社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－GN－63 (23)
2007／3／22

ニュースを選別する手間を軽減させるRSSリーダの提案

高林 光† 古井 陽之助† 速水 治夫†‡

†神奈川工科大学 情報学部
‡神奈川工科大学 大学院 工学研究科 情報工学専攻

概要

Webサイトを要約するために，RSSというデータ形式のフィードを用いる．RSSフィードはWebサイトの更新情報を要約し，配信する．RSSリーダと呼ばれるアプリケーションによってRSSフィードを受信することで，直接Webサイトを閲覧することなく，更新状況を確認できる．しかし，昨今の情報爆発により，更新情報だけでもその量は膨大になり，必要な情報をえり抜く労力は，RSSリーダ利用者の負担となる．本研究では，利用者のコンテンツ閲覧履歴から導き出した利用者自身の嗜好を用いて，収集した膨大な量の情報を取捨選択するアルゴリズムをRSSリーダに組み込み，評価実験を行った．

An aggregator that reduces the time and effort required to pick out news articles

Kou Takabayashi† Younosuke Furui† Haruo Hayami†‡

†Kanagawa Institute of Technology
‡Graduate School of Engineering, Kanagawa Institute of Technology

Abstract

RSS is a family of web feed formats used to summarize websites. A reader can use an aggregator, which is an application program for retrieving the summary (or RSS feeds), to reduce the time and effort needed to regularly check the websites for updates. Today, however, there is a large amount of RSS feeds because of the information explosion, and a reader has to spend much time and effort to pick out interesting articles. We designed an algorithm for analyzing each reader's access history to acquire his/her preference and pick interesting titles out of the RSS feeds. We also integrated the algorithm into our aggregator, and conducted an experimental evaluation.

1 はじめに

情報発信手段の敷居が低くなり，総表現社会化によって，Webの情報量は急激に増加した．[1] このような情報爆発時代であっても，情報の意味を記述するメタデータを活用することで，大量情報の中から必要な情報を適切なタイミングで利用することができる．そこで武田らは，「メタデータをどう活用するか」が鍵になると，メタデータの重要性を訴えている．[2] メタデータの一例として，Webサイトのコンテンツを配信するデータである**RSSフィード**が挙げられる．

RSSリーダと呼ばれるアプリケーションを用いてRSSフィードを受信することにより，RSSフィード配信元Webサイトの更新状況を確認することができる．

しかし，昨今の情報爆発により，更新情報だけでもその量は膨大になり，そこから必要な情報のみを選り抜く労力は，全てRSSリーダ利用者の負担となる．受信するRSSフィードの量が増えるほど，こうした負担は大きくなる．

既存のRSSリーダの一部には，利用者全体の傾向

をもとにして，コンテンツを推薦する機能をもつものもあるが，それらは利用者自身の嗜好に合ったコンテンツを提示するまでには至っていない．

そこで本研究では，利用者自身の嗜好に合ったコンテンツを提示する RSS リーダを試作した．

2 従来の技術

2.1 RSS

RSS は，Web サイトのコンテンツのメタデータ (フィード) を記述するための形式の一種である．RSS フィードは，Web サイトのコンテンツの要約である．ただし，RSS には現在以下の 3 種類の規格がある．

- Really Simple Syndication(RSS2.0)
- RDF Site Summary(RSS1.0)
- Rich Site Summary(RSS0.91)

RSS は，XML(eXtensible Markup Language) を利用したデータ形式である．XML は，開始タグ (<~>) と終了タグ (< / ~>) で文書を修飾することによって意味づけを行ったり階層構造を表現したりすることのできる言語である．RSS フィードには，こうしたタグによって，「Web サイト自体の情報」がひとつと，「Web サイト内のコンテンツ」が複数，記述されている．RSS2.0 において定義されている，基本的な階層構造を図 2.1 に示す．

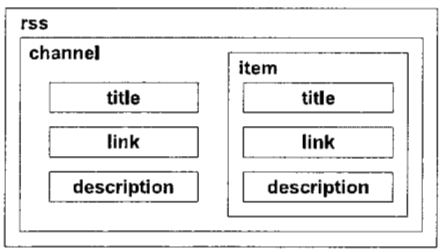

図 2.1 RSS2.0 における基本的な階層構造

図 2.1 の各要素は，RSS2.0 で定義されている属性である．各要素の意味を以下に列挙する．

- rss:自身が RSS であることを表す．
  - channel:Web サイト自体の情報．
    - title:Web サイトのタイトル．
    - link:Web サイトのアドレス．
    - description:Web サイトの説明．
    - item:Web サイトのコンテンツ．
      - title:コンテンツのタイトル．
      - link:コンテンツのアドレス
      - description:コンテンツの説明．

