるでも瞬時に取り上げられ、活発な議論が繰り広げられるので、既存のマスメディアに頼らずとも、ほとんどすべての情報が掲示板で得られるようになってきた。一般利用者のこれほどまでの増加は2ちゃんねるがすでに一定の社会的認知を獲得し、旧来の権威に対するオルタナティブな情報源となっている証拠である。実際、就職活動や市場調査などに利用されるケースが増えている。

2ちゃんねるの運営スタッフには、当初から匿名掲示板にありがちなアングラ臭を払拭し、マニアだけでなくより広い一般ユーザを獲得しようという意図があったように思われる。しかしながら未だに2ちゃんねるはたわいもない雑談や、真偽の確かでない噂話、単なるテレビ

番組の実況、誹謗中傷、意図的に事実をねじ曲げようとする大量投稿などにあふれている。フィルターにかけられていないありとあらゆる「生の情報」が2ちゃんねるには洪水のように流れており、2ちゃんねる独特の持ち味を活かしたまま、その有益な上澄み部分だけを抽出するのは困難に見える。しかしながら運営側やユーザの地道な努力はたゆみなく続けられており、雑談と上質な議論の棲み分けは目立たなくとも確実に進展している。それは何よりもそのような言論の場が必要とされ、誰がというのでなしに、利用者がみなで育てようとする暗黙の力が働いているからだと思う。

筆者らはすでに2年近くにわたって2ちゃんねるを研究し、観察し続けてきたが、その間にも2ちゃんねるの社会的地位はますます向上してきたように思われるし、また筆者らが2ちゃんねる研究の中ではどちらかと言えば先駆的でありまた継続的でもあることに自信を持ちつつもある。筆者らはWeb開発という技術的な側面から2ちゃんねるを研究しているわけだが、2ちゃんねるはもっと人文・社会学的にも幅広く研究対象とされるべきだと感じている。

2ちゃんねるのような匿名掲示板における個々の発言は単発的で取るに足りないものかもしれないが、そのサンプル数が膨大であるために、全体としては統計的に貴重な資源となっている。その統計的価値に着目した先駆的研究もいくつか存在する[3-8]。

文章を書く訓練を受けた専門家によって、社会規範に従って書かれたマスコミの記事とは対蹠的に、いわゆる「一般大衆」の生の感情の発露を2ちゃんねるに見ることができる。すなわちマスコミ主導で形成された世論とは異なる言論の場であるところに2ちゃんねるの真価がある。また、社会的地位や肩書き、権威、先入観ではなく、言論だけが説得力を持ち世論形成に影響を与えるという意味では公平であるという見方もあり得る。またいかなる人も自由に議論に参加でき、その中にはいわゆる有識者も相当数含まれていると考えられる[44-45]。

今更説明するまでもないことだが、Webの特長は「速報性」「多様性」「非拘束性」を兼ね備えていることであると言える。2ちゃんねるもまた同様である。以下に主要な他メディアと比較して、2ちゃんねるが優れている点を列挙してみる。

5.1.1 テレビとの比較

テレビは確かに速報性がある。特に、テレビ局の内部にニュース提供者がいる場合、実況中継などの場合などに。しかし、ニュースがロイターやAPなどから配信される場合にはWebと同程度の速さしか望めない。またテレビで実況されていても2ちゃんねるには数分遅れでニュースが流れるので、テレビとWebの時間差はそれほど大きな問題ではない。地震速報などがその好例である。

テレビでほんとうに新しいニュースを見るためにはテレビの前に常に張り付いていなくてはならないが、多くの場合それは時間の無駄である。当然ながら、ニュース以外の番組が流れているときにはニュースを見ることはできないし、その順番も構成もメディアまかせである。このように報道の形態が冗長なので、時間がかかる。また、せいぜい数種類のキー局がニュースを配信するだけで、しかもニュースソースを明示しないことが多いので、多面的な判断材料を入手するには向いていない。ただ単に速ければ良いというのではなく、素早くいろいろなソースのニュースを蒐集し比較できなくてはニュースとしての価値に乏しい。

