1G1-03

Semblog: RDF メタデータによる Web 情報の共有支援プラットフォーム

Semblog: Web Contents Distribution Platform with RDF Metadata

大向 一輝[*1] 武田 英明[*1*2]
Ikki OHMUKAI Hideaki TAKEDA

[*1]総合研究大学院大学 [*2]国立情報学研究所
The Graduate University for Advanced Studies National Institute of Informatics

We propose a personal knowledge publishing system called *Semblog* that provide an integrated environment for distributing small contents and making human relationship seamlessly. It enables people to exchange information and knowledge with easy and casual fashion in degrees of personal interest, e.g. checking, clipping, and posting. Semblog extends Weblogs by adding semantic tags to Weblog sites and entries with RSS/FOAF aggregators, for an egocentric search method and recommendation. We design a new metadata module to define personal ontology that realizes semantic relations among people and Weblog sites.

1. はじめに

本研究ではセマンティック Web 技術を利用した個人のためのコンテンツ流通支援システムを提案する．コンテンツの流通プロセスは，その作成および公開だけにとどまらず，前段階における情報収集を含めたサイクルとして捉える必要がある [1]．しかしながら，現在の Web 環境においては流通プロセス全体を統合的に支援する枠組みが用意されていない．

このような問題に対し，セマンティック Web では，Web 上のコンテンツについて機械可読なメタデータを付加し，エージェントによる情報処理を促進することを目標としている [2]．現状のセマンティック Web 技術が抱える問題点として，ユーザに対しどのようにメタデータを記述させるかといういわゆるオーサリング技術の不足が指摘されている．セマンティック Web の要素技術である Resource Description Framework（RDF）や RDF Schema（RDFS），あるいは Web Ontology Language（OWL）は，XML に由来する記法の複雑さだけではなく，適切な語彙の選択の難しさを内包している．これらの言語を専門家だけではなく一般のユーザへ普及させるためには，アノテーションやマークアップを容易にすることが重要である．

本システムでは，有効なコンテンツ流通を提供するための基盤として，RDF Site Summary（RSS）を利用する [3]．RSS は Web サイトの概要を記述するために提案されたメタデータ規格である．RSS には Web サイトのタイトルや作成者といった全体的な属性と，サイト内の各コンテンツの概要や更新時間を記述することができる．すでに一部の Web サイトでは RSS の配信が行われており，これを利用して各サイトが配信する RSS を収集し，これを整形することで多くの情報を短時間に閲覧するアグリゲータと呼ばれるアプリケーションやサービスも生まれている．

このように，RSS によって情報の取得コストは減少したと思われるが，その際に情報の選別は行われておらず，結果として得られた情報には多くのノイズが含まれていると思われる．また，上記の枠組みには新たな情報の生産活動に結びつけるといった視点がないために，先に述べた情報流通プロセス全体を支援しているとは言えない．

連絡先: 大向 一輝，総合研究大学院大学，〒 101-8430 東京都千代田区一ツ橋 2-1-2 国立情報学研究所，Tel: 03-4212-2681，Fax: 03-3556-1916，i2k@grad.nii.ac.jp

2. Weblog による情報流通

本研究では，上で述べた問題へのアプローチとして Weblog による情報流通について検討を行う．近年の Web において，個人が運営する Weblog サイトが注目されている．Weblog についての明確な定義は存在しないが，一般的には雑記や他サイトへのリンク，それに関するコメントが日々更新されるようなサイトの総称であるとされている [4]．

Weblog サイトでは，一定の読者層を想定して体系化されたコンテンツではなく，書き手が興味の赴くままに記述したスモールコンテンツを配信する形態となっていることが多い．スモールコンテンツの内容は多種多様であり，日記から批評，他サイトの紹介などフォーマットも大きく異なる．中でも，他サイトのコンテンツ紹介とそれに関するコメントは量が多く，頻繁に更新されているコンテンツの一種である．紹介はハイパーリンクやコンテンツ自体の引用によってなされ，その対象は通常の Web サイトやニュースサイト，他の Weblog サイトまでと多岐に渡る．すでにアメリカでは数十万，日本でも 10 万近い Weblog サイトが存在するともいわれる．Weblog は情報の受け手であった人々を，再編集という手順を通して情報の送り手に変えるという働きを持っているといえる．

