RDF メタデータに基づく情報流通環境の構築と運用

Development and Operations of Information Distribution Environment based on RDF Metadata

安達 真 †
Shin ADACHI
† 早稲田大学理工学部情報学科
Dept. of Information and Computer Science, Waseda University
shn@yama.info.waseda.ac.jp

大向 一輝 ††
Ikki OHMUKAI
†† 国立情報学研究所
National Institute of Informatics
i2k@nii.ac.jp

本研究では，RSS や FOAF といったメタデータを利用し，個人間の情報流通を支援するためのアプリケーション「glucose」について議論する．glucose は情報の閲覧だけでなく，情報発信や Web サービス統合の基盤として機能する．本論文では，glucose の実装の過程で得られた知見や運用結果について述べるとともに，今後の目標であるセマンティック Web 環境の実現のために必要な要素について考察を行う．

1 はじめに

本研究では，Web コンテンツ収集から生産，公開までを統合的に支援し，個人単位の情報流通をさらに活性化させるべく，RDF メタデータを利用した情報・コミュニケーション支援環境である「glucose」の提案を行う．

glucose では，セマンティック Web および Weblog（Blog）の要素技術を利用し，Web における個人の存在を明確にすることで，HTML ページ単位であった Web コンテンツを個人の単位で集約し，情報の粒度を高めることを目指す．また，コンテンツの書き手，読み手，あるいは編集主体としての個人同士が相互にコミュニケーションを行うための基盤を提供し，個人が提供するコンテンツの多様化を進めることや，コンテンツ間のリンクに対する詳細化を実現する．

本論文では，glucose の実装の過程で得られた知見や運用結果について述べるとともに，今後の目標であるセマンティック Web 環境の実現のために必要な要素について考察を行う．

2 メタデータによる情報流通

Web における諸技術の中で，近年注目されているのが XML の普及とそれに関連するさまざまな規格やサービスである．XML は汎用的なコンテンツ記述フォーマットとして提案されたものであるが，この XML が本格的に Web で利用できるようになりつつある．

2.1 RSS

Web 上における一般的な XML の利用形態として，RSS（RDF Site Summary）を挙げることができる [1]. RSS は，XML もしくは RDF（Resource Description Framework）を用いて，サイト名や更新日時，本文の概要といった，各 Web サイトに共通する属性を記述するための統一規格である．RSS はその構文の複雑さから，ユーザー自身が記述するには向いていない．しかしながら，個人用のコンテンツマネジメントシステムとも呼ぶべき Blog ツールにより，ユーザはテキストを入力するだけで，HTML 形式への変換・配信が可能になるとともに，同じデータから RSS を自動生成することができる．Blog サイトの増加に伴って RSS は普及を始め，現在では複数の Web サイトから RSS を読み込み，記事を一覧表示する RSS アグリゲータと呼ばれるアプリケーションも登場している．

RSS および RSS アグリゲータ，そしてブロードバンド接続環境によって，ユーザは定期的にチェックされた各サイトの記事をまとめて閲覧するというプッシュ型の情報閲覧モデルが実現している．また，各サイトのコンテンツが RSS として提供されていることを利用し，必要な情報だけを選択・変換し，自分の Web サイトに掲載するということも可能になっている．このように，RSS は情報の再編集のためのインフラストラクチャーとしても機能している．

2.2 FOAF

Blog と同様に流行の兆しを見せているのがソーシャルネットワーキングサービス（SNS）である．SNS は，参加者同士が明示的にリンクし合うことでパーソナルネットワークを構築し，そのネットワーク上で限定的な情報流通を行うためのサービスである．SNS はクローズドなサービスとして提供されているが，近年では Blog 間を SNS と同様にリンクし，情報流通に付加価値を与える方法が模索されている．その際に用いられるメタデータが FOAF（Friend Of A Friend）である [2].

FOAF は，RSS と同様の RDF 構文を用いて人間関係を記述するメタデータフォーマットである．FOAF は各サイトごとに分散管理される．FOAF の分散モデルでは，双方向のリンクを表現するために両者が互いのリンクを記述する必要があるが，筆者らは以前に「FOAF TrackBack」を提案し，両者が 1 クリックで FOAF リンクを付加しあうことを可能にした [3]. これにより，オープンな分散環境であっても，既存の SNS と同様の利便性を確保できる．

3 システム概要

本研究では，情報流通の活性化を目的として，前述の Blog ツールや RSS アグリゲータ，SNS 等を組み合わせることによって，人間関係やトピックの類似性をもとにしたコンテンツの自動収集やコンテンツ配信が可能となるような基盤を提供する．また，個人が明示的に「書いた」コンテンツだけではなく，カテゴリ分けなどの「編集」プロセスや，どのサイトをチェックしているかといった「興味」を Web 上に表明していくことにより，多様な情報を配信・入手できるようにする．

本研究では，これらを実現するために RSS アグリゲータ「glucose」を実装し，配布を行っている．

3.1 glucose の基本機能

glucose は Weblog ツールや各種サービスと連携するクライアント型の RSS アグリゲータである．個々のサービスとのデータの交換は RSS によって行い，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[1]など，RSS ならびに XML-RPC をサポートしている既存のシステムを利用する．glucose のスクリーンショットを図 1 に示す.