RSS は，他にも画像情報など，様々な種類のデータを表現することができるが，本研究では利用しないため，省略する．

2.2 様々な RSS リーダ

RSS リーダの基本的な動作は，RSS の各要素を抽出して，画面に表示するというものである．利用者は，表示にもとづいて閲覧したいコンテンツを選び，閲覧する．

RSS リーダの形態は様々であるが，代表的なものを以下に列挙する．

a) スタンドアロンで動作する RSS リーダ
- ウィンドウ型
- ティッカー型

b) ブラウザやメーラに付属する RSS リーダ

c) Web サーバで動作する RSS リーダ

a) のウィンドウ型は，アプリケーション・ウィンドウで動作する RSS リーダである．また，ティッカー型はデスクトップに常駐し，テレビのテロップのようにコンテンツを表示する．b) は，Web ブラウザやメーラのツールバーなどに，ひとつの機能として付属している．c) は，Web ブラウザを用いて利用する，Web アプリケーションの RSS リーダである．[3)]

なお，本研究の試作 RSS リーダは c) の形態である．

3 既存の RSS リーダの問題点

RSS リーダには，数多くの RSS フィードを登録すると，受信した RSS フィードに記述されている情報の収拾がつかなくなるという欠点がある．

この原因は，既存の RSS リーダは，RSS フィードに記述された情報を表示するのみであり，その取捨選択は利用者自身が行わなければならないという点にあると，本研究では考えた．利用者が RSS リーダを使い込めば使い込むほど，多くの RSS フィードを登録するため，コンテンツの取捨選択にかかる労力も増加する．

そのような利用者の労力を軽減するため，既存の RSS リーダの中には，利用者にコンテンツを薦める

ものもある．

例えば，ブログラインズ[4]というRSSリーダは，利用者が購読中のRSSフィードと似たものを，別の複数の利用者が購読しているRSSフィードから選択し，推薦する．

また，ニューズモンスター[5]というRSSリーダは，コンテンツの関連と評価を報告するシステムを組み込むことで，コンテンツをレーティングする．

しかし，これらは利用者全体の利用特性を用いているため，利用者個人の嗜好を十分に反映していない．

4 問題を解決するためのアルゴリズム

前節で述べた問題点を解決するため，本研究では，利用毎のコンテンツ閲覧履歴を分析し，その結果として利用者毎の嗜好を反映したアクセス傾向（利用特性）を得て，それをもとにコンテンツを取捨選択するアルゴリズムを考案した．このアルゴリズムは以下の通りである．

まず，利用特性を決定するために，試作RSSリーダは，あらかじめ利用者によって閲覧されたコンテンツのタイトルに含まれる単語を抽出し，単語毎に出現回数を数える．出現回数が多い単語は利用者の嗜好に合っている可能性が高い．これらの単語と出現回数の組の集合が利用特性である．

次に，タイトル及び説明に，利用特性として得られた単語が含まれているコンテンツを利用者の嗜好に合っていると判別する．

最後に，利用者の嗜好に合っていると判別されたコンテンツのタイトルを利用者に提示する．利用者は，提示されたタイトルを画面上でクリックするなどの操作によって，そのコンテンツを閲覧することができる．

5 試作RSSリーダ

5.1 システム概要

本研究で試作したRSSリーダの構成図を図5.1に示す．

試作RSSリーダは，事前にWebからRSSフィードを受信・解析し，その結果をデータベースへ格納する（**RSS受信・解析部**）．

利用者は，Webブラウザを用いて試作RSSリーダが存在するサーバへアクセスし，IDとパスワードを登録した後，ログインする．この操作により，試作RSSリーダは利用者を識別し，利用者毎に利用特性等のデータを管理する（**利用者管理部**）．