タレントによるバラエティ番組仕立ての討論番組に至っては、2ちゃんねるにおける不特定多数の匿名参加者による討論にはるかに及ばない。2ちゃんねるのような不特定多数が関わる議論の場としての代替メディアは存在しない。議論はどちらも主観的なものなので、どちらが特に信頼できるということは言えない。であれば、特定少数の意見しか得られない点、番組上の脚色やスポンサーの意図などが排除しきれない点などからテレビ番組の方が議論としての質は低いだろう。

5.1.2 新聞との比較

新聞は速報性の点ではWebに劣るが一覧性、可読性、詳細さにおいてまさる。その代わり読む必要のない記事を多く読まされるとも言える。圧倒的物量をもってして一面的な論述を展開する危険性もある。かといってさまざまな新聞を毎日熟読することは不可能であるから、結局はある特定の視点の意見と情報を大量に仕入れることになる。

5.2 2ちゃんねるのメタサマリー化

今日、ほとんどの電子掲示板はRSSを配信している。phpBBやブログに付属しているBBSなどはもちろん、比較的古くから運用されている2ちゃんねるなどの掲示

板も例外ではない。問題なのはそれらのRSSが単に掲示板システムが保有しているデータをRSS化しただけのものであり、メタサマリー化されてはいないということである。

2ちゃんねるに限らないが、RSS配信元は、他のサイトのRSSとの整合性とか、RSSはそもそもどういう要素と記述を共有すべきなのかという発想を、少なくとも現時点ではもっていない。RSSはどうあるべきかという理念を共有していない。とりもなおさず、このことは、RSSリーダの提供者もそのユーザも、RSSどうしのフォーマットがてんでばらばらでもかまわずにいる証拠である。

ニュースサイトの記事と匿名掲示板の関連スレッドを結びつけるということはおそらく筆者らが2ちゃんねる可視化計画において初めて試みたことである。最初は、スレッドタイトルに出現する単語を検索語としてGoogle AJAX Search APIを用いて関連ニュース記事や関連するWebページ、動画、ブログなどをスレッドとともに表示していた。次に筆者らは発想を逆転して、ニュースサイトやブログのRSSを解析して、RSSの中に出現する単語を検索語として関連するスレッドを表示することにした。筆者らはこれを2ch Feed APIと呼んだ。

2ch Feed APIの考え方をさらに進めれば、RSS配信されるサイトであればいかなるサイトでも、相互に関連する記事やスレッドを比較・抽出できるという考えに至るし、また現状のRSSの記述をそのまま使うにはあまりにも不備が多いため、メタサマリー化する段階が必要であることに思い至る。

筆者らが2ch Feed APIを作る前に、2ch検索および2ch検索をAPI化したものはすでに存在していた（ただし2ch検索APIは現在稼働していない）。2ちゃんねるではないがGoogle AJAX Search APIやWikipedia APIのような先行事例もいくつか存在した。しかし、RSS配信を解析することによって、半自動的に、ブログやニュースサイトの連携を図るものはそれまでなかったといえる。筆者らはこのような試行錯誤を経て、2ちゃんねる、ブログ、そしてニュースサイトのRSSは、異種メディア統合のためにその共通項としてのフォーマットを持つべきであり、また現時点で共通フォーマットを持っていないのであれば、フィルターを通してメタサマリー化すべきであるという結論に到達した。

筆者らは現在、既存のRSS規格のどれかをメタサマリーの形式として採用するか、あたらにXMLのフォーマットを決めるべきか検討している。おそらくしばらくは既存のRSS 2.0を利用して、のちに独自のXML形式を採用することになると考えている。