多くの Weblog サイトでは Weblog ツールと呼ばれるコンテンツマネジメントシステム（CMS）が導入されている．Weblog ツールは Web ブラウザ上でのコンテンツ記述・編集を可能にし，その結果は即座に HTML 化されて公開される．多くの Weblog ツールは MVC（Model/View/Controller）モデルという Web アプリケーションの基本概念を踏襲しており，書き手は一度 View テンプレートを定義しておけばその後は HTML タグ等の記述をすることなしにコンテンツを公開することができる．これにより，情報公開のためのコストは従来の HTML マークアップと FTP 等によるファイルのアップロードによる方法と比較して劇的に低減する．このコストの低減が，スモールコンテンツの生産を可能にしているといえる．

また，Weblog ツールは HTML と同様に RSS を自動生成することが可能である．作成者等の属性はあらかじめユーザに初期設定として Weblog ツールに入力させたものを埋め込み，各コンテンツの概要，更新時間および RSS が指す HTML ファイルの URI 等はコンテンツが入力された際に自動的に記述される．Weblog ツールによって，新たなコストをかけることなく RSS を配信することができるため，一般のユーザにおいてもメタデータの効用が得やすくなっている．

3. Semblog プラットフォーム

本研究ではセマンティック Web 技術と Weblog ツールを用いてユーザの情報収集から生産，公開までを統合的に支援するための「Semblog: Semantic Weblog」システムを提案する．このシステムを利用することで，ユーザは各人の視点に基づく情報収集および情報発信を容易に行うことが可能になる．

3.1 Degree of Interest

本研究では，情報収集および情報発信に際して”Check”，”Clip”，および”Post”という 3 レベルの興味の度合い（Degree of Interest）を定義し，興味の強さに応じて情報の配信プロセスを変える．

最も弱い”Check”レベルとは，ユーザーが特定の Web サイトや情報ソースに日常的にアクセスすることを意味する．ユーザはその Web サイトのコンテンツ内容をあらかじめ知っているわけではないが，過去の更新履歴からどのような情報が掲載されうるかを知っている．本研究では，このような知識が情報流通において重要な働きを持つと考え，ユーザが日常的に巡回する Web サイトのリストを公開することで，そのユーザがどの分野に興味を持っているかを表明するための支援を行う．このリストには，サイトの URI やタイトルの他に，サイトに含まれるコンテンツの概要が記述されている．概要部分は登録先のサイトが更新するたびに変化するため，リスト自体が動的なコンテンツとして他の閲覧者によって頻繁にアクセスされる可能性が高まる．

次の”Clip”レベルとは，ユーザが閲覧したコンテンツの中でとくに興味があったものを指定し，保存することで，後日同じコンテンツに再びアクセスしやすいようにすることを意味する．本システムでは，”Check”レベルで登録された Web サイトに含まれるコンテンツの中でユーザがとくに興味があるものを指定することで，指定されたコンテンツのリストを作成，公開する支援を行う．Cilp レベルの情報は個々のコンテンツへの興味を表しているため，”Check”レベルでのサイト全般へのリンクよりも強い意志を表明していると考えられる．また，Check レベルのリンクはリンク先の内容が日々変わっていくが，Clip レベルのリンクは Permalink と呼ばれる永続的なリンクであり，内容が変化しない．

最後の”Post”レベルとは，非常に強い興味を持つコンテンツに対してそれを引用し，コメントを付加して新しい情報として発信することを意味する．ここでは，単なる興味だけでなく，それに伴う意見の表明がなされる．本システムにおいては，”Post”レベルの情報発信は Weblog ツールが担うものとし，その前段階の情報収集プロセスについての支援を考える．

3.2 システム構成

本システムの構成を以下に示す．本システムはサービス型とクライアント型の 2 種の RSS アグリゲータおよび検索用プログラム，そして Weblog ツールから構成される．個々のモジュールは RSS によってデータの交換が行われる．また，動的に他のモジュールを呼び出す場合には XML-RPC プロトコルによる通信を行う．Weblog ツールには MovableType などの既存のシステムを利用する．これらのツールは RSS ならびに XML-RPC をサポートしている．

3.3 RNA: RSS 収集・配信サービス

RNA は Perl で記述された CGI プログラムである．ユーザは自身が持つ Web サーバにこれを設置して運用することができる．スクリーンショットを図 1 に示す．

RNA のユーザは最初に RSS の登録を行う必要がある．他サイトが配信している RSS の URI を設定すると，RNA は HTTP 通信によってファイルを取得する．登録サイトには分類のためにカテゴリを設定することができる．登録サイトのリストは RSS 化され，他のアプリケーションで使用することができる．また，アグリゲータのサイトリストの標準フォーマットである OPML の読み込み，書き出しにも対応している．

