RSS feed, XML, HTML,
news syndication

Katarzyna WEGRZYN-WOLSKA*

PUBLISHING AND SYNDICATION INFORMATION ACROSS THE EDUCATION'S SITES USING THE RSS FEED

SHORT NOTE

With the increasing number of demand for access to education, we need the new technologies to facilitate learning. The good opportunity is to use the Web, which is an enormous and unlimited source of useful and varied kinds of information. Online learning and teaching is becoming more and more popular. Besides the teaching and learning techniques the education environments also consists a communication tool which allows easy publications and syndications the information (news) across the different education Web sites. This paper presents some statistics related to news syndication and the data format used by publishers for news publication. We describe two of the most popular formats currently used to publish and retrieve news information; the HTML non-structured data format and the XML-based RSS-feed format.

1. INTRODUCTION

This paper describes some general problems of publishing, retrieving and filtering news information across the educations Web sites. It introduces the principle of the news publication, explains the main problems of exploration and filtering data from the Web pages and the problems of the content syndication on the Web sites. Moreover, the RSS feed data format, as well as the possibilities and advantages of this format for publishing and syndication the information across the Web sites are presented. The conclusion summarizes the problems and solutions presented.

2. E-LEARNING ENVIRONMENT

Online teaching and learning provide great opportunities in terms of availability of information and resources, synchronous and asynchronous communication via the Web. This communication by using a discussion board, chat and course by e-mail is a priority in an online learning environment. In general, a didactical E-learning environment firstly consists of the students and teachers with whom the students' can interact during the learning process [4],[5]. But we cannot forget that it is also very important in didactical environment to allow teachers and students access in the very simply way to the most important necessary fresh information like news. This information could be published in the format, which allows reusing the data and syndication automatically on different educational information sites.

* Ecole Supérieure d'Ingenieurs en Informatique et Génie de Télécommunication, 77-215 Avon Fontainebleau, France

3. NEWS PUBLISHED ON THE WEB

There are a lot of Web sites, which publish news. The news' sites publish different kinds of information in different presentation forms [3], [11]. News is a very dynamic kind of information, permanently updated. In Table 1 updating frequency for some Web news sites is presented. This information is provided and confirmed by the sites administrators. Some parameters are significant for automatic treatment, like frequency of data updating, size of the transferred data, as well as extraction and filtering facilities on news sites.

Table 1. Example of updating news frequency

Service news	URL	Update
French Google	http://news.google.fr	about 20min
Google	http://news.google.com	about 20 min
Voila actuality	http://actu.voila.fr	every day
Voilanews info	http://actu.voila.fr/depeche	instantaneously
Yahoo!News	http://fr.news.yahoo.com	instantaneously
TF1 news	http://new.tf1.fr/news	instantaneously
News now	http://ww.newsnow.co.uk	5 min
CategoryNet	http://www.categorynet.com	every day
CNN	http://www.cnn.com	instantaneously
Company news groups	http://www.companynewsgroup.com	about 40 per day

3.1. UPDATING FREQUENCY

We have done some statistical tests to evaluate the updating frequency [10,11] of news. The results show the different behaviour of interrogated sites. The sites can be classified into two categories depending on the news updating period; very regular with a constant update time, and irregular - information updated when present. The sites can be also classified depending on the refresh time; slow - with a refresh time greater then 10 minutes, fast - information refreshable even about 10 seconds. Our results confirm that the content of the news sites change very often. This is one of the most important reasons for carefully optimising the data flux format.

3.2. PAGE SIZE

The traffic generated on the internet by news is high and it is desirable to optimise it. We have done some comparative tests of the transferred data size for the news presented in HTML and in RSS. The results of comparison are presented in Table 2.

Table 2. Comparison of data size in HTML and RSS

Service news	HTML	RSS
Business Week	47K	9K
TF1	32K	3,7K
Dallas News	22K	0,5K
IOL	14K	2,3K
Prime Minister	37K	2,5K

4. NEWS DATA FORMATS

There are two kinds of data format, most frequently used for publishing of news; the HTML non-structured data format and the dedicated format, called RSS-feed. These formats have different features for publication and presentation, as well as for retrieving and exploring the data by other tools such as catalogues, search-engines and meta-search tools.