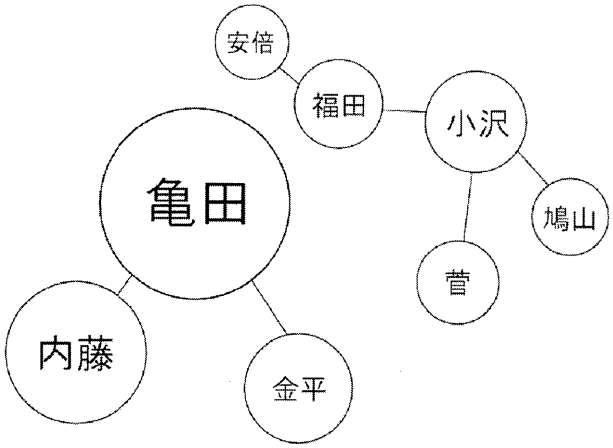

図4　記事に含まれる固有名詞相関の可視化

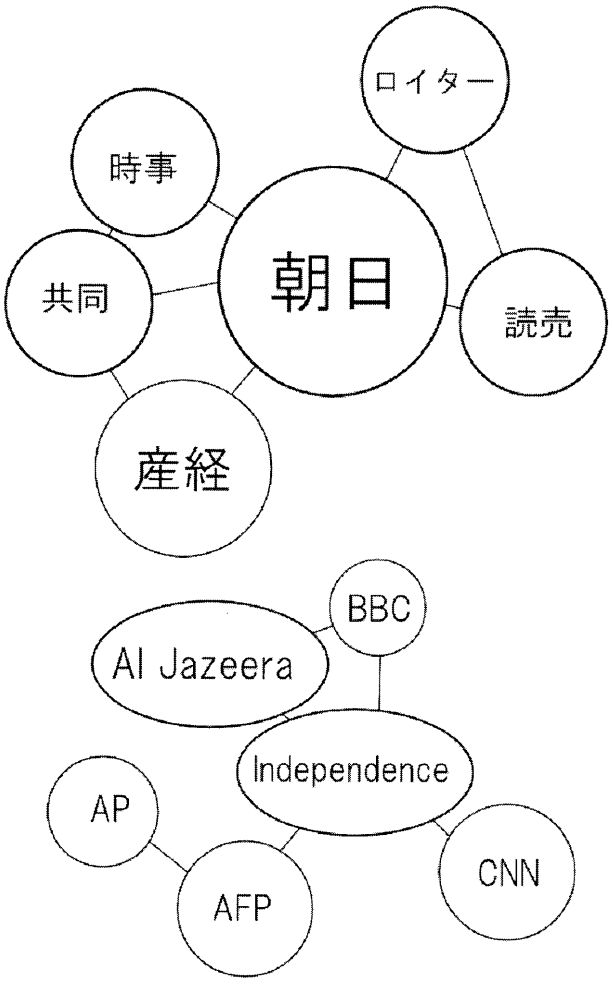

図5　ニュースサイト相関の可視化

5.3 固有名詞相関の可視化

記事中に含まれる人名などの固有名詞を抽出することは、形態素解析ソフトを使えば比較的簡単に実現することができる。今、記事のメタサマリーが得られているとすれば、同一の記事の中に出現する人名どうしの相関を計測できるので、図4のような相関図を描くことが可能である。筆者らはこれを仮に「gossip detector」と呼んでいる。これまでに筆者らは、短期間に特定の人名でニュース記事が満ちあふれることを何度か観察している。その人名はあるときは「安倍」であったり「亀田」であったりする。このような出現頻度が極端に多い単語は多少精度が低くとも、きわめて容易に検出が可能であって、現在はまだ企画段階であるが、早期実現をめざしている。

5.4 ニュースサイト相関の可視化

ニュース記事のメタサマリーが得られているとき、同じニュースを配信しているRSSサイトどうしは同様の傾向のニュースを好むと判断することができる。あるいは、同じニュース配信会社からニュースを取得している可能性がある。また、同じニュースを共有する度合いによって相関が定まるために、図5のような相関図を描くことが可能である。

現在、国内外の主要なニュースサイト（朝日、毎日、産経、NHK、TBS、FNN、時事、ロイター、AP、Al Jazeera、BBC、Independent、CNN、Washington Post、International Herald Tribune、Los Angels Times、New Zealand Herald、USA Today、Wall Street Journal、Choson Ilbo、など）およびいくつかの国内地方新聞（山陰中央新報、デーリー東北新聞、四国新聞、静岡新聞、山陽新聞、佐賀新聞、信濃毎日新聞、琉球新報など）、および、いわゆるネットニュース（Yahoo!、Livedoor、goo、2ちゃんねる、J-CAST、など）の間の相関の定量化を行っている段階であるが、さほど顕著な相関は得られていない。わずかに、BBCとAl Jazeeraの記事に重複が多いこと、2ちゃんねると産経新聞の記事にも重複が多いことなどがある程度読み取れる。