RNA は登録された RSS を取得後，パース処理を行い，複数の RSS ツリーから 1 つの「global」RSS ツリーを構築する．global RSS ツリーは取得された全ての情報が格納されている．次に，RNA はコントローラの要求に応じて global ツリーを加工し．部分ツリーを生成する．ここでは，サイトごとの最新記事を抽出したもの，サイトにかかわらず更新時間順にコンテンツを並べるものといった 3 種類のツリーを生成する．また，ユーザはルールを記述したプラグインスクリプトを用意することで自由に部分ツリーを生成することができる．

生成された部分ツリーは，そのまま新しい RSS として配信するほか，XSL スタイルシートを用いて Web ブラウザ側もしくはサーバ側の XSLT エンジンによって可視化することが可能である．また，RNA 内部の HTML 変換エンジンによって，ユーザがテンプレートファイルを用意することで部分ツリーを HTML 化することも可能である．ここで用いられるテンプレートは HTML と類似したものになっており，XSL スタイルシートよりも理解しやすく一般ユーザにもカスタマイズしやすいものになっている．

RNA で表示するコンテンツのうち，ユーザが興味を持ったものに対しては，1 クリックでクリップリストに登録することができる．クリップされたコンテンツは独自の RSS ツリーに格納され，その他の RSS と同様に配信される．通常のツリーは内容が刻々と変化していくが，クリップのツリーからは情報が消されることはない．

RNA は取得したコンテンツのそれぞれについて後述の Track-Back リンクの有無をシステムに問い合わせ，存在する場合にはこれを抽出する．また，Description 内に記述されているハイパーリンクを同様に抽出する．抽出されたリンク情報は新たなメタデータとして配信時に追加される．

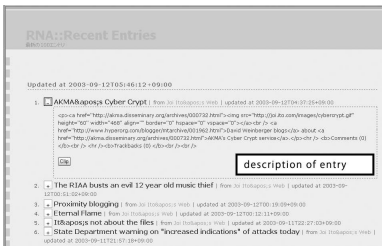

図 1: RNA: Snapshot

3.4 glucose: クライアント型 RSS アグリゲータ

glucose は Windows PC 上で動作するクライアント型 RSS アグリゲータである．既存のクライアント型アグリゲータと異なり，glucose では RNA との連携によって情報の流通プロセスを支援することを目指して開発されている．スクリーンショットを図 2 に示す．

ユーザは RNA と同様に他サイトが配信する RSS の URI を登録する．OPML の入出力にも対応する．また，RSS を配信していないいくつかのニュースサイトについてはセンサープラグ

インという Python スクリプトによって記事を切り出し，RSS 化することが可能である．

glucose によって取得された RSS は展開され，3 ペインのインターフェイスによって表示される．左ペインは RSS を配信するサイトのリスト（チャンネル）である．右上のペインには各コンテンツのタイトル，更新日時，サイト名等のリストが表示されており，各項目によってソートすることが可能である．右下のペインには選択されたコンテンツの内容が表示される．また，ティッカー（電光掲示板）機能により，ユーザに対してプッシュ形式で情報を伝えることも可能である．

RNA と同様に各コンテンツについて TrackBack を抽出することが可能である．抽出されたリンクは右上のペインでメーラの「Re:」表示と同じように表示される．また，リンク先のコンテンツは glucose が先読みすることで，快適に閲覧することができる．

興味のあるコンテンツについてユーザ自身の Weblog に記事を追加する場合には，glucose の Weblog インターフェイスを用いて直接ポストすることができる．このインターフェイスには XML-RPC を利用している．

Weblog へのポスト機能と同様に，ユーザの持つ RNA のクリップに情報を追加することができる．

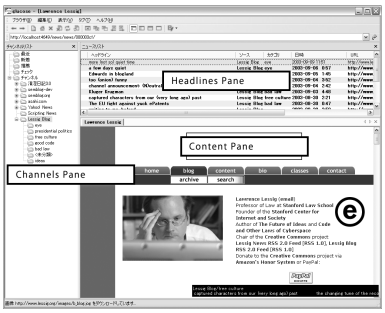

図 2: Glucose: Interface

4. Semblog プラットフォーム上のアプリケーション

われわれは Weblog ツールと RNA および glucose によって構築される RSS 流通環境を Semblog プラットフォームと呼ぶ．Semblog プラットフォームでは RSS を用いた情報収集から Check 型，Clip 型，Post 型の情報配信を行い，その結果が再び RSS として流通するというプロセスが作られる．このような RSS による情報流通プラットフォーム上での応用例として，複数の RNA を用いた情報推薦手法を提案する．

