ClipperRSS: A Light-Weight Prototype for the Cross-language Exploitation of Syndicated Feeds

Roderick Holland
Information Technology Center
The MITRE Corporation
Bedford, Massachusetts 01730, USA
`rholland@mitre.org`

Brenden Keyes
Information Technology Center
The MITRE Corporation
Bedford, Massachusetts 01730, USA
`bkeyes@mitre.org`

Abstract

Syndicated feeds in RSS, Atom, and related formats have emerged as ubiquitous information sources in World Wide Web language communities including Arabic, Farsi, Chinese, and others, providing subscribers with timely updates on topics of particular interest. We have modified an existing Open Source RSS reader, Sage, for cross-language use, permitting English-speakers to discover, subscribe to, update, and browse RSS feeds in ten languages. This early prototype, called ClipperRSS, has been integrated with the Clipper cross-language information retrieval tool. The integrated system provides English-speakers with an effective means of exploring the potential of foreign-language syndicated feeds in their domains of interest.

1 Syndicated Feeds and the Polyglot Web

RSS ("Really Simple Syndication") and related XML data distribution formats (Atom, etc.) have developed over the last decade[1] as change notification and summarization mechanisms on the World Wide Web.

Most often associated with news websites and blogs, an RSS feed provides timely distribution of updates to the site it is associated with through a publish-subscribe mechanism. Users can discover, subscribe, and read these feeds using a variety of RSS readers and aggregators.

RSS feeds consist of a series of items, each item typically consisting of a title, a link to an HTML page, a text description, and a date-time stamp. A user with an RSS reader can request updates on the feed, and receive the n-most-current items. At the user's discretion, each item's link may be followed to the associated HTML page.

We were made aware of the potential need for cross-language support for RSS feeds by Jason Bruzdzinski[2], who pointed out their frequent use on the Chinese Web. Subsequent investigation revealed significant—and at times, surprising[3]—RSS feeds in use in association with Web sites in Chinese, Arabic, Farsi, Bahasa Indonesia, Russian, Spanish, and a range of other languages, in domains including military science, terrorism, technology, public health, commerce, etc.

The most important aspect of syndicated feeds is that they provide a convenient notification mechanism for changes to sites of interest. This is a significant contribution to the currency of information derived from Web sources.

2 The ClipperRSS Prototype

Of the many possible approaches to cross-language exploitation of RSS feeds, we chose what may well

---

[1] Useful treatments of the RSS specification and its history can be found at http://www.rss-specifications.com/ and http://cyber.law.harvard.edu/rss/rss.html.

[2] Bruzdzinski, Jason. Private conversation, January 26, 2006.

[3] For example, a frequently-updated Arabic blog self-attributed to The Al Qaeda Organization in the Arabian Peninsula (http://amrallam.maktoobblog.com/), with an associated RSS feed (http://www.maktoobblog.com/amrallam/rss.xml).

be the simplest. We modified an existing Open Source RSS reader, Sage[4]. Sage runs as an "add-on" to the Firefox browser.

Modification was straight-forward: we examined each portion of the Sage user interface, and transformed it for cross-language use by making calls to a web service abstraction of machine translation, OpenMT[5], adjusting the user interface for cross-language use, as necessary.

A screenshot of a ClipperRSS session[6] can be seen in Figure 1.

Figure 1 ClipperRSS prototype.

2.1 Display of an RSS Feed

In Figure 1, the central area of the display shows one Chinese language feed. The Chinese feed title

and description is shown at the top, followed by the English translation of the feed title ("Avian flu") and a translation of its description. Both versions of the feed title are hyperlinks to the HTML page referenced by the feed; clicking on either will launch Clipper[7] on that page.

Immediately following is a two-column display of the five most recent feed items (a value set by the server). For each numbered feed item, the original (Chinese) title, description, and date/time stamp are shown in the left column, with a machine translated English version of the same information shown in the facing entry on the right column. Each item title, in either the original or translated version, is a hyperlink to the HTML page referenced by that item; clicking on either will launch Clipper on that page.