これらニュースサイトのRSS配信方法もまちまちであり、朝日は単一のRSSにすべてのカテゴリをまとめて配信する。しかし、フジサンケイグループは産経新聞、IZAニュース、MSN産経ニュースなど複数のニュースサイトでカテゴリに分けた大量のRSSを配信している。ところがそれらフジサンケイのRSSには多くの重複がある。地方新聞社は地域のニュースだけをRSS配信したり、共同通信をソースとする全国ニュースをそのまま流したりもする。このようなメディアごとの差異を吸収するためにもRSSのメタサマリー化が必要になる。

主要なニュースサイトであっても、未だにRSS配信を行っていないサイトが多い。たとえば読売は2007年11月現在でRSS配信は行っていない。日経は英文のみRSS配信を行っている。

5.5 メディア横断人気順RSSリーダ

YahooやGoogleなどのポータルサイトもしくはサーチエンジンサイトもニュースを配信しており、ニュースソースも明記してあるが、その配信元には偏りがあることが多い。たとえばYahoo.comは主にAPとAFPからのニュースを配信している。

RSSリーダに多くのニュースサイトを登録してそれらの全部に目を通せば良いかもしれないが、読むのに非常に時間がかかる。筆者らはニュース記事のメタサマリーに多く含まれる記事の順に表示する、いわば「メタRSSリーダ」と呼び得るものを試作した[29〜30]。これよって、多くのサイトをいちいち巡回する手間が大幅に省ける。

5.6 ブログパーツもしくはAPIとしての提供

上述したようなアプリケーションをブログパーツもしくはWebのAPIとして提供することを検討している。ブログパーツそのものの機能は限定的なものでよく、多くのブログに組み込んでもらうことによって筆者らが提供するWebサービスに誘導することも可能になる。筆者らがすでに提供しているものとしては先に述べた2ch Feed APIがある。

6. いくつかの問題

本章では、筆者らがこれまで取り組んできた事例の中で気づいた具体的な問題点について、まとめて述べる。

6.1 ブログのメタサマリー化

ニュース記事は相関が高く、同一時期の記事には少なからず同じ内容の記事が含まれると考えられる。5.3でも述べたように固有名詞などの特徴的な単語が、短い期間に集中して現れるために、それを検出することもたやすい。これに対してブログはめいめいが勝手に自分の関心の赴くままに執筆するために、相関は非常に低い。おそらくは相当多くのブログを集めなくては有意な相関は見いだせない。ニュースとブログの相関もまた一般には高いとは言えない。またブログのカテゴライズは容易ではない。それぞれが自分勝手にカテゴライズ（タグ付け）

しているだけなので、それらからメタサマリーを構成することは非常に難しい。筆者らはRSSの揺籃でもあったブログも同じようにメタサマリー化して扱いたいとは考えているが、今後も未知の可能性として残されるものと思われる。

6.2 RSSの著作権的問題

RSSが著作物であるとすれば、RSSを解析してメタサマリーとして再構成することは複製権や同一性保持権、財産権などを侵害する可能性がある。現在のRSSリーダの利用形態では許容されていても、筆者らが提案する「メタRSSリーダ」では認められないという問題も起こり得るかもしれない。

RSSが著作物であるかどうかに関しては、読売新聞社（YOL、Yomiuri On Line）の見出し記事に対しての著作権侵害裁判の判例[40]が参考になると思われる。少々長くなるが、判決文「当裁判所の判断」を引用してみる。

> 一般に，ニュース報道における記事見出しは，報道対象となる出来事等の内容を簡潔な表現で正確に読者に伝えるという性質から導かれる制約があるほか，使用し得る字数にもおのずと限界があることなどにも起因して，表現の選択の幅は広いとはいい難く，創作性を発揮する余地が比較的少ないことは否定し難いところであり，著作物性が肯定されることは必ずしも容易ではないものと考えられる。
>
> しかし，ニュース報道における記事見出しであるからといって，直ちにすべてが著作権法10条2項に該当して著作物性が否定されるものと即断すべきものではなく，その表現いかんでは，創作性を肯定し得る余地もないではないのであって，結局は，各記事見出しの表現を個別具体的に検討して，創作的表現であるといえるか否かを判断すべきものである。