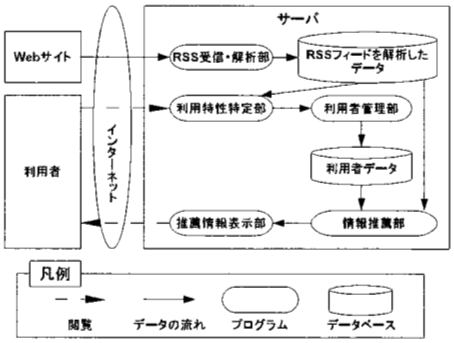

図5.1 試作RSSリーダの構成図

利用者によるログイン操作後，試作RSSリーダは，RSS受信・解析部が格納したデータを用いて，利用者毎の利用特性を求めデータベースへ格納する（**利用特性特定部**）．試作RSSリーダは，得られた利用特性をもとに，推薦すべきコンテンツを決定し（**情報推薦部**），表示する（**推薦情報表示部**）．

5.2 システム詳細

5.2.1 RSS受信・解析部

RSS受信・解析部は，RSSフィードを受信・解析して試作RSSリーダが必要とする情報をデータベースへ格納し，それ以外の情報を破棄する．

試作RSSリーダは，RSSフィードに記述されている情報のうち，Webサイト自体の情報と，コンテンツの情報のタイトル及び説明のみを利用する．

5.2.2 利用特性特定部

利用特性特定部は，利用者の利用特性を抽出する．

まずRSS受信・解析部で格納した全ての情報を表示し，その中から，利用者にとって興味のあるコンテンツを利用者に閲覧してもらう．

次に，利用者が閲覧したコンテンツのタイトルを形態素解析する．形態素解析とは，自然言語で書かれた文書を，言語の中で意味を持つ最小単位形態素の列に分割し，品詞を特定する処理である．[6]

最後に，形態素解析によって分割した単語の出現回数を数え，利用特性としてデータベースへ格納する．

5.2.3 利用者管理部

利用者管理部は，利用者のID，パスワード及び利用特性を利用者毎に管理する．

5.2.4 情報推薦部

情報推薦部では，利用者の利用特性に合ったコンテンツを推薦する．本研究で考案したアルゴリズムにおけるデータの流れを図 5.2 に示す．

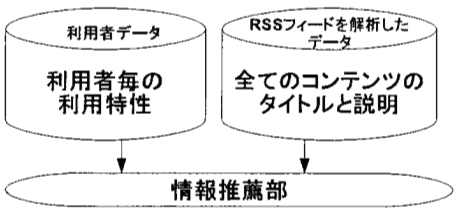

図 5.2　アルゴリズムにおけるデータの流れ

情報推薦部は，利用特性のうち単語の出現回数が多い上位 10 単語を用いて，全てのコンテンツの中から，タイトル及び説明に上述の単語が含まれているコンテンツを抽出する．

5.2.5 推薦情報表示部

推薦情報表示部は，情報推薦部によって抽出されたコンテンツを表示する．

5.3 システム画面

試作 RSS リーダの画面遷移を図 5.3 に示す．

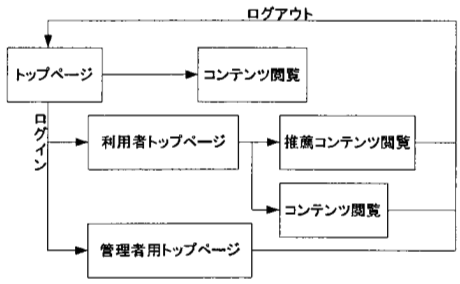

図 5.3　試作 RSS リーダの画面遷移図

試作 RSS リーダは，利用者がログインせずともトップページからコンテンツを閲覧できる．

利用者トップページを図 5.4 に示す．

試作 RSS リーダは，フレームにより 3 つに画面が分けられている．図 5.4 の各フレームの役割を列挙する．

① 操作ボタン群
② 補助的な情報表示部
③ コンテンツ表示部（主表示部）

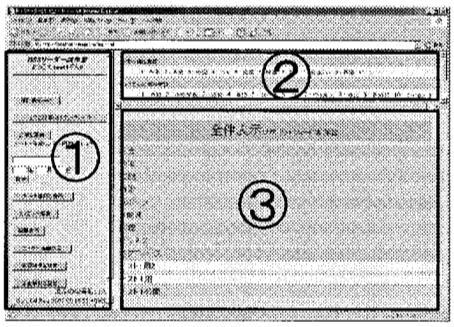

図 5.4 利用者トップページ

① の操作ボタン群には，利用特性特定部，推薦情報表示部などを利用するためのボタンが配置されている．

② の補助的な情報表示部は，利用者毎の利用特性として格納した単語の出現回数のうち，出現回数が多い上位 10 単語が表示されている．

③ のコンテンツ表示部は，利用特性特定部及び推薦情報表示部が，コンテンツを表示するために用いる．

6 評価実験

6.1 評価尺度

試作 RSS リーダの性能を評価するために，再現率と精度を評価尺度として用いる．

再現率は試作 RSS リーダに格納されている全てのコンテンツのうち，利用者の興味があるコンテンツに占める，試作 RSS リーダが提示したコンテンツの割合，**精度**は試作 RSS リーダが提示したコンテンツに占める，利用者の興味があるコンテンツの割合である．

