Archival of Articles via RSS, and Datamining Performed on Stored Articles

Car Bauer

January 18, 2005

Abstract

RSS (Really Simple Syndication, encompassing Rich Site Summary and RDF Site Summary) is a web syndication protocol used by many blogs 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 the RSS metadata. As the articles linked may move or be eliminated 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, 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, if all goes well, implement some sort of datamining capability as well.

1 Introduction

This paper is intended to be a detailed summary of all of the author's findings regarding the archival of articles in a linkable, searchable database via RSS.

1.1 Background

1.1.1 RSS

RSS stands for Really Simple Syndication, 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.

1.1.2 Archival Options Available in Existing RSS Aggregators

1.1.3 Data Mining

Data mining is the searching out of information based on patterns present in large amounts of data. //more will be here.

1.2 Purpose

The purpose of this project is to create an RSS aggregator that, in addition to serving as a feed reader, obtains the text of the documents linked in the RSS feeds and places it into a database that is both searchable and linkable. In addition to this, the database is intended to reach an implementation wherein it performs some manner of data mining on the information contained therein; the specifics on this have yet to be determined.

1.3 Scope

2 Development

3 Results

4 Conclusions

5 Summary

6 References

1. "RSS (protocol)." Wikipedia. 8 Jan. 2005. 11 Jan. 2005 <http://en.wikipedia.org/wiki/RSS_%28protocol%29>.

2. "Data mining." Wikipedia. 7 Jan. 2005. 12 Jan. 2005. <http://en.wikipedia.org/wiki/Data_mining>.