HTML2RSS: Automatic Generation of RSS Feed based on Structure Analysis of HTML Document

Tomoyuki Nanno
Interdisciplinary Graduate School of Science and
Engineering, Tokyo Institute of Technology.
4259 Nagatsuta-cho Midori-ku
Yokohama Kanagawa JAPAN 226–8503
nanno@lr.pi.titech.ac.jp

Manabu Okumura
Precision and Intelligence Laboratory,
Tokyo Institute of Technology.
4259 Nagatsuta-cho Midori-ku
Yokohama Kanagawa JAPAN 226–8503
oku@pi.titech.ac.jp

ABSTRACT

We present a system to automatically generate RSS feeds from HTML documents that consist of time-series items with date expressions, e.g., archives of weblogs, BBSs, chats, mailing lists, site update descriptions, and event announcements. Our system extracts date expressions, performs structure analysis of a HTML document, and detects or generates titles from the document.

Categories and Subject Descriptors

H.5.4 [**Information Systems**]: Hypertext/Hypermedia; H.3.5 [**Information Systems**]: Online Information Services

General Terms

Management

Keywords

RSS, Atom, feed, document analysis, syndication

1. INTRODUCTION

Almost all the information on the World Wide Web is unstructured [1], and as such, its layout description languages are hard to analyze automatically. In recent years, there has been a lot of effort to find ways to provide structured information, such as the Semantic Web and web services such as Google APIs and Amazon Web Services. For computers to use the web effectively, it is thought that the use of metadata will be inevitable.

RSS (RDF Site Summary / Really Simple Syndication) in particular has received a lot of attention. For example, RSS and/or Atom feed are used for metadata distribution in almost all weblogs. Since RSS is designed for distributing update information on a web site (e.g. title, description and last update), many web sites other than weblogs, such as news web sites or BBSs (Bulletin Board System), are using it for update notification.

Many applications to handle RSS feeds have been developed. For example, by using an RSS reader application we can easily track hundreds of web sites every day because we can browse at a glance the updated information on many web sites.

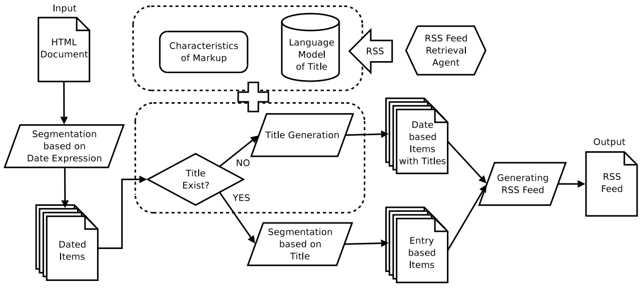

Figure 1: Flowchart of our system

However, the web sites that can be subscribed to with an RSS reader are restricted to web sites that have been built up with CMSs (Content Management Systems) such as blogging softwares, because if an author builds a web site with CMS, there are no additional costs to generate the RSS feed; otherwise the author has to generate the RSS feed manually and update it accordingly. Another problem is that we cannot read past contents via RSS feeds, because the feeds usually include only the n-most recently published contents [2].

To solve these problems, we construct a system to automatically generate RSS feeds from HTML documents that consist of time-series items with date expressions, e.g., archives of weblogs, BBSs, chats, and mailing lists, update descriptions on a site page, and announcements of events. Our system extracts date expressions, analyzes the structure of a HTML document, and detects/generates titles from the document.

By using our system, a user can read any web pages with an RSS reader even when they provide no RSS feed, and authors of web pages that have no RSS feed can easily make them "RSS-Autodiscovery" ready. Moreover, our system can be applied to an archive page of time-series items, enabling a user to easily get a "back-issue" RSS feed.

2. OVERVIEW OF OUR SYSTEM

Figure 1 shows the flowchart of our system. To generate the appropriate RSS feeds, we need to extract the shadowed

WWW 2006, May 22–26, 2006, Edinburgh, Scotland.
ACM 1-59593-332-9/06/0005.

