Archival of Articles via RSS and Datamining Performed on Stored Articles

Caroline Bauer
2004-2005
Period 1
Computer Systems Lab

Abstract

RSS is a web syndication protocol used by many weblogs and news websites to distribute information; it saves people having to visit several sites repeatedly to check for new content. At this point in time there are many RSS newsfeed aggregators available to the public, but none of them perform any sort of archival of information beyond archival of the RSS metadata. As the articles linked may move or be removed from the internet at some time in the future, if one wants to be sure one can access them in the future one has to archive them oneself; furthermore, should one want to link such collected articles together (based on subject matter, for instance), it is far easier to do if one has them archived. The purpose of this project is to create an RSS aggregator that will archive the text of the actual articles linked to in the RSS feeds in some kind of linkable, searchable database, and implement some sort of datamining capability as well.

Process

In order to do this, I researched RSS, databases, and data mining. I also picked up perl along the way to use, as it had been pointed out to me that perl would be perfect for what I was trying to do. I found a perl module, XML::RSS, that made what I wanted to do a great deal easier; it made it possible to easily extract just the link (or any other specific field) from the RSS data. I modified a perl script to collect the links from a command-line argument RSS feed URL, then get the text from the articles and save it as a flat text file on the hard drive of the machine. I was not able to pick up enough SQL in time to have the script put the articles into a database, but if I knew how it would be a very minor modification to the script. Datamining has also not yet been implemented, and there are various things I would like to make to the code, such as parsing the HTML tags out of the data; at this point it is a relatively simple script.

Background

RSS stands for Really Simple Syndication and/or Rich Site Summary (it depends who you ask), a syndication protocol often used by weblogs and news sites. Technically, RSS is an XML-based communication standard that encompasses Rich Site Summary (RSS 0.9x and RSS 2.0) and RDF Site Summary (RSS 0.9 and 1.0). It enables people to gather new information by using an RSS aggregator (or "feed reader") to poll RSS-enabled sites for new information, so the user does not have to manually check each site. RSS aggregators are often extensions of browsers or email programs, or standalone programs; alternately, they can be web-based, so the user can view their "feeds" from any computer with Web access.

Data mining is the searching out of information based on patterns present in large amounts of data.

Results/Conclusion

At this point in time the script is in a far more rudimentary form that I had hoped to accomplish this year. This is a project I started because I wanted a system for archival, so I will continue this project after the school year is up, but I had hoped to get more done this year. At this point what it does is relatively limited—it simply archives as flat text files the articles linked in the RSS data. I still need to modify it to parse out the HTML (or other) tags, put the article text into a database, put together a simpler interface for the user(s), and then work on datamining. I had hoped to get some datamining done this year but it has seemed a very daunting task. I did not know perl at the beginning of this project, and I still am not very good with it, so I think that had I known perl better at the beginning of this project I would have gotten a lot further during the school year.