として、個別の具体例について検討した結果

> そうすると，YOL見出しの性質や作成過程等について控訴人が種々主張するところを考慮しても，控訴人作成のYOL見出しについて一般的に著作物性が認められると断ずることはできない（後に判示するように，控訴人が多大の労力，費用をかけて取材し，記事を作成し，YOL見出しの作成に至っているからといって，そのことゆえに，当然にすべてのYOL見出しに創作性があるというべきことにはならない。）。

このように今回の事例では見出しに著作権は認めないとしたが、

> 必ずしも著作権など法律に定められた厳密な意味での権利が侵害された場合に限らず，法的保護に値する利益が違法に侵害がされた場合であれば不法行為が成立する

とし、

> 見出しのみでも有料での取引対象とされるなど独立した価値を有するものとして扱われている実情があるゆえに、
>
> 見出しは，法的保護に値する利益となり得る

とし、

> 無断で，営利の目的をもって，かつ，反復継続して（中略）見出しが作成されて間もないいわば情報の鮮度が高い時期に（中略）特段の労力を要することもなくこれらをデッドコピーないし実質的にデッドコピー（することは）社会的に許容される限度を越えたものであって，控訴人の法的保護に値する利益を違法に侵害したものとして不法行為を構成する

としている。

要するにWeb上の記事の見出しが著作物であるかどうかは個別に判断するしかないとして、見出しが著作物かどうかの定義は避けている。さらに今回の事例では見出しに創作性も著作物性も認められないとしている。さらに著作権侵害ではないが、無断で、営利目的で、反復継続的に、かつ記事が新鮮なうちにデッドコピーを利用することは社会的に許容される範囲を逸脱しているので不法行為に当たる、としている。

報道機関などのメディアは公共のものであるがゆえに「報道の自由」「信条の自由」などが認められており、その他さまざまに社会的な特権が認められている。であるがゆえに、ニュース報道には著作権的にも例外が認められている。従ってニュース記事の見出しには著作権は認められないが、多大な営為の産物であるから、法的保護の対象にはなり得る、と解釈できるだろうか。

RSSはWebページに公開された見出しとは違い、配信元が自発的かつ能動的に配信しているものであるから、Webページをクロールして切り出すのとは少々事情が異なるが、上記の判例で「見出し」を「RSS」と読み替えることは可能だろう。筆者らとしては、「営利目的でなく」「社会的に許容される限度を逸脱しない程度」にRSSを利用することは可能であると解釈してもよかろうと思う。

6.3 解析精度の問題

先に4.3.3および4.3.4で述べたように、RSSの解析精度を上げるためには形態素解析の精度を上げなくてはな

らず、そのためには「辞書を鍛える」必要がある。現在入手可能な辞書には最新の俗語や隠語、流行語などがほとんど含まれていないので、その蒐集と辞書登録作業は必須になると思われる。

6.4 可視化の問題

仮に完璧な辞書が得られて、自然言語特有の曖昧さもある程度克服されたとしても、思わしい結果が得られない場合がある。それはすなわち、先に4.冒頭でも述べたように、クライテリアが適切でない場合である。ここでクライテリアというのは、新聞社や出版社においてオーナーや編集長やあるいは現場の編集者が主観によって記事を取捨選択するのではなく、何らかのアルゴリズムを用いて、客観的な基準を定めて、数値によって自動的に抽出するのだから、そのクライテリアの決め方が、企画的に重要になるという意味である。企画（すなわち、クライテリア）には（人文科学であろうと自然科学であろうと）人間の主観は当然入ってくるが、企画自体が面白くなければ面白い結果が出るわけがない。

また、5.3と5.4で若干ふれたように、ゴシップ記事のように多くの人の関心が高いものもあれば、可視化しても必ずしも目立った特徴が得られなかったり、人の関心をひくほど面白いとは限らないことがあり得る。

データのかたまりを解析して、ある現象を数値化できたとして、それを可視化しても面白くないのならばあえて可視化する価値があるとは言えない。というのは、そもそも可視化するということは、多くの人々にとっての関心事をわかりやすく伝達するために行うものであり、伝達すべき対象が不在なのに単に学術的意味だけで可視化しても意味がないと思われるからである。もっとあからさまな言い方をするならば、可視化とはテレビのバラエティ番組と同じたぐいのものである。

