การพัฒนาอัลกอริทึมสำหรับจัดการข้อมูลจากอาร์เอสเอส

Development the Algorithm for archiving and retrieving RSS feeds

วฤษาย์ ร่มสายหยุด
ภาควิชาวิศวกรรมคอมพิวเตอร์, คณะวิศวกรรมศาสตร์
มหาวิทยาลัยสยาม เลขที่ 235 ถนนเพชรเกษม เขตภาษีเจริญ
กรุงเทพมหานคร 10163
โทร (+662) 457-0068 ต่อ 210, โทรสาร. (+662) 457-3982
Walisa.r@siamu.ac.th

Abstract: Most people are interested in many websites whose content changes on an unpredictable schedule. Examples of such websites are news sites, community and religious organization information pages. The user would repeatedly checking each website to see if there is any new content, many web sites/pages are provide what are called RSS feeds, which are descriptions of the new data is a lightweight XML format but they are not able to retrieve information that was updated during the time period when the reader is offline. In this paper, I proposal an algorithm to support retrieving, archiving, and synchronize RSS feeds for the updated information through a network.

Key Words: RSS, Feed Reader

Introduction

RSS is a Web content syndication format. Its name is an acronym for ***R**eally **S**imple **S**yndication*. All RSS files must conform to the XML 1.0 specification, as published by the World Wide Web Consortium (W3C)[1]. RSS is used to provide items containing short descriptions of web content together with a link to the full version of the content. This information is delivered as an XML file called RSS feed, RSS stream, or RSS channel.

RSS feeds offer different kinds of news in specific channels. People interested in specific news subscribe such channels using feed readers that look for new contributions in these channels so that subscribers can read these new contributions immediately after the update of the items in the channels.

An RSS feed is written in XML. A feed comprises a channel, which has a title, description, etc. followed by a series of items; a sample channel looks something like this:

```
<?xml version="1.0"?>
<rss version="2.0">
  <channel>
    <title>Liftoff News</title>
    <link>http://liftoff.msfc.nasa.gov/</link>
    <description>Liftoff to Space
Exploration.</description>
```

```
   <language>en-us</language>
   <pubDate>Tue, 10 Jun 2003 04:00:00
GMT</pubDate>
   <lastBuildDate>Tue, 10 Jun 2003 09:41:01
GMT</lastBuildDate>

<docs>http://blogs.law.harvard.edu/tech/rss</do
cs>
   <generator>Weblog Editor 2.0</generator>

<managingEditor>editor@example.com</managi
ngEditor>

<webMaster>webmaster@example.com</webMa
ster>

   <item>
     <title>Star City</title>

<link>http://liftoff.msfc.nasa.gov/news/2003/news-
starcity.asp</link>
     <description>How do Americans get ready to
work with Russians aboard the
       International Space Station? They take a
crash course in culture, language
       and protocol at Russia's Star
City.</description>
     <pubDate>Tue, 03 Jun 2003 09:39:21
GMT</pubDate>

<guid>http://liftoff.msfc.nasa.gov/2003/06/03.html
#item573</guid>
   </item>
   <item>
     <title>Space Exploration</title>
     <link>http://liftoff.msfc.nasa.gov/</link>
     <description>Sky watchers in Europe, Asia,
and parts of Alaska and Canada
       will experience a partial eclipse of the Sun on
Saturday, May 31st.</description>
     <pubDate>Fri, 30 May 2003 11:06:42
GMT</pubDate>

<guid>http://liftoff.msfc.nasa.gov/2003/05/30.html
#item572</guid>
   </item>
   </channel>
</rss>
```

A RSS reader (RSS-aware program or aggregator) is a program that reads RSS feeds and turns the XML data into a format that is meaningful to the application at the client's site [10]. A reader has a local cache of data and works much like an email client.

A main idea with RSS readers is that they can only retrieve information that a site is making available. This means that if the reader is offline for an extended period of time, it may miss some update data. This presents a need to archive items for future use. In this paper, I present an algorithm for archiving and retrieving RSS feeds. This algorithm keep one set of update data and make it available to many different clients, all of whom have subscribed to the same RSS feeds. A

server should also be able to connect to other servers to get update information, requiring there to be only one set of data for each feed across the entire network of servers.

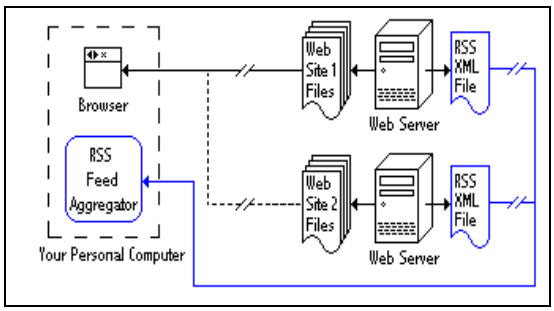

Figure 1: The RSS feed XML Files.

This figure is showing how the websites, the RSS feed XML files and a personal computer are connected by a web browser being used to read first Web Site 1 over the Internet and then Web Site 2. It also shows the RSS feed XML files for both websites being monitored simultaneously by an RSS Feed Aggregator.

The remainder of the paper is structured as follows: in section two, I present a BARF. Section three focuses on the proposed an algorithm. The paper ends with a conclusion and future work.