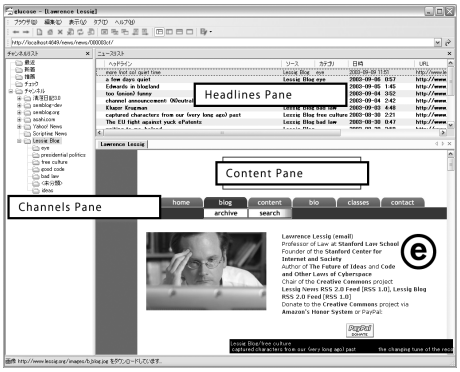

図 1: glucose

ユーザは，最初に他サイトが配信する RSS の URI を登録する．RSS サイトリストの標準フォーマットである OPML（Outline Processor Markup Language）の入出力にも対応する．また，RSS を配信していないサイトについてはセンサーと呼ばれるプラグインによって記事を切り出し，RSS に変換することが可能である．

glucose によって取得された RSS は内蔵のデータベースに格納され，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，ユーザに対して擬似プッシュ形式で情報を伝えることも可能である．

取得した各コンテンツについて TrackBack が存在する場合には，Blog ツールにこれを問い合わせ，コンテンツを抽出する．抽出されたリンクは右上のペインでメーラの「Re:（返信）」表示と同様に表示される．

また，glucose は Blog 編集インターフェイスを備えており，ユーザ自身の Blog ツールに記事を追加する場合には，直接投稿することができる．このインターフェイスには XML-RPC を利用している．これにより，ユーザは，興味のある記事の発見からそれに基づく新しい記事の作成，引用，TrackBack を一括で行うことができる．

[1] http://movabletype.jp/

3.2 glucose の拡張機能

glucose では，前節で述べた RSS アグリゲータとしての基本機能に加え，独自の拡張機能によって個人間の情報流通を支援する．これらの機能の詳細について以下に述べる．

- RSS 検索エンジンとの連携

  一般的な RSS アグリゲータでは，ユーザがあらかじめ登録したサイトのみを対象に記事一覧の表示を行う．これに加えて，glucose では外部の RSS 検索エンジンと連携し，ユーザがキーワードを設定することで，glucose に登録されていないサイトからの情報を提示することが可能になっている．実装としては，RSS 検索エンジンである「Bulkfeeds」[2]提供する API に対応し，glucose 側でユーザが入力したキーワードに合致する記事を取得して RSS を生成している．これらの作業を定期的に行うことで，ユーザは特定の話題に関して最新の情報を取得することができる．

- ソーシャルブックマークとの連携

  近年,「del.icio.us」[3]や「はてなブックマーク」[4]といったブックマーク共有サービスが注目されている．これらのサービスでは，参加者のブックマーク情報を集約することで，流行の話題を提示することや，キーワードを利用したコンテンツに対する集団的なアノテーション機能を提供している．glucose では，こういったサービスと連携し，閲覧中のコンテンツに対する付加情報の提示機能を実現している．

- 個人単位の情報の組織化

  SNS では，クローズドな環境の中で，ユーザが自身のページを持ち，会員同士が日記やレビューなどを公開しあうことでコミュニケーションを取る．しかしながら，複数の SNS が存在する場合には，アカウント情報が散逸し，同じユーザが持つ情報を取得するために各サイトにアクセスしなければならないなど，非常に煩雑である．

  そこで，glucose では，パーソナルチャネルという概念を取り入れ，友人・知人の 1 人 1 人を 1 つのチャネルとして扱うこととした．パーソナルチャネルには複数の SNS の ID や Blog の ID 等を格納することができる．glucose は，このパーソナルチャネルの情報に基づいて各サイトにアクセスし，得た情報をパーソナルチャンネル単位で表示する．これにより，さまざまな情報をより理解しやすい単位で組織化することが可能になる．なお，パーソナルチャネルは内部表現として FOAF を利用している．

- クライアントによるサービス統合

  Blog の普及とともに，新たな広告モデルやサービスが多数登場している．これらによって Blog のユーザ数は飛躍的に増加したが，現状のサービスでは，ユーザが個々のサイトにアクセスし，必要な HTML テキストをコピー・ペーストするなど，便利であるとは言えない．

