*Proceedings of the First Young Vietnamese Scientists Meeting (YVSM '05), Nha Trang, June 12-16, 2005*

# Building an e-news office supporting news collection from other websites

N.V. Thanh[a], H. M. N. Hai[b], N. D. Hiep[b], P. N. Cuong[a]
[a] University of Natural Sciences, Vietnam National University
227 Nguyen Van Cu St., Q5, Hochiminh City, Vietnam
[b] Pyramid-Consulting Vietnam,
Block 37-37, 364 Cong Hoa St., Q.Tan Binh, Hochiminh City, Vietnam
{nvthanh, pncuong}@fit.hcmuns.edu.vn
{heart_socker, nguyenduyhiep}@yahoo.com

Our main research goal is to propose a framework for and, moreover, develop a typical online newspaper office that can collect news from other websites. This office aim at small companies to help them building e-newspaper with little cost and human resource. We use an open source portal, DotNetNuke, as a base for the office and support a typical news producing process. Moreover, this web-based office can collect information from pre-defined news sources which don't support RSS. In this paper, we describe briefly the structure, function and mechanism of this office and discuss the beneficial meaning of this office with regard to the need of small companies or organizations. We also report our experiences when deployment this office for the Faculty of Economy, Vietnam National University, Hochiminh city. We analyze the results and the problem we encountered, as well as offer constructive solutions.

## 1. Introduction

Without a shadow of doubt, World Wide Web grows rapidly and provides many public information sources like online newspapers. Recently, in Vietnam, electronic newspaper is more and more common. With many advantages, e-newspaper can satisfy many demands of readers and give them the last-updated information. Therefore, constructing an e-news website is an essential requirement of companies to broadcast their related and interested information to customers. However, companies, especially small companies, face with many problems when building their own rich e-news website. The most important problem is the lack of human resource to manually collect, edit and represent useful information on their website. Therefore, an e-news office that can collect desired news from pre-defined source is very useful.

However, we must regard that most of web pages are in the form of Hypertext Markup Language (HTML), suitable for viewing purpose not for automated extraction. Moreover, most of Vietnamese websites do not support Really Simple Syndication (RSS), a standard lightweight XML format for syndicating news and the content of news-like sites, news-oriented and personal weblogs. These make structured data extraction from Internet is not trivial task.

In this paper, we propose a framework for and, moreover, develop a typical e-news office that can collect news from other websites which do not support RSS. We use an open source portal, DotNetNuke, as a base for the office. Using an interactive tool, users will define which information they want to collect from many sources. This tool will auto generate the XML Path Language (XPath, W3C Recommendation 1999) of the user-desired information. Then, the office will regularly and automatically collect news from these pre-defined sources by XPath and simply classify them into user-defined category.

The organization of this paper is as follows. After this introduction, we will review some related works. Section 3 expresses our rational components of the system. In the next section, we describe our deployment for the system. In the final section, we conclude and provide directions for future research.

## 2. Related Work

Data extraction from Internet for further tasks is wide and well research. N. Maria and M.J. Silva (2000) extract documents from web publications and perform a support vector machine learning method to categorize them into multiple categories. Z. Zheng (2001) attempts to build an international news classification and retrieval system. Using a rule-based filtering approach, this system extracts relevant news stories and categorizes them by using naïve Bayes classification.

In Vietnam, with the rapid development of Internet and online newspaper, many recent efforts in extracting web data have done. N.T.Long and N.P.Binh (2002) build an information collecting and parsing client helping personal. Meanwhile, N.V.Thanh et al (2003) construct a