The Study

They present a distributed platform for archiving and retrieving RSS feeds. They call it BARF (Braf Archives RSS Feeds) server.

The BARF server consists of several components, called managers. Each of these managers is responsible for carrying out a task with minimal assistance from the others, providing a modular design that can be easily expanded upon. The overall architecture of the server managers is shown in Figure 2.

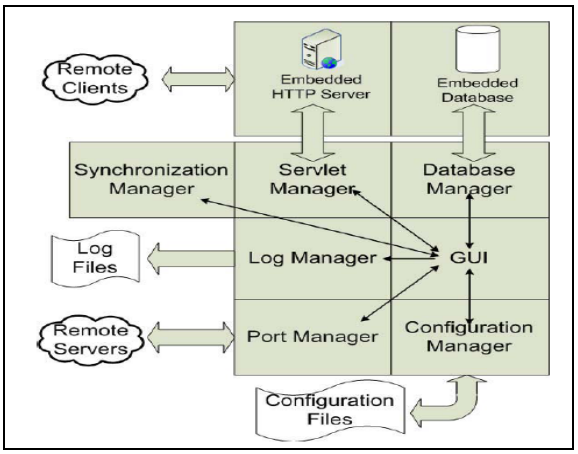

Figure 2: Components of the BARF server.

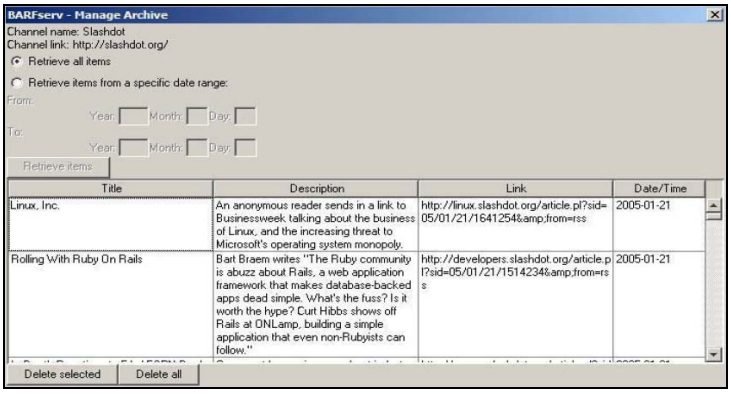

Figure 3: Feed Management Dialog.

BARF server has a rich set of options for administrating the server in a variety of different environments, provided in a graphic user

interface. The two major options are Action and Configuration is shown in Figure 3.

Our Algorithm for the RSS Reader

I define the following in different modules for our algorithm as following:

Phase 1: Add_URL

- Create a RSSHandler object.
- Get the channel.
- Parse the RSS feed at a given URL using the RSSHandler object.
- Add a server to the current server list.

Phase 2: Add_Server(on current server list)

- Create a portManager and a serverPacket.
- Set up attributes (address, port, return Channels, etc.) for the request.
- If the server already exists, return.
- Otherwise, send a request to the server to get information.
- Add the server and its channels to the database.
- Configuration for the server.

Phase 3: Sync_RemoteServer

- Synchronize all channels if channel id is 0 (recursively)
- Get information (name, address, port, etc.)
- Poll the server for information if it's up.
- Update the database with remote channel information.

Phase 4: Sync_URL

- Get some information about the channel we're to synchronize.
- Get server name and id.
- If this is a locally cached feed, connect to the remote site to get information about the feed's items.
- Update the channel with the new item information.

Conclusions

This paper has proposed an approach for a developer in the development of RSS Reader system base extend from BARF server.

I plan to prove the performance of this algorithm with popular RSS Reader: Firefox, Flock, and Firefox2.

Future work shall concentrate on implementation of this algorithm as open source software.

References

[1] Peter J. A. Ewusch, Bastian Stoll, Torsten Schulwandt, Pawel Serwatowski,"New Communication Concepts Based upon Advanced RSS Feeds", IEEE Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Application, 2005.

[2] Mortbay Consulting Ptr. Ltd. (2004). "Jetty HTTP Server 5.1.2," http://jetty.mortbay.org/javadoc/index.html

[3] My Netscape Network, http://www.purplepages.ie/RSS/netscape/rss0.90.html.

[4] Pilgrim, M. (2002). "What is RSS?" http://www.xml.com/pub/a/2002/12/18/dive-into-xml.html

[5] Quadcap Software. (2004). "Quadcap Embeddable Database," http://www.quadcap.com/products/qed/docs/index.html

[6] Resource Description Framework, http://www.w3c.org/RDF

[7] RSS Advisory Board. (2005). "RSS 2.0 Specification: RSS at Harvard Law," http://blogs.law.harvard.edu/tech/rss

[8] RSS Info: News and Information on the RSS Format, http://blogspace.com/rss/resources

[9] Stealthp.org. (2004). "RSSLib4J: rsslib4jRSS Parser API Documentation," http://rsslib4j.sourceforge.net/javadoc

[10] David Chmielewski and Gongzhu Hu, "A Distributed Platform for Archiving and Retrieving RSS Feeds",In Proc. Of the Fourth Annual ACIS International Conference on Computer and Information Science (ICIS'05), 2005