Web-log(ブログ)を活用した
分散デジタルアーカイブとセマンティックディレクトリの試み
Distributed Knowledgebase
using Web-log and Distributed Semantic Directory

グローマン・ヒロスケ
Hirosuke GROHMANN
Telecommuting-Lab.

要旨

すっかり社会に定着した感のある Web-log(ブログ)は、インターネット上に個別のアドレスを持ちトラックバックなど標準化された相互参照機構を持っている。これは膨大な記録、知識と経験のアーカイブ(archive of memory)であると考えることができる。この特性をふまえた上で、現実的な運用を考慮した適切な分散意味検索機構を用意することにより、分散したサーバ上に散らばるさまざまな知識や経験記事を有機的に結びつけて引き出すことができ、さまざまな分野における研究活動、企業活動のための、有効な情報発信、ナレッジ共有のインフラとなることを期待できる。分散アーカイブとしての Web-Log の特性を明快にするとともに、より有用なリソースとして活用するための補完手段としての意味検索システムに関する当研究所の試みを紹介する。

1．はじめに

Web2.0 的なものの１つといわれる CGM(Consumer Generated Media)は、ここ数年ですっかりインターネットの世界に定着した。Web デザインに習熟していないＰＣユーザが、ウェッブ上に体系的に情報発信できる環境が整ってきたということだ。ブログなど SNS(social network system)のサービスが、フリーウエアや安価なサービスによって定着し、技術に過度に関わることなく、純粋に情報内容(contents)そのものの発信に注力できる時代が始まったと言える。各分野における、知識や経験が時系列やキーワードを伴った記事として発信され、それを閲覧することができるようになった。現実の世界の事象が投射された「アーカイブ」がネット上に実現しつつあるということだ。これは、特定の組織が作成した集権的なアーカイブではなく、世界中のサーバに分散された「分散アーカイブ(distributed archive)」の出現である。

そこで問題になるのが、その情報の信憑性である。マスメディアの資本力や IT 技術に長けているもののみが発信していた偏った情報以外に、いままで発信されなかったさまざまな分野の専門家の意見や、マスメディアに拾われない情報を得る機会が提供されたことで、総体としての情報量が増えた反面、あまりにも膨大な情報量ゆえその価値の判断がますます難しくなっていることも事実である。Google のような商業検索サービスによって膨大なディレクトリが構築され、目的の情報を高速で探しだすことは可能だが、商業的な意図がそこに介入してくることを避けることはできない。また、情報の重要度は人によりより違っており、一元的な評価はできないことにも注意する必要がある。

そしてこれらの情報にアクセスするために何らかのディレクトリサービスが必要になってくる。分野ごとの流儀、目的による選択などにより、その前提は大きく変わってくるので、それぞれの分野における専門家の見解を容易に取り込める柔軟でセマンティックなディレクトリが提供されなければならない。そのための柔軟で拡張性のあるシステムの構築と配布を、当研究所では計画している。

本稿では、Web2.0 時代の、Web-log を題材としたメディアとしての特性の分析を行い、Web-log などのメソッドで発信されたインターネット上のリソースを「分散アーカイブ」として捉えて活用するための方法論を、「セマンティックディレクトリサービス」として提案し、解説する。

2. CGM(Customer generated Media)の特性

CMG の典型として Web-log（ブログ）を取り上げ、その特性を整理する。CMG は、従来のように HTML や Web メディア作成ツールのハンドリングに習熟した業者や技術者、Web デザイナに、インターネット上への記事の掲載を依頼するのではなく、情報の生産者、発信者自らが、直接 Web 上に記事を掲載し、情報発信するスタイルを、指している。

従来の方法論との大きな違いは、リアルタイム性である。他人に依頼すると、どうしても記事の掲出までに時間がかかり、事象と掲載のタイミングにギャップが生じる。その点、Web-log などの CGM は、情報の発生時に直ぐに記事を発信することができる。携帯電話と連携すれば、限りなくリアルタイムな発信をえる。従来も掲示板(BBS)などの仕組みでリアルタイな情報掲出が可能だったが、掲示板の場合は、記事１つ１つにアドレスが付いておらず、発言者の立場も明確でないため、多くの場合文脈が飛んでしまい、まとまった情報を取りだすことは難しい。その点、Web-log は、１つ１つの記事(log)に特有のアドレスが振られ、匿名であっても発言者の立場が明確にされており、利用価値はより高い。

Web-log が時系列な、いわば日記の形態であることで価値が低いように見なす見解もあるが、実際我々の日常の情報活動は、時系列であることが多い。メディアにおけるニュースの発信は当然として、業務報告書、航海日誌、研究における観察記録など、時系列な情報収集、記録、発信が情報行動の基本である。もちろん、熟考した記事作成も重要だが、大抵、毎日の記録を元にして、そういった洗練した記事の作成が行われるという順序だ。これら生データ(raw data)およびその導出データ(edited data)の特性把握は情報をより正確に扱うため重要である。また、日本文学史上、「枕草子」「徒然草」など日記形式の作品が多いことからも、日本人の思考形態にとっても自然であることが推察できる。