  そこで，glucose では，複数のサービスを適切に利用するために，プラグインによって特定のサイトのアクセスおよびコンテンツの RSS 化だけではなく，サービス側サイトに含まれれる検索エンジンの呼び出しと結果の処理や，RSS に対応したサービスから得た情報の前処理など，さまざまな機能を実現する．これらを適切に組み合わせることにより，現在 Blog ユーザの間で流行しているアフィリエイト（E コマースのサイトと個人サイトが連携した広告）の実現に必要な工数を最小限に抑えることなどが可能になる．

- コミュニティ・イントラネットへの対応

  glucose をさまざまなコミュニティに適用させるための方法として，サイトリストの自動配信メカニズムを実装した．これは，コミュニティの管理者がユーザの利用する glucose に対して自由に RSS チャンネルを設定できるというもので，これによって RSS の課題であったプッシュ型の情報配信が可能になる．

  企業内・組織内での利用を活性化するためには，認証機構を強化し，イントラネット内に存在する RSS にアクセスすることが可能になった．これによって，普及が進んでいる Blog を利用した組織内のナレッジマネジメントを支援することが可能になる．

4 システムの設計および実装

glucose の開発に際して，以下の設計指針に基づき実装を行った．

[2] http://bulkfeeds.net/
[3] http://del.icio.us/
[4] http://b.hatena.ne.jp/

glucose は，Windows の基本ライブラリに加え，多数のオープンソースライブラリから構成されている．Web サービスやメタデータを取り扱うにあたっては，新たなサービスや規格が頻繁に登場するため，これに合わせたライブラリの更新が必須である．オープンソースライブラリはコミュニティベースで開発が行われており，機能向上やセキュリティ対策などのアップデート頻度が高い．

glucose では，メタデータの処理系に XML パーサの MiX[5]，データベースに SQLite Database[6]を用いている．これらは Windows アプリケーション用に DLL として提供されている．また，開発環境として Boost Library[7]，Window Template Library（WTL）[8]を利用している．なお，HTML の表示部には Microsoft が提供する Internet Explorer コンポーネントを利用している．

glucose に特有の機能としては，スクリプト言語 Python の処理系を内蔵することで，ソフトウェアにさまざまな機能拡張を施すことが可能である．前述の，RSS 非対応サイトから RSS を生成するセンサープラグインは，この機能を利用しており，Python スクリプトとして提供されている．センサー仕様は公開されており，ユーザが自由に作成し，実行することが可能である．

以上のように，公開ライブラリを利用することは，機能やメンテナンス性の向上のために重要であるが，RSS アグリゲータという特定の目的に対して十分な機能が提供されていないことがある．このような場合には，独自のライブラリを実装している．独自ライブラリの機能を以下に述べる．

- HTTP 通信ライブラリ

  RSS アグリゲータは，登録されたサイトに対して定期的にアクセスし，情報を取得するため，サイト数が多い場合にはクライアント側，サーバ側の双方に対して大きな負荷を与える可能性がある．この問題を軽減するため，RSS 取得時の HTTP 通信に関しては，ファイルの更新日時を細かくチェックし，前回の取得時から変化がなければ新たに取得しないといったアクセス制御機能を持つ通信ライブラリの実装を行った．

- XML/RDF アグリゲーションライブラリ

  XML/RDF アグリゲーションライブラリは，RSS だけでなく，FOAF や Atom などのメタデータフォーマットに対応するために独自実装としている．このライブラリでは，前述の Python 処理系ならびにセンサープラグインと連携し，HTML の解析結果を RSS や FOAF に変換することが可能である．

- RDF データベース

  データベース部では，RDF の記法に基づいて主語・述語・目的語のトリプル形式でメタデータを格納している．データベースに対するクエリ処理等については，RDF に適したインターフェイスライブラリを実装している．また，登録されたサイトのリストは，データベースではなく OPML 形式で保存している．

- HTTP サーバ

  glucose は，クライアント PC で動作する HTTP サーバを搭載し，ブラウザコンポーネントとの通信を行うことができる．これにより，本体に対するデータの入出力インターフェイスをブラウザコンポーネントに委譲することが可能になり，その結果ユーザインターフェイスの開発コストを軽減することができる．また，Web サービスとの親和性の高いことや，既存の Web アプリケーションの組み込みが可能であるなど，拡張性の高さに寄与している．

5 システムの運用と知見

glucose は 2004 年 7 月より配布を開始し，現在までのダウンロード数は 30 万を超える（派生バージョンを含む）．以下では配布・運用の過程で得られた，メタデータ流通に関する課題や知見について述べる．