4.1. HTML PRESENTATION AND DATA EXTRACTION PROBLEMS

There is a lot of news Web sites that present news using only the standard HTML page form. This page contains a lot of diverse data, not only lists of selected news items but also much more additional information. This additional information is completely useless for the retrieving tools (Search Engine, Meta-Search Engine, etc.). The searching news' agent needs to extract only the significant data. Additional and non-essential data increases the complexity of analysis [1]. The HTML pages, which include the informational and presentational data, are not optimal for data extraction and information content updating. The HTML pages are also not optimal for data transfer because of their size.

The format of the HTML news pages is not standardized. Their form does not lend itself to information extraction. There are two kinds of extraction problems. Firstly, finding all of the news description with their links included in the news page and then identifying only the pertinent ones. The most important difficulties in information extraction are: complex linking, difficulties in recognizing and following links to framed pages and then extraction of the information in this frame, difficulties in identifying links in image maps and in the script code sources like a JavaScript, etc. There is also some information, which is not static, for example, thematic publicity selected automatically and frequently changed. The retrieving tool has to distinguish which piece of data is the news information and select only the significant links.

4.2. 4.2 RSS FEED AND ITS ADVANTAGES

RSS [6]-[9] *Really Simple Syndication* is an XML-based special format that enables web developers to describe and syndicate web site content. RSS provides a static and well-structured format for all the textual documents. The RSS file contains only the informational data formatted in a standardized format without any presentation parts. With RSS files one can create a data feed that supplies all kinds of data: headlines, links, and article summaries from a web site, etc. It is possible to analyse, monitor and to extract data automatically. Because of the advantages of RSS, this format may be attractive for such educational tasks when information is frequently and simultaneously introduced, modified, updated, and exchanged. The RSS-format is also more users' friendly what is important when less experienced users (e.g. students) are involved into the news creation process. The RSS-formatted documents are also well suited for automatic intelligent analysis and processing [11].

5. CONCLUSION

At the beginning, RSS feed files were used only for the news sites. Now, with thousands of RSS-enabled sites, this format has become more popular, perhaps the most widely seen kind of XML. RSS has democratised news distribution and soon the RSS-like portals will democratise the another kinds of site; news services, databases, weblogs, search results, calendars, etc.RSS-feed is also easy to use and well optimized to retrieve the news from the source sites. That is why it is useful to publish the news in two formats; firstly in HTML dedicated to visual presentation and secondly in XML-based RSS-format, which is more useful for information retrieval tasks.

BIBLIOGRAPHY

[1] BERKA P: Intelligent systems on the internet, online: http://lisp.vse.cz, 98.

[2] Berkman Center. Rss 2.0 specification, on-line: http://blogs.law.harvard.edu, 2004.

[3] ASSELIN Ch.: Chercher dans l'actualite recente ou les archives d'actualites francaises et internationale, on-line http://c.asselin.free.fr, 2004.

[4] CAO Y, J. GREEN: Facilitating Web-based Education using Intelligent Agent Technologies. In Proceedings of The Second International Workshop on Web-based Support Systems, In IEEE WIC ACM WI/AT'04, Pekin, 2004.

[5] MARKLAND M. and KEMP B.: Integrating Digital Resources into Online Learning Environments to Support the Learner, In Proceedings of Networked Learning Conference, Lancaster University, April 2004

[6] W3C Consortium. Rdf site summary (rss) 1.0 offi- cial specification. W3C Recommendation, on-line: http://www.w3.org, 2000.

[7] W3C Consortium. Rdf/xml syntax specification (revised). W3C Recommendation, on-line: http://www.w3.org, 2004.

[8] HAMMERSLEY B.: Content Syndication with RSS. Oreilly, 2003.

[9] Webreference. Introduction to rss, on-line: http://www.webreference.com, 2003.

[10] WEGRZYN-WOLSKA K.: Le document numerique: une etoile filante dans l'espace documentaire. Colloque EBSI-ENSSIB; Montreal, 2004

[11] WEGRZYN-WOLSKA K., SZCZEPANIAK P.S.: Classification of RSS-Formatted Documents Using Full Text Similarity Measures, In Proceedings of 5th International Conference, ICWE 2005, Sydney, 2005, pp 400-405