可視化は往々にしてデータの取捨選択や冗長化を伴う。もちろんすべての可視化がそうだとは言わない。たとえば医療用画像の可視化などには臨床的な厳密さや見た目の現実感が求められるが、一般的に可視化とはそこまで要求水準は高くなく、どちらかと言えば専門家に見せるものというよりは素人にわかりやすく見せるためのものであって、従ってエンターテインメント目的のものだと言える。そのような可視化に特有の性質を見極めた上でうまく扱う必要があると思う。

7. おわりに

これまでRSSの謳い文句は、Webサイトをブログにすればrssが自動配信され、人の目に触れる機会が増えるので儲かりますよ、という程度のものだった。本稿は、匿名掲示板などのWeb情報のメタサマリー化を促進すれば、大手メディアや匿名掲示板やブログなどの異種メディア間の相互比較と監視、および補完が実現できるというものであって、これら異種メディア間の共通項としてRSSは有効であると主張するものである。

2ちゃんねるに代表される匿名掲示板、ブログ、wikiなどは、商業的大手メディアとは異なるメディアを通じて、無数の参加者によって自己組織的に形成された言論の場であるところに存在意義がある。これらのメディアが発達するにつれて、旧来の電波・出版などのマスメディアの相対化、メディアの多極化がますます進むだろう。メディアのオンライン化やWeb化は時代の趨勢であって、旧メディアもいつまでも電波や紙媒体だけに依存し続けることはできない。たとえ望んでいなくとも、RSSによるメタ化と相対化に巻き込まれていくだろう。

マスメディアからの報道を一方的に受け止めてきた視聴者、またはマスメディアを通じて情報を一方的に発信してきた大手企業の側も、メディアが相対化し、あるいは情報を発信し伝播する「担い手」が多様化するにしたがって、情報提供者の意図がおのずから明らかになる。送り手も受け手も情報の主体的取捨選択を迫られるようになるだろう。そのためのメディアリテラシー、および、データベースやGUIの開発が急務であると考えられる。

従来は、本稿で述べたようなメディアリテラシー的手法を、実際のメディアの出版物や放送に対して実時間で適用することは物理的に不可能であったが、今やニュースのオンライン化やRSS配信、Webサービスやデータベースシステムの発達によって可能になった。必ずしも筆者らのみならず、今後このような取り組みがいろいろな場面で本格化していくものと考えられる。現在Googleは、おそらく膨大な費用と労力を投じて、ニュースを蒐集して自動的に関連記事をカテゴライズし配信している。しかし、その著作権的問題はクリアではないし、メディアの相対性や多極性に着目しているのでもない。私たちは、RSS配信とその解析というもっとシンプルな手法で、さまざまなメディアをすばやく一覧し、メディアの受け手の立場で主体的に記事を比較・判断したいのである。現今のGoogleやYahoo!などの大手ネットニュー

スはそのようなニーズを満たしているとは言い難い。

最後に教育上の問題について一言言わせてもらうならば、このようなWeb開発は、Webでメディアリテラシーを実践したいという文系の学生、Webデザインがやりたいという芸術系の学生、そしてWebのシステム開発がやりたいという理系の学生が共同で学び得る学際の場であり、本学のような工芸融合を謳う大学に非常に適した研究・教育分野であると考えられる。いまだ手つかずの問題が多く残されていて、多少のプログラミングのスキルがあれば、最新の研究テーマにアクセスすることができる。他の分野であれば、相当にプログラムができても分野自体が成熟しているためになかなか目に見える結果は出せないだろうと思う。もっと注目されてもよい分野であると考えている。