5.1 RDF メタデータ流通における課題

RSS は仕様策定の歴史的経緯より，4 つのバージョン（0.91/0.92/1.0/2.0）が独立に存在し，Blog ツールやニュースサイトの対応はそれぞれについて異なる．また，近年では RSS に代わるメタデータフォーマットとして Atom が提案されている．これらの規格は用途が同じであるにもかかわらず，XML 文書としての構造が異なるために，アグリゲータ側で処理を振り分ける必要がある．また，運用側の誤解から，ある規格の文書に別の規格の記法が混在しているなど

[5] http://mix.sourceforge.jp/
[6] http://www.sqlite.org/
[7] http://www.boost.org/
[8] http://sourceforge.net/projects/wtl/

の例も多数見受けられ，これらを適切に処理するための例外系の実装コストが増加する．

また，規格に沿ったメタデータであっても，規格自体が解釈の多様性を許している場合には，アグリゲータの設計時に起こりうるすべての可能性について考慮する必要がある．

例として，RSS 1.0 における channel 要素の主語は RSS 自身の URI であるべきか，サイトの URI であるべきかは厳密には言及されていない．また，RSS 検索エンジンなどで利用される，リソースの取得元を示すメタデータ ag:sourceURL[9]も同様である．こういった場合，実装では channel 要素のリソースを取得し，内容を読み込んで RSS か HTML かを判別した後，後者である場合は RSS Auto Discovery の規格に則って RSS の URI を得るという手順を取っている．こういった作業は実装時ならびに実行時の双方でコスト要因となる．

また，FOAF の規格では，個人の特定にメールアドレスを利用することとなっているが，Web コンテンツからのメールアドレスの抽出や，メールアドレスが存在しない場合における複数コンテンツの作者の同一性判定には，多段階の処理が必要となり，実行速度が低下する．このため，glucose の内部では仮想的にユーザ用の URI を生成することで問題に対処しているが，本来であればメタデータフォーマットでの規定と運用方針の統一が必要であると思われる．

より低次の問題としては，プロパティの誤りも数多い．なかでも文字コードの誤りは顕著である．XML 宣言部に記述された文字コード名と実際の文書の文字コードが異なる事例や，部分ごとに使用されている文字コードが異なる事例が見受けられる．後者の問題は，RDF メタデータの流通によって複数サイトから配信されたメタデータを 1 つのファイルに統合する際に生じやすい．

glucose では，XML 宣言部の文字コードと，glucose のテキスト処理系を通して推定された文字コードを比較して，可能性の高い方をメタデータに埋め込む．また，文書内で文字コードが変化する問題に対しては，文書内の全てのブロックについて文字コード判定を行い，UTF-8 に統一する処理を行う．

その他の単純な文法エラーは多数の事例がある．これらについて，処理系の機能追加によって対処を行うのか，配布側に修正を要望するかという判断を行うことは難しい．ユーザの視点では前者が求められるが，過度に対応を行うことでメタデータフォーマット自体の信頼性が低下する可能性がある．

[9] http://web.resource.org/rss/1.0/modules/aggregation/

5.2 セマンティック Web 環境の実現に向けて

今後，glucose はセマンティック Web のためのプラットフォームとして，さらに汎用性を高めることを目標としている．そのためには，メタデータを効率的に管理するためのデータベース機能の強化が必要である．具体的な課題としては，RDF Scheme（RDFS）の利用および RDF の推論（RDF Query）への対応が挙げられる．

現状の glucose では，RDF をトリプルとしてデータベースに格納しているため，目的のグラフを得るためには複数回のクエリを発行する必要がある．また，RDFS の処理系を持たないため，RDF のバリデーションを用途に応じて実装しなければならないなど，拡張性に問題がある．

これらを実装することで，Blog 記事を格納した RSS データベースより，FOAF に登録された友人に関するものだけを抜き出すなど，多様な用途を提供することが可能になると期待される．

6 まとめ

本論文では，RSS や FOAF といったメタデータを利用し，個人間の情報流通を支援するためのアプリケーション「glucose」について述べた．glucose は情報の閲覧だけでなく，情報発信や外部の Web サービスを統合するための基盤として機能する．glucose は現在までに多くのユーザに利用されているが，運用の過程ではメタデータが内包する多くの問題が露わになった．今後は，こういった問題に対してソフトウェアの改良を施すだけではなく，コミュニティ活動を通じてメタデータ環境全体の改善を図ることを目指す．

参考文献

[1] B.Hammersley. *Content Syndication with RSS*. O'Reilly & Associates, 2003.

[2] D.Brickley and L.Miller. FOAF Vocabulary Specification. http://xmlns.com/foaf/0.1/, 2004.

[3] I.Ohmukai, H.Takeda, K.Numa, M.Hamasaki, and S.Adachi. Metadata-driven Personal Knowledge Publishing. *Proceedings of the Third International Semantic Web Conference (ISWC2004)*, 2004.