再現率と精度は，図 6.1 に示す記号を用いて，以下の計算式 (式 6-1)，(式 6-2) で算出できる．

$$\text{再現率} = \frac{D}{B} \quad \textbf{(式 6-1)}$$

$$\text{精度} = \frac{D}{C} \quad \textbf{(式 6-2)}$$

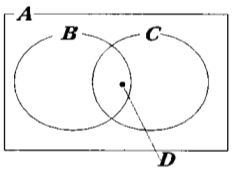

図 6.1　再現率と精度

ただし，

A :試作 RSS リーダが格納した全てのコンテンツ．

B :利用者の興味がある全てのコンテンツ．

C :試作RSSリーダが推薦した全てのコンテンツ.

D :B 及び，C の共通部分.

6.2 実験方法

今回の評価は，実験参加者3名 (α, β, γ) に，様々なニュースサイトのニュースを，利用特性特定部を用いて，1日分閲覧してもらった．これにより，試作RSSリーダが実験参加者の利用特性を特定し，コンテンツの推薦が可能となったため，別の日に，6.1節の評価尺度にもとづいて評価を行った．

6.3 実験結果

実験の結果を表6.1に示す．表6.1の A, B, C, D は，図6.1の A, B, C, D に対応している．

表 6.1 実験結果

参加者	A	B	C	D	再現率	精度
α	111	33	9	9	0.27	1
β	278	23	14	4	0.17	0.28
γ	278	71	46	9	0.12	0.20

6.4 考察

6.4.1 実験結果からの考察

6.3節の結果より，実験参加者によって再現率および，精度は大きく異なる事が分かった．利用者 α は，精度が高いが再現率は低かった．また，利用者 β, γ は再現率及び精度が共に低かった．

よって，試作RSSリーダは利用者の嗜好を利用特性として完全には捉えることができていないということが分かった．その原因は，利用特性を単語の出現頻度だけで決定している事にあり，改善の余地がある．

6.4.2 実験方法についての考察

今回の実験では，実験参加者に，事前に格納されている全てのコンテンツを閲覧してもらい，興味の有無を判別してから，コンテンツの推薦表示を行った．今後，6.2節の実験方法だけではなく，試作RSSリーダが推薦したコンテンツが利用者の嗜好に合っているかどうかを，実験参加者に推薦後も判別してもらい，精度を算出するなど，多角的な評価を行うことで，試作RSSリーダの性能の向上に役立てる．

7 おわりに

本研究では，既存のRSSリーダでは膨大な量のコンテンツを取捨選択するために利用者にかかる負担が大きいと言う問題を解決するため，利用者のコンテンツ閲覧履歴から調べた利用特性を用いて，自動的にコンテンツを取捨選択するアルゴリズムを検討した．また，このアルゴリズムを取り入れたRSSリーダを開発し，評価実験を行った．

今回の実験結果を見る限り，このアルゴリズムではまだ十分な性能が得られなかった．今後はアルゴリズムの改良と，それによる試作RSSリーダの性能向上を目指す．

参考文献

1) 梅田望夫：ウェブ進化論-本当の大変化はこれから始まる，ちくま新書 (2006.05)
2) 武田英明，大向一輝，福原和宏，Sakya Aman，沼晃介：メタデータの構造に関する研究，情報爆発時代に向けた新しいIT基盤技術の研究，文部科学省科学研究費補助金特定領域研究平成18年度概要，p. 34(2007.01)
3) 水野貴明：詳解RSS，株式会社ディー・アート (2006.05)
4) Bloglines(http://www.bloglines.com/)
5) Newsmonster(http://www.newsmonster.org/)
6) 三枝優一，古井陽之助，納富一宏，速水治夫：形態素解析におけるデータベース解析方式の提案，マルチメディア，分散，強調とモバイルシンポジウム論文集 (II)，Vol. 2006，No. 6，pp. 573-576(2006.07)