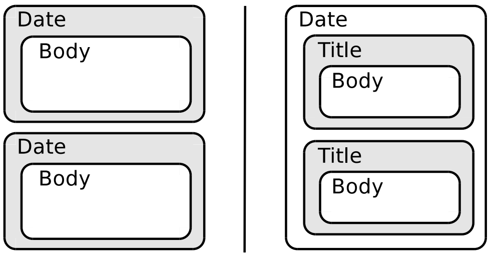

Figure 2: The structure to be extracted

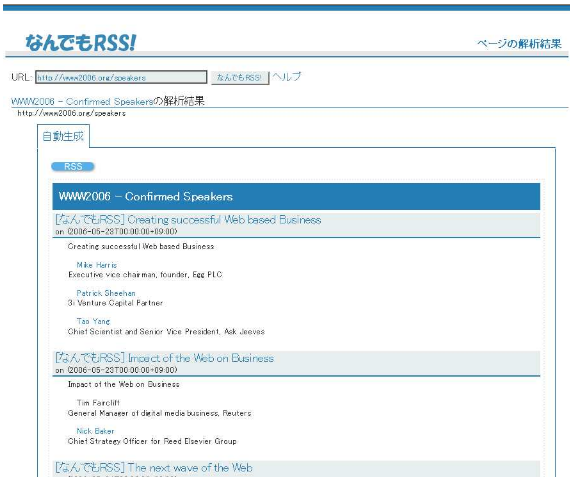
なんでもRSS!
ページの解析結果
URL: http://www2006.org/speakers なんでもRSS! ヘルプ
WWW2006 – Confirmed Speakersの解析結果
http://www2006.org/speakers
自動生成
RSS
WWW2006 – Confirmed Speakers
[なんでもRSS] Creating successful Web based Business
on (2006-05-23T00:00:00+09:00)
Creating successful Web based Business
Mike Harris
Executive vice chairman, founder, Egg PLC
Patrick Sheehan
3i Venture Capital Partner
Tao Yang
Chief Scientist and Senior Vice President, Ask Jeeves
[なんでもRSS] Impact of the Web on Business
on (2006-05-23T00:00:00+09:00)
Impact of the Web on Business
Tim Faircliff
General Manager of digital media business, Reuters
Nick Baker
Chief Strategy Officer for Reed Elsevier Group
[なんでもRSS] The next wave of the Web

Figure 3: Output for "Confirmed Speakers" in WWW2006 site

boxes in Figure 2 as items. Therefore, our system first segments a web page into dated items by detecting and using date expressions. After that, our system tries to detect titles in each item. If each dated item includes more than one title, our system segments it further by using the titles, otherwise it generates a title for each dated item, because the RSS feed needs a title for each item. To detect/generate the title, our system uses many RSS feeds that have been collected from the web to construct a language model of title expressions.

3. HTML2RSS: RSS FEED GENERATOR

Here we explain our web application "RSS Feed Generator[1]", which is freely available. We released the system on 9th May 2005, and had 11.73 million accesses since then (as of the end-October 2005). Lately our system generates RSS feeds for 11 thousand different URLs and delivers more than 90 thousand feeds a day.

In the "Confirmed Business Track Speakers" section of "WWW2006: Confirmed Speakers" page[2], the speakers are described with date expressions and session titles. Figure 3 shows the output of our system for this page [3]. Our system found these date expressions and session titles automatically, and extracted the appropriate part of the HTML document that each date expression and title annotate. It also visualized all the possible RSS feed patterns so that a user could select the one which she wanted, because a web page can contain multiple time-series items.

If the user clicked the link "Subscribe This RSS Feeds", she would obtain the RSS feeds and thereby track the schedule/speaker change by combining our system and the RSS Reader together.

The author of a page can also use the system in another way. When she wants to make her page "RSS-Autodiscovery" ready, she has only to copy the URL of the link "Subscribe to This RSS feed" and pastes it into the document header. By doing so, when the author updates the "Schedule" information, she needs no additional work to maintain the RSS feed, because the system re-generates the feed automatically.

These results showed user's demands for tracking much information on the web via an RSS reader even though it does not provide its content with RSS. RSS has become popular day by day, though current users might not be ordinary people.

4. CONCLUSION AND FUTURE WORK

We presented a system that automatically generates RSS feeds from HTML documents. Our method is based on HTML document structure analysis and detection of date expressions and titles.

Although metadata distribution by RSS feed has become popular, the web sites that can be subscribed to with RSS reader are restricted to a fraction of web sites built up with CMS, because if an author builds a web site with CMS, there are no additional costs to generate the RSS feed; otherwise the author has to generate the feed manually and update it accordingly.

The preliminary results show that our system works well. We obtained suitable RSS feeds from many non-RSS-ready web pages and users of our system are increasing. However, we consider that quantitative evaluations for title detection/generation are needed.

As future work, we plan to automatically annotate other information to RSS feeds, such as feed type, i.e., blog feed, BBS feed, etc., or the item category by using machine learning. We are also developing a search engine, called "blogWatcher[4]", which can search time-series information collected by this method [3].

5. ACKNOWLEDGMENTS

This work was supported by The 21st Century COE Program, "Framework for Systematization and Application of Large-scale Knowledge Resources", of the Japan Society for the Promotion of Science.

6. REFERENCES

[1] M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engine. In *Proc. International Joint Conference on Artificial Intelligence*, pages 1573–1579, 2003.

[2] D. R. Karger and D. Quan. What would it mean to blog on the semantic web? In *Proc. Third International Semantic Web Conference*, pages 214–228, 2004.

[3] T. Nanno, Y. Suzuki, T. Fujiki, and M. Okumura. Automatic collection and monitoring of japanese weblogs. In *WWW2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2004.

[1] http://blogwatcher.pi.titech.ac.jp/nandemorss/

[2] http://www2006.org/speakers/

[3] http://blogwatcher.pi.titech.ac.jp/nandemorss/index.cgi?url=http://www2006.org/speakers

[4] http://blogwatcher.pi.titech.ac.jp/