参考文献

1）永江孝規：インターネットにおける自発的コミュニティの形成、特にWeb日記に関して、1999、尚美学園短期大学紀要。
2）http://freeman.media.t-kougei.ac.jp/~nagae/pub/wdc/
3）長　健太：2ch各板不徹底図解
http://www.asahi-net.or.jp/~cs8k-cyu/chmap/
4）郷田まり子：ふいんきり～だ～
http://www.madin.jp/finky/api.html
5）郷田まり子：機械的に2ちゃんねるの関連板にご案内
http://www.madin.jp/itaguide/
6）郷田まり子：2ちゃんねる板別スレッド一覧単語出現頻度ランキング
http://www.madin.jp/docs/word2ch1.html
7）郷田まり子：2ちゃんねる漢字の分布からみた類似板リスト
http://www.madin.jp/docs/vector2ch.html
8）郷田まり子：2ちゃんねる板別スレッド一覧漢字出現頻度ランキング
http://www.madin.jp/docs/kanji-2ch.html
9）Google: AJAX Search API
http://code.google.com/apis/ajaxsearch/
10）SimpleAPI: Wikipedia API
http://wikipedia.simpleapi.net/
11）2ちゃんねる：http://www.2ch.net/
12）3ちゃんねる：http://www.3ch.jp/
13）4ちゃんねる：http://www.4ch.in/
14）10ちゃんねる：http://www.10ch.net/
15）Level-3 BBS: http://www.lv3.net/
16）まちBBS: http://www.machi.to/
17）したらばBBS:
http://rentalbbs.livedoor.com/jbbs/
18）2NN＋2ちゃんねるニュース速報＋ナビbeta:
http://www.2nn.jp/
19）すずめ：http://stats.2ch.net/suzume.cgi
20）2ちゃんねるページビュー観測所：http://pv.40.kg/
21）スレッドランキング：http://www.bbsnews.jp/
22）レス数いろいろ集計：http://sabo2.kakiko.com/bbspost/
23）2ch鯖監視係：http://sv2ch.baila6.jp/
24）2ちゃんねる熱いスレッド（毎分更新）：http://stats.2ch.net/
25）2ちゃんねる検索：http://find.2ch.net/
26）有限会社未来検索ブラジル：Senna組み込み型全文検索エンジン
http://qwik.jp/senna/FrontPageJ.html
27）永江、比嘉：2ちゃんねる等可視化計画：
http://v2ch.media.t-kougei.ac.jp/
28）工藤　拓：MeCab: Yet Another Part - of - Speech and Morphological Analyzer
http://mecab.sourceforge.net/
29）比嘉、永江：「RSS配信の類似度を用いたニュースサイト分析」ITE冬季大会2007
30）比嘉：「RSS配信の類似度を用いたニュースサイト比較サービス」町田市立国際版画美術館学生メディアアート展2007.11.17～2
31）永江：半自動辞書編纂システムSADE
http://sade.media.t-kougei.ac.jp/
32）Google: Maps API
http://www.google.com/apis/maps/
33）F.Ko-Ji: gmapper.js
http://blog.fkoji.com/2006/09140124.html
34）The Online Plain Text English Dictionary
35）The Electronic Dictionary Research and Development Group, Monash University: JMdict/EDICT
http://www.csse.monash.edu.au/~jwb/edict_doc.html
36）The Electronic Dictionary Research and Development Group, Monash University: ENAMDICT/JMnedict
http://www.csse.monash.edu.au/~jwb/ enamdict_doc.html
37）The Electronic Dictionary Research and Development Group, Monash University: KANJIDIC2
http://www.csse.monash.edu.au/~jwb/kanjidic2/
38）国立国語研究所：形態素解析辞書UniDic
http://www.tokuteicorpus.jp/dist/
39）KAKASI-漢字→かな（ローマ字）変換プログラム
http://kakasi.namazu.org/
40）平成17年（ネ）第10049号　著作権侵害差止等請求控訴事件
41）比嘉習子、永江孝規：「Web掲示板の実時間解析と可視化手法—動的データベースとしての巨大掲示板群の有効利用—」、映像情報メディア学会技術報告、Vol. 30、No. 24、pp. 21～24、2006.3
42）比嘉習子、永江孝規：Web掲示板群の適応的巡回と再構成手法、NICOGRAPH秋期大会、2007
43）比嘉習子、永江孝規：Webユーザビリティを考慮した巨大掲示板群の解析と表示、15-6、ITE年次大会2007. see also RSS Juicer
http://juicer.media.t-kougei.ac.jp/
44）西村博之：2ちゃんねるはなぜ潰れないのか？、扶桑社新書2007.6.29.
45）ネット右翼ってどんなヤツ？嫌韓、嫌中、反プロ市民、打倒バカサヨ、別冊宝島2008.1.21.
46）中馬清福：新聞は生き残れるか、岩波新書2003.4.18.
47）歌川令三：新聞がなくなる日、草思社2005.9.15.
48）朝比奈アンテナに関するドキュメント
http://kohgushi.fastwave.gr.jp/hina-doc/