RNA では XML-RPC プロトコルによって格納された情報の入出力が可能である．これを利用して，複数の RNA の連携による情報推薦を行うことを考える．

ここでは，個々の RNA を識別するために，Friend Of A Friend（FOAF）形式のメタデータを用いる [5]．FOAF は RDF によって人間関係を記述するためのメタデータフォーマットである．FOAF には本人の名前，メールアドレス，Web ページの URI といった基本要素とともに，あるユーザ A が別のユーザ B を知っている状態を `A knows B` という形式で記述する．このリンク関係は一方向である．RNA では，FOAF による人間関係ネットワークを容易に拡張するために，1 クリックでこのリンクを張ることができる．ユーザはこの作業を繰り返し，自身の RNA を中心としたスター型ネットワークを構築することができる．

このネットワークを利用して，個々の RNA に登録されているサイトもしくはクリップの違いに基づく情報推薦を行う．以下に手順を述べる．まず，自身の持つ RNAR_0 と，パーソナルネットワーク上の RNA$R_1, \ldots, R_n$ との類似度 S_i を以下の式で求める．

$$S_i = \frac{C_i}{N_0 + Ni}$$

ここで N_i は R_i に含まれるサイト数を示し，C_i は R_0 と R_i に共通なサイト数を示す．個々の RNA はサイトの URI リスト $R_i = \{u_0, \ldots, u_k\}$ を持つ．システムは，これらの URI に対して推薦スコア $V(u)$ を以下の式で与える．

$$V_i(u) = \begin{cases} S_i & \text{if } u \in R_i \\ 0 & \text{if } u \notin R_i \ (i = 1, \ldots, n) \end{cases}$$

$$V(u) = \frac{\sum_{i=i}^{n} V_i(u)}{n}$$

u_i が R_0 に含まれていない場合には，システムは URI のリストをこのスコア順にソートしたうえでユーザに提示する．ユーザはこれらのサイトを 1 クリックで自身の RNA に登録することができる．また，クリップされたコンテンツの RSS ツリーに対しても同様の手法を適用することが可能である．

5. パーソナルオントロジーの構築

スモールコンテンツを多様な形で処理するには，オントロジーを用いたセマンティックマークアップが必要不可欠である．オントロジーの構築については様々な手法が提案されているが，精密なオントロジーをトップダウンに構築するためには，専門家の知識が必要であるとともに，それらの知識を矛盾なく組織化するためのコストが非常に大きくなる．本研究では，日常的な分類行為のうちに個人の知識体系が表出するとの考えから，そういった知識体系同士の連携という形でグローバルな意味体系をボトムアップに構築することを考える．そして，これらを実現するために，RSS および FOAF を利用して個人の知識体系を記述する枠組みを提案する．図 3 にパーソナルオントロジーの概念図を示す．

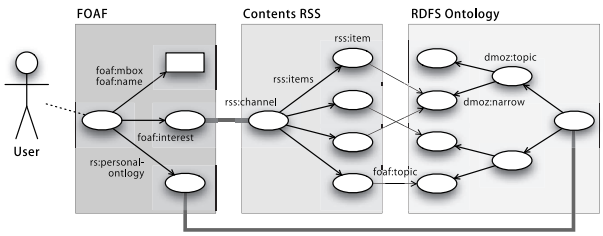

図 3: Personal Ontology Framework

本研究では，パーソナルオントロジーを「ツリー構造を持ったカテゴリの体系」であると定義する．パーソナルオントロジーは各個人が持つものであるとし，ユーザは日常的な作業として記述もしくは収集したコンテンツをカテゴリに分類する．各カテゴリのラベルは任意である．既存のオントロジーと異なり，パーソナルオントロジーをメタデータで記述するた

めには，それを作成した人との関係を示す必要がある．そこで FOAF の語彙を用いて人とオントロジーの間の関連づけを行う．

パーソナルオントロジーは個人を示す FOAF，カテゴリの構造を示す RDFS オントロジー，収集および記述したコンテンツ集合を表現するコンテンツ RSS の 3 つから構成される．

パーソナルオントロジーで用いる FOAF には，基本的なモデルに加えて`<foaf:interest>`，`<rs:personalontology>`の 2 つの要素を追加する．`<foaf:interest>`はコンテンツ RSS を示すための語彙である．`<rs:personalontology>`は RDFS オントロジーを示すために，本研究において新たに定義した語彙である．この語彙は `dc:relation` のサブクラスとして定義されており，ドメインは `foaf:Agent`，レンジは `rdf:resource` である．

RDFS オントロジーの記述形式は Open Directory Project（`http://www.dmoz.org/`）に準ずる．各ノードにはフラグメント ID を付加する．