技術的な観点からみた場合、重要な特性は、「平易で標準的な交換フォーマットとプロトコルが確立されている」ことである。Web-log や Web-log にインスパイアされたニュースやコンテンツの掲出システムでは、RSS(RDF Site Summary)を代表とする XML ベースの情報交換が行われるため、異なったＯＳやミドルウエアで開発されたアプリケーション同士で、トラックバックなどによって記事同士を相互参照したり、RSS によってそのディレクトリを相互に開示し交換することができる。

3. 分散アーカイブとしての Web-log

前章の考察から、インターネット上に Web-log による記事の膨大な集積があり、それらが情報発生から比較的短い時間内に記録された生データであることが分かる。時間的な遅れが少ないということは、事象の生々しさがもっとも反映されフィルタのかかっていない現場のデータであるということだ。フィールドワークや実験の観察を主体とする研究分野、膨大なデータから情勢を見極めて判断を下さなければならないビジネスにおいても利用価値が高いことが推察できる。これらのデータを元に、多面的な評価を行い整理すれば、キーワードやテーマごとにより完成度の高い二次データを作り出すことができる。

つまりこれらは、インターネット上に分散集積された、経験、そして知識の巨大なアーカイブ(保管庫)であると考えることができる。つまり我々の記憶(memory)の一部が投影された「記憶のアーカイブ(archive of memory)」ということである。多くの見解や多分野の情報が、サーバ上に分散されて存在し、集権的に発信される「マスメディア」情報ではなく、クローズに存在しているだけの「プライベート」情報でもない、「ミドルメディア(middle media)」としての情報の集積が行われている。

では、ミドルメディアの特徴とは何であろうか？分散されて運用されていることにより、種々雑多なデータが存在しているということは、ノイズが多く運営が不安定という欠点がある反面、それだけ多面的な評価や判断を行う材料が揃っており、かつそのデータ量が圧倒的に多いという利点があるということである。少なくともデジタル化されているデータの中では、従来までは、私的に消費されてしまうか、誰の目にも留まらず、評価の舞台にも上がらなかった、しかし誰に対してでもではないにせよ、人によっては非常に有用と考えられる情報がミドルメディアとしてインターネットの上に閲覧可能な状態で掲示されるようになったことの意義は大きい。

4. Web-log の将来

Web-log という名称が、この先も残るかどうかは分からないが、CGM としての特性は、将来も、発展し続けると思われる。その典型的な特徴を以下に列挙した。この点については、情報システム学会の以前の論文(参考文献１)でも触れているので、詳細はそちらを参照されたい。。

4.1 マルチパーパス

会社における報告業務を、地域限定、専門分野、少数意見などの報道の手段、研究におけるフィールドノートや測定データの記録など、多くの目的に応用する

4.2 マルチメディア

記事として。テキストだけではなく、音声や映像など、多彩なメディアを統合して利用する

4.3 マルチターミナル

Web-log の配信先として、携帯電話、携帯プレーヤ、ゲーム機など、さまざまな端末を想定する

4.4 マルチランゲージ

複数の自然言語、言語コードを扱って言語の多重化を行い、自動翻訳機能などを狙う

4.5 キーワードディクショナリ

Web-log の中で頻繁に使われるキーワードをカタログ化し、アクセシビリティの向上を計る

4.6 グラフィックインデックス

文字以外に、地理情報や画像情報などを連携させ空間的な広がりを持たせる

4.7 データマイニング

Web-log の記録をさまざまな確度から集計分析する

4.8 イベントハンドリング

投稿された記事をトリガーとして特定の情報処理を行い、特定の情報処理の結果を自動投稿する

5. ミドルメディアの信頼性について

「ミドルメディア」有効活用する上での問題はそのデータが信頼おけるものであるかどうかという点である。情報量が膨大ということは、ノイズも多く含まれるということを意味し、その評価を行わなければならないということ。適切なデータを膨大な情報から拾いだして多面的にそれらの関連性を評価する必要がある。その分の手間や道具立てが必要、ということだ。

もちろん、商業的な検索サービスを利用することもできるが、多くの検索サービスがマーケティングという視点からの整理を行っており、必ずしもミドルメディアの視点から価値を評価することに焦点があてられていないことに注意する必要がある。また、知らないうちに、メディアによって行われる商業的な「誘導」によって、フィルタがかかってしまうことから逃れることも難しくなる。