Display of a feed is activated by a mouse-click on one of the list of subscribed-to feeds in the center pane on the left edge of the display. In the example, the "Avian flu" feed is seen to be selected.

2.2 Launching Clipper on a Title

As noted, both feed titles and item titles are hyperlinks that serve as launch-points for Clipper, a cross-language information retrieval tool. The act of clicking on either sort of title, in either the original or translated form, will launch Clipper on the associated webpage. An example of the resulting display is shown in Figure 2.

Figure 2 Clipper display of an HTML page launched from within ClipperRSS.

---

[4] http://sage.mozdev.org/; the Release 1.3.10 source archive was used.

[5] OpenMT is a simple web services abstraction of machine translation services in the REST idiom. The salient feature of OpenMT is that it can be invoked through HTTP POST and GET operations, making it highly suitable for embedding machine translation in existing applications.

[6] View of a Chinese-language RSS feed on avian influenza (http://www.who.int/feeds/entity/csr/disease/avian_influenza/zh/rss.xml) associated with a World Health Organization web page (http://www.who.int/csr/disease/avian_influenza/zh/). Translations in this example were performed with the Language Weaver 5 Chinese-English machine translation engine.

[7] Clipper is a cross-language information retrieval (CLIR) prototype tool previously developed by the MITRE team. Clipper is essentially a flexible interface for integrating commercial machine translation engines with existing Internet search engines with specific language indexes. CLIR is accomplished by query translation, using machine translation backed by a user dictionary.

The integration of ClipperRSS with Clipper is accomplished by passing the desired URL and language parameters as a switch to the Clipper launch URL; this brings up Clipper with the proper page. In the Clipper two-column display, the original (Chinese) page is shown in the left column, with a machine-translated English counterpart shown in the right column. Links may be followed within either the original or translated views, resulting in a similar bi-lingual display of the linked material.

In the bottom-left pane is a separate list of the titles, in the original language (Chinese), for items in the current feed. These can be used as an alternative set of launch points, once the user has begun to use Clipper to explore pages addressed by the feed. An English translation of a item title in this pane is shown when the user drags the mouse cursor over the item title (not shown).

2.3 Feed Discovery, Subscription, and Updating

Before viewing feeds, it is first necessary to discover and subscribe to them. The Discover Feeds button (shown as a Magnifying Glass icon) can be used to discover any RSS feeds associated with the currently displayed page. As modified in Clipper RSS, it will provide English translations of the feed names. This is shown in Figure 3. Once feeds are discovered, they may be selected and subscribed to using the Add Feed button.

Once feeds have been subscribed to, the user can obtain notifications of updates to the feeds by using the Check Feeds button (shown as a rotation icon, immediately to the left of the Discover Feeds button). Each feed to which the user has subscribed will be checked for updates. Feeds with unread updates will be indicated in bold-face type in the feeds pane.

Figure 3 ClipperRSS feed discovery.

As a practical matter, web pages with associated RSS feeds can often be found with Clipper queries of the form

Topic AND RSS

in the target language. This is illustrated by Figure 4. In this case, the query

"avian influenza" AND RSS

was used against Chinese sources with the Baidu search engine.

Figure 4 Discovering pages with RSS feeds using Clipper.

3 Conclusions and Next Steps

Though extremely simple, the ClipperRSS prototype has allowed the MITRE development team, and others, to explore the potential of RSS feeds in

multiple language communities and topical domains. We are convinced that this represents a promising new source for those interested in timely notification of new information from regional Web sources.

In addition to a list of pending technical improvements to the ClipperRSS prototype itself—automatic language identification, better formatting of translated item descriptions with complex content, etc.—we see the opportunity to develop more sophisticated systems for cross-language RSS exploitation. A logical next step would be the construction of an RSS aggregation service with cross-language capabilities. Such a system would permit tailored user profiles to be used to deliver topically precise notifications from a large number of feeds, in multiple languages.