コンテンツ RSS は既存のものと同様に記述する．既存の RSS では `dc:subject` を用いてリテラルでカテゴリを表現する場合が多い．これに対してパーソナルオントロジーで用いる RSS では，`<foaf:topic>`を用いて RDFS オントロジーのフラグメント ID を指す．なお，RSS が指し示すカテゴリは必ずしもユーザ自身の持つ RDFS オントロジー内のものでなくともよく，他人の RDFS オントロジーや，その他のグローバルオントロジー内のカテゴリを示す場合もある．

このように，FOAF，コンテンツ本体およびオントロジーをそれぞれ別のファイルに分離して管理することで，既存のモデルやアプリケーションとの後方互換性を確保し，また多様な意味を表現することが可能になる．

このフレームワークによって以下のことが可能になる．ブックマークやディレクトリを対象とする 2 つのツリーの比較手法によってカテゴリ間の類似判定とインスタンス（Web ページ）のマッピングが可能になる [6]．これによって，意味的なリンクを利用したコンテンツ検索・推薦が可能になる．また，`<foaf:knows>`のネットワーク上で上のような検索を行い，ネットワーク距離に応じたスコアを付加することでエゴセントリック検索も容易に実現することができる [**?**]．

また，図 4 に示すように，パーソナルオントロジーと ODP，Wordnet のようなグローバルオントロジーとのマッチングをあらかじめ計算しておき，このグローバルオントロジーを介して複数のパーソナルオントロジー間の類似度計算を行うことも可能である．グローバルオントロジーはパーソナルオントロジーと同じ構造をしているため，アルゴリズムを変更する必要はない．

この手法では，それぞれのユーザはコンテンツのフォルダ分け以外に特別な作業（アノテーションなど）を行う必要がない．また，十分な量のコンテンツが分類された後には，それを教師データとする学習手法を導入し，自動分類を行うことも可能である．そのような状態では，ユーザは一切の作業を行わずに新たなコンテンツが推薦される，いわゆるクエリーフリー検索が実現される．他にも，自らが記述したコンテンツに対して，グローバルオントロジーの分類キーワードを自動的に付加させるなど，新たなコンテンツのメタデータの表現力を高める働きも期待できる．

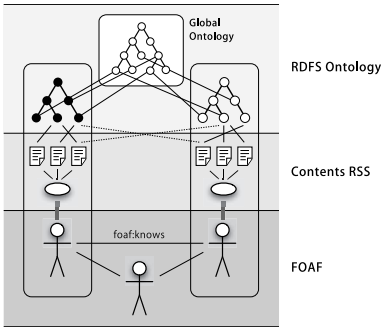

図 4: Bottom-up Ontology

6. おわりに

本研究では，セマンティック Web 技術と Weblog を利用した情報流通プラットフォームについて提案を行った．RDF に基づくメタデータを普及させるために，提案システムでは Weblog ツールによってユーザに負担をかけることなく RSS や FOAF 情報を配信する．また，Web 上のコンテンツを多様化するために，”Check”，”Clip”および”Post”という 3 種の興味に応じた情報配信や，パーソナルオントロジーの構築を行う．提案システムは 2 種の RSS アグリゲータと Weblog ツール，および応用サービスから構成される．本研究で開発されたアグリゲータは 15000 以上のダウンロード数を記録している．今後は，パーソナルオントロジーが情報流通に与える影響について実証実験を行う予定である．

参考文献

[1] Ben Shneiderman. *Leonardo's Laptop: Human Needs and the New Computing Technologies*. MIT Press, 2002.

[2] Tim Berners-Lee. A roadmap to the Semantic Web. `http://www.w3.org/DesignIssues/Semantic.html`, 1998.

[3] RDF Site Summary 1.0 Specification Working Group. RDF Site Summary (RSS) 1.0. `http://web.resource.org/rss/1.0/spec`, 2001.

[4] Rebecca Blood. *The Weblog Handbook: Practical Advice on Creating and Maintaining Your Blog*. Perseus Publishing, 2002.

[5] Dan Brickley and Libby Miller. FOAF Vocabulary Specification. `http://xmlns.com/foaf/0.1/`, 2002.

[6] M.Hamasaki and H.Takeda. Experimental Results for a Method to Discover Human Relationship based on WWW Bookmarks. *Proceedings of the Fifth International Conference on Knowledge-Based Intelligent Information Engineering Systems & Allied Technologies (KES2001)*, pp. 1291–1295, 2001.

[7] I.Ohmukai, K.Numa, and H.Takeda. Egocentric Search Method for Authoring Support in Semantic Weblog. *Workshop on Knowledge Markup and Semantic Annotation (Semannot2003)*, 2003.