学内ポータルサイト構築のための分散Web情報の一元化

渦尾, 秀勝
九州大学大学院システム情報科学府 : 修士

甲斐, 啓文
九州大学大学院システム情報科学府 : 修士

伊東, 栄典
九州大学情報基盤センター : 准教授 : 情報学

廣川, 佐千男
九州大学情報基盤センター : 教授 : 情報学

出版情報：情報処理学会 マルチメディア, 分散, 協調とモバイルシンポジウム (DICOMO 2005), pp.465-468, 2005-07. 情報処理学会
バージョン：
権利関係：

学内ポータルサイト構築のための分散 Web 情報の一元化

渦尾秀勝 †　甲斐啓文 †　伊東栄典 *　廣川佐千男 *

† 九州大学大学院システム情報科学府
* 九州大学情報基盤センター
〒 812-8581 福岡市東区箱崎 6-10-1, (092)-642-4037
{h-uzuo@matu,itou@, hirokawa@,kai@}.cc.kyushu-u.ac.jp

Intra-school Web Information Integration for University Portal

Hidekatsu Uzuo†, Hirofumi Kai†, Eisuke Itoh*,　Sachio Hirokawa*

† Graduate School of Information Science and Electrical Engineering,
* Computing and Communications Center,
Kyushu University.

1　はじめに

近年、インターネットの普及によってユーザーが閲覧する Web ページが増加の一途をたどっている。著者らの所属する九州大学内部でも各部局ごとに Web ページを提供しており、Web ページが増加している。九州大学トップページの学部、大学院紹介ページだけでも、約 66 サイト、約 1200 ページ存在する。本論文は、Web ページ閲覧に便利なポータルサイトにおける学内情報の掲示手法の考察と実装について述べる。

膨大な量の Web ページには利用者にとっても、管理者にとっても問題点がいくつかある。Web 情報を利用する側（利用者側）と提供する側（提供者側）の観点から問題を述べる。

利用者側からの問題点は二つあり、一つ目は Web ページを巡回（お知らせなどの自分に必要な情報を閲覧）するのに手間がかかるということである。従来、大学事務からのお知らせは、掲示板に掲載されることがほとんどである。掲示板は、場所が物理的に固定されており、その場に行かなければ掲示板情報を見逃すことがある。近年、Web の発達により、事務情報を Web サイトに掲載されることも多くなっている。しかし、各部局ごとの情報は各部局 Web ページにしか掲載されないので、組織内のすべてのお知らせを知ろうとすると手間がかかってしまう。

二つ目は情報過多による閲覧性の低下である。総合大学大では学内サイトが、学部、学府、施設、サークルなど多数になる事がある。さらに各サイトに多数のページがあるため自分の必要な情報がどこにあるのか分かりにくく、探し出すのが困難である。

次に情報提供者側からの問題点を述べる。大学の情報提供者は事務員である事が多いが、情報技術の専門家とは限らない。ワープロを用いて印刷物作成は、日常的に行われているが、Web に掲示情報を載せるためには、Web の仕組みをある程度理解する必要がある。情報技術の非専門家は、この作業が困難である。リンク作業など Web ページを適切管理することは大変な労力と時間がかかる事も問題になる。常に新鮮な情報を提供するためには情報提供作業を専従して行う技術作業者が必要になる。一方、Web デザイナーなどの専門家へ Web ページ作成依頼も考えられる。しかし Web ページ作成には費用がかかる為、頻繁な依頼は困難である。

以上の問題点の解決方法として、利用者側からは一元的に情報を提供するポータルサイト、提供者側からは連絡情報を半自動的まとめるシステムが考えられる。ポータルサイトを作るためのシステムがいくつか存在する。例えば uPortal(`http://www.uportal.org/`) というシステムがあり、米国を中心に約 80 サイトで使われている。しかし、これらのシステムは分散した Web 情報を一元化する機能を持たない。