ただし、適切な方法論や道具(tool)の助けを借りることにより、それがうまく達成できれば、マスメディア発の情報をブランド価値だけから鵜呑みしていたときに比べ、遥かに大きな発見を、また、他人に先んじた独自のアイデアや着眼点を、生み出すことができる。また、その成果や分析を、自身が発表することもできる。そして、情報評価の連鎖がだんだんとミドルメディアを豊かにしていくだろう。

情報の信頼性は、メディアとしてのブランド保証によって生まれるのではなく、多くの生データを元にした多面的な考察から積み上げられるものだと思う。それがマスメディアや権威機関に在籍する個人の誠実さや情報バランスによって生み出される場合もあるが、組織故というわけはない。インターネットの普及によって、誰もがミドルメディアに触れることができる機会が増大するにつれ、そこにアクセスするための適切な方法論が案出され、ツールやサービスとして提供される必要がある。

6. セマンティック・ディレクトリ・サービスの必要性

ミドルメディアにアクセスするには、まず手がかりになる何らかの索引「ディレクトリ(directory)」が必要である。商業的な検索エンジンはその一つだが、前章で注意したように、あくまでも手段の１つと

考えたほうが良い。ディレクトリの作成には、情報の発信者自身がディレクトリを作る流れ、そして第三者が、情報を評価し、ある「分類基準」「編集基準」に基づいてディレクトリ作成をする流れの２つがある。また、技術的には、「手動生成」と「自動生成」という要素ある。

発信者自身がそれを作成すれば、意図が良く伝わり、きめ細かなフォローができるが、反面、統一的な分類基準、概念モデルがなく多くの記事を横断的に検索するような場合、類似のデータを関連づけることが難しくなる。第三者がそれを行う場合、ある程度のものを作ることが可能だとしても、どんな概念をも包含する分類基準を作ることは難しい。

けっきょく、ある程度緩い基準で概念整理をした枠を用意し、そこに情報発信者が参加して概念をさらに深化させるといったハイブリットモデルがディレクトリ実現にはもっとも有効だと考えられる。また、概念モデルは、特定の専門家やインターネット技術者が作るべきではなく、あくまでも、その専門分野の人間が自身で作らなければ本当に有用な分類基準はにはなり得ないので、ある程度の自動化の仕組みを含めた支援環境を用意して、記事の投稿のタイミングに、簡単な手順で定義の追加ができることが望ましい。後付けしようとすると、結局は機会を逃してしまう。

7. セマンティック・ディレクトリ・サービスの具体化

分散したキーワード辞書を、運営の協力者に設置してもらい、キーワードとキーワード同士の関係をできる範囲で定義してもらうことで、キーワードツリーのメンテナンスに参加してもらう仕組みを案出した。単なるキーワードの固まりではなく、多くの参加者のセマンティックな解釈を含んだ「セマンティックなディレクトリ(semantic directory)」として分散運用され成長する仕組みである。

ディレクトリの緩い枠として、ある程度の分量(最初から網羅的なものである必要はなく、必要な分野から少しづつ構築していければ良い)の概念キーワードを提供し(情報の発信者自身も整理と定義に参加できるものとする)、辞書を、DNS のような手法で、１つの評価ツリーとして統合した上で、ツリーを辿って順次セマンティックな評価 wp しながら、キーワード検索を行い情報本体にアクセスする。掲出情報と定義の作成者が近い場合、それに合った重みづけを行う。ディレクトリサービスは、明快でオープンなプロトコルを持ち、プラットフォーム非依存に実現する。フリーウエアで配布を念頭に置く。

ベースになるキーワードの階層概念分類に、既存の公開シソーラスを下敷きに使うことは可能だろう。また、Wikipedia などとの連携を探ることも考えられる。

この方法論を、具体的な実験を横浜市の委嘱の「市民デジタルアーカイブによる調査研究」に適用して実験している。その成果を合わせて発表する。

8. まとめ

Web-log を実現するフリーウエアや RSS プロトコルの普及は目覚ましく、新しメディアのあり方がオルタナティブメディアやミドルメディアとして登場してきた。ただ、これらの膨大だが有益なアーカイブに対して、商業的な検索サービス以外に深くアクセスする手法やツールについてはまだまだ供給が十分でないと感じている。そのため、フリーウエアとして配布することを念頭に、その一旦を担う方法論を展開してツールを研究開発しており、その一旦をこの場を借りて紹介した。

参考文献／情報

[1] Web2.0: http://radar.oreilly.com/archives/2006/07/levels_of_the_game.html
[2] Weg-log:**Rebecca Blood** /2002 年刊 “**the weblog handbook**”, http://en.wikipedia.org/wiki/Blog
[3]RSS:http://web.resource.org/rss/1.0/,http://blogs.law.harvard.edu/tech/rss , http://blogs.law.harvard.edu/tech/rssVersionHistory
[4] 情報システム学会第一回研究発表「**Web-log(ブログ)など新しいメディア(ミドルメディア)を使った学会運営について**」