本研究では、各部局の Web サイトで分散して提供される「お知らせ」を収集および抽出するシステムの構築を目的とする。抽出した情報は、学内ポータルサイトの中で利用する。利用者はポータルサイトを見ることで各部局のお知らせを閲覧する事が可能になる。国内で総合ポータルサイトを持つ大学が３０大学あることは確認できている。 このように学内総合ポータルが増えていてその必要性が認識されるようになっ ている。しかし、その構築は人手によっていたり、更新の頻度が低かったり、解決しなければならない問題が多い。

2　Web 情報の一元化

2.1　Web からの情報抽出

ポータルサイトを構築するには二つの方法が考えられる。分散した Web 情報を統合して掲示する方法と複数の情報提供者に統一的な DB に登録してもらう方法である。今回は前者を採用した。

Web 情報を統合するためには、分散している Web 情報を抽出する事が必要である。Web サイトから有用な情報を抽出し統合する研究は様々ある [1][2]。例えば複数の商取引サイトから一つの商品についての価格、商品説明などを取得することが行われてきた [3]。

Web 情報抽出の一つの方法としては、各 Web サイトの HTML 構造を解析し、タグ情報をもとにそれぞれの抽出プログラムを作る方法がある。ただしこの方法は提供者側ページの HTML タグが一つ変更されるだけで、プログラムは誤動作を起こす可能性がある。

そこで、Web サイト間の処理をシームレスに繋ぐために Web サービスシステムが開発されている [4]。しかし、各部局の Web サーバを Web サービス対応にするにはシステムを入れ替えるがかかるため今回は考慮しない。

2.2 情報の提示

次に抽出および収集した情報の提示方法について考える。各部局の Web サイトから抽出すべき情報には二つの種類がある。一つは比較的長期にわたり固定された情報で、教員リスト、講義時間割、教室の位置、事務室への連絡方法などがある。これらは年度中に変更されることはあまりない。

もう一つは、お知らせやニュースのような短期的な情報である。お知らせは随時提供されるため、頻繁に更新される。そのため、ポータルサイト側でも、更新に即座に対応する事が必要である。

長期的な情報については、情報の存在する URL を人手で集め、テキストファイルなどで保存しておき、ポータルサイト側で提示してやればよい。URL の先のページが変更されたり、削除された場合にのみ対応すればよい。短期的な情報については即応性が重要であるため、RSS 形式で保存するのがよい。

RSS とは RDF Site Summary の略で、XML でサイトの概要を記述するフォーマットであり、Netescape 社が 1993 年 3 月に公開した [5][6]。最近ではニュースのヘッドラインを表現するのに使われる事が多い。図 1 に RSS の例を示す。この例は@IT（http://atmarkit.co.jp）で提供されている RSS 文書である。

図 1: RSS 文書の例

この RSS を利用する代表的なアプリケーションとして、複数のサイトの RSS 文書を読み込んで新着情報を横断的に表示する RSS リーダーがある。RSS リーダーには「SharpReader」[1]、「FeedReader[2]」、「NewsMonstar」[3]などがある。SharpReader の画面を図 2 に示す。これらを利用すると簡単に更新情報を得る事ができる。

[1]http://www.sharpreader.net/
[2]http://www.feedreader.com/
[3]http://www.newsmonster.org/

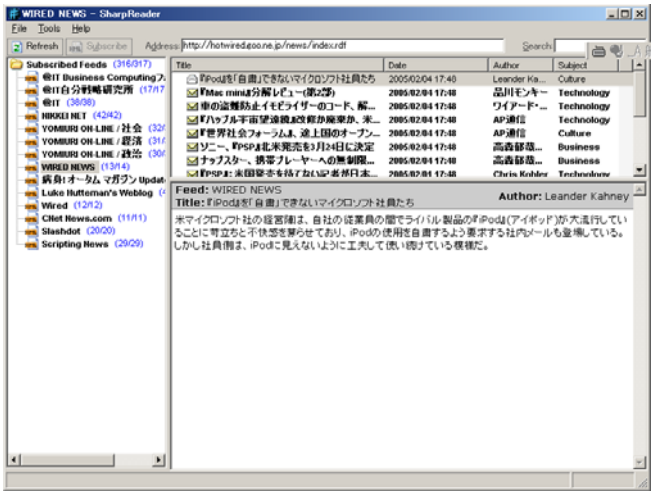

図 2: SharpReader の画面

本研究で作成したシステムでも、お知らせについては、更新日時、タイトル、Web ページの URL を元にし RSS を作成する。

3 試作システムの概要

本研究は、Web 上に公開されている九州大学各部局のお知らせ情報を自動収集する事を目標とする。紙に印刷されたお知らせから情報を取り込むのは容易ではないが、現在では多くのお知らせが Web 上で閲覧できる。よって、URL さえ分かれば比較的容易に収集することができる。本節では、本研究で作成したシステムについて述べる。

3.1 システム構成

本研究で構築したシステムは図 3 のように「Web ページ収集部」「RSS 変換部」「データ掲示部」の三つの部分から構成される。それぞれの動作を以下に示す。本システムは、プログラム言語は perl バージョン 5.8.5 を用いている。

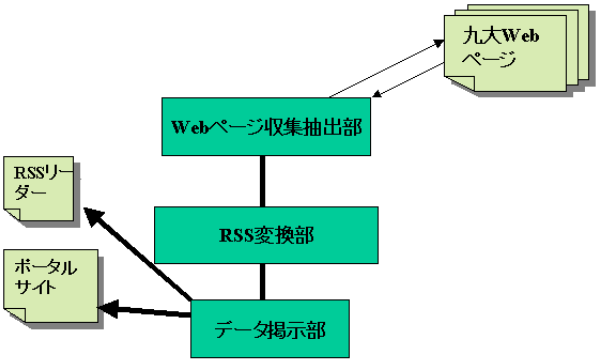

図 3: システム構成図

3.2 Web ページ収集、抽出部

Web ページ収集部では、Web 上に公開されている九州大学の Web ページを収集し、お知らせ情報の抽出を行う。収集する Web ページを特定する作業は今のところ手作業である。

この作業には、二通りの手法を用いて抽出した。以下に二つの手法について述べる。

3.2.1 手法１

手法 1 では、タグ名だけを用いている。ここでは, HTML::TokeParser モジュール [7] を用いた。このモジュールを使うと、HTML テキストを先頭から辿りながら、指定したタグにより特定されるコンテンツ部（テキスト部分や URL 部分）を抜き出す事が可能になる。

各部局サイトの「お知らせ」情報を抽出するために、以下の三つの作業を行った。

- 特定した Web ページを入手
- 取得しようとしているお「知らせ」情報に用いられているタグを特定
- 特定したタグを元に「お知らせ」情報を抽出するプログラムを作成

3.2.2 手法２

手法 2 は XPath を用いた収集、抽出である。XPath とは、元来 XML データ内の特定部位を指し示すための表示法である。XML データはタグ（要素名）や属性名、属性値の関係を木構造で表現することができる。HTML データも内部のタグコンテンツを木構造で表現することができ、XPath の表記方法で HTML ファイル内の特定部位を指定することができる。

例えば、HTML コンテンツを取得したい場合、/html/head/title と指定すると、html 要素をたどり、 head、 そして目的の title 要素を辿り、title 要素のテキスト部分を取得できる。を取得できる。

同様に各サイトのページのお知らせ部分を XPath で指定し、お知らせ情報を抽出する。この時に、XML::LibXML モジュール [7] を用いた。

3.3 RSS 変換部

抽出した情報から、タイトル、URL などの RSS 文書に必要な情報を抜き出し RSS 文書を作成する。RSS の生成には、XML::RSS モジュールを用いた。

3.4 データ掲示部

Web 収集抽出部、RSS 変換部でで収集したお知らせ情報を掲示する。収集した情報は、ポータルサイトとして、掲示するか、もしくは RSS に変換したものは RSS リーダーで閲覧することができる。他にも RSS データを HTML に変換する事でポータルサイトの Web ページとして表示することもできる。

4 実験および評価

九州大学公式サイトにある各学部大学院へのリンクページには、約 66 のリンクがある。それらのサイトを調べたところ、45 サイトのみお知らせ情報を発信していた。今回の実験では表 1 に示す五つのサイトについての、HTML ファイルを取得し、「お知らせ」の部分を特定するためのタグ情報、及び XPath を調査した。また特徴的なタグと XPath も表 1 に示す。

表 1: 実験対象サイト

部署名	手法 1:タグ	手法 2:XPath
九大トップ	$<$ $spanclass$ $=$ $"text12" >$	/html/body/table[2]/ tr[2]/td[2]/table[2]/ tr/td[1]/table/tr
文学部	特徴的タグなし	/html/body/div/div/ ul/li
21 世紀プログラム	特徴的タグなし	/html/body/div/div/ div[1]/ul/li
システム情報	特徴的タグなし	/html/body/table[1]/ tr[3]/td[2]/table[1]/tr
情報基盤センター	$< p >$	/html/body/table/tr[3]/ td[2]/table/tr[1]/td/table/ tr[2]/td/table/tr/td/p

4.1 実験結果

表 1 で示したサイトについて、手法１（第 3.2.1 節）および手法２（第 3.2.2 節）でのお知らせ取得を行った。その結果を表 2 に示す。

表 2: 実験結果

部署名	手法 1(第 3.2.1 節)	手法 2(第 3.2.2 節)
九大トップ	10/10	10/10
文学部	-	3/3
２１世紀プログラム	-	6/6
システム情報	-	0/5
情報基盤センター	6/5	0/5

実際に取得できたお知らせ個数/実際のお知らせ個数

4.2 評価

表 2 の結果について評価する。

- 九大トップページ、文学部のサイトでは問題なくお知らせを抽出する事ができた。
- 21 世紀プログラムのサイトにおいては、手法１（タグを用いたもの）では抽出できなかった。その理由は、ページを記述する HTML 構造において, 「お知らせ」情報を特定するような特徴的なタグが存在しないためである。手法２では問題なくお知らせ情報を抽出できた。
- システム情報科学府のサイトは、21 世紀プログラム同様特徴的なタグがないために手法１では抽出できなかった。

- 情報基盤センターのサイトでは、手法１では一つだけ余計な情報を抽出した。それは「過去のお知らせ」というリンクのついたコンテンツである。これは、余計な一つの情報にもお知らせ情報と同じタグを持つためである。

システム情報科学府、情報基盤センターでは、手法２による取得ができなかった。その理由は日本語文字コードの処理の不具合であると思われる。手法２の XPath 指定による情報抽出では、Perl5.3 の LibXML モジュールを使用している。このモジュールが、ある程度の日本語文字コードに対応してないようで、プログラムが動作しなかった。

4.3 RSS への変換

抽出したお知らせ情報を RSS 形式のファイルとして格納した。作成したファイルを RSS リーダーで読み込み、各サイトのお知らせ情報を表示させた。図 4 に RSS リーダーでの表示画面を示す。

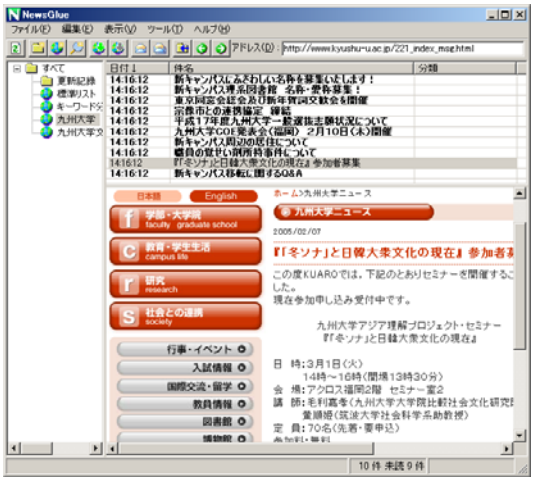

図 4: RSS リーダー画面による表示

5 終わりに

本研究では、学内ポータルサイト構築のために、学内に分散したお知らせ情報を収集、抽出するシステムについて検討した。

お知らせ情報の収集、抽出については２つの手法を考えた。一つ目の手法は特徴的な HTML タグを指定する方法である。二つ目の手法は XPath によるお知らせ情報の位置の指定である。これらの手法を九州大学内の５つのサイトに適用しお知らせ情報を抽出する実験を行った。その結果、お知らせを抽出しやすい構造をもつサイトや抽出しにくい構造をもつサイトが存在する事が分かった。また抽出したお知らせ情報の RSS 形式への変換も行った。

既に様々な大学ポータルが稼働している。実際、「オキナビ －大学ポータルサイト－」(http://oki-nabi.com/zenkoku.html) には、大学ポータルサイトへのリンクが約 30 存在する。九州大学でも大学ポータルへの要求が高まっており、それらの構築に向けての準備を進めている。

総合大学のように、複数の部局が存在する組織では、情報は分散管理せざるを得ない。ポータルサイトへ情報を統合するために、本論文では各部局のサイトから情報を抽出する方法を検討した。

一方、各部局側でお知らせを提供するとともに、そのお知らせについての RSS ファイルを用意する方法も考えられる。IBM 社の「ホームページビルダー」や、Adobe 社 (旧マクロメディア社) の「Web Publishing System」には RSS ファイル生成機能が備わっている。これらを用いることで、今回検討した内容は比較的簡単に実現できる可能性はある。

ポータルサイトは様々な情報システムへの入口となることが期待されているため、お知らせ情報を提供するだけでなく、認証機能や、電子メールなども必要である。Japan OCW[8] に見られるように、今後は各大学における e-Learning[9] の充実が重要になる。Japan OCW では、大阪大学、京都大学、慶應義塾大学、東京工業大学、東京大学、早稲田大学の 7 大学が、講義資料の公開を行なっている。各大学でも e-Learning システムの構築が進んでおり、大学ポータルサイトは、e-Learning 教材提供システムとの連携が必要であろう。この様に実際の大学ポータルサイト構築には、様々な機能の充実が必要となる。

今後はより多くのサイトからお知らせ情報を抽出するようにシステムを拡張する予定である。また実際に、学内ポータルサイトを作成する予定である。

参考文献

[1] 伊東栄典, 竇ギョク峰, 廣川佐千男：“Web シラバス群のデータ形式統合に関する考察”, 情報処理学会 マルチメディア, 分散, 協調とモバイル (DICOMO 2004) シンポジウム論文集, pp.345-348, 2004.

[2] Y. Matsunaga, S. Yamada, E. Ito, S. Hirokawa : “*A Web Syllabus Crawler and its Efficiency Evaluation*”, International Symposium on Information Science and Electrical Engineering 2003 (ISEE 2003), pp.565-568, 2003.

[3] 坂本比呂志, 有村博紀：“Web マイニング”, 人工知能学会誌, 特集「テキストマイニング」, Vol.16, No.2, pp.233-238, 2001.

[4] 本俊也：“最新 Web サービス”, 秀和システム社,2004.

[5] RDF Site Summary(RSS)1.0, `http://web.resource.org/rss/1.0/spec`

[6] RSS–サイト情報の要約と公開, `http://www.kanzaki.com/docs/sw/rss.html`

[7] CPAN : `http://www.cpan.org/`

[8] Japan OCW : `http://www.jocw.jp/` , 2005.

[9] 情報処理振興事業協会, 先端学習基盤協会：“*e* ラーニング白書 *2002/2003* 年版”, オーム社, 2002. (ISBN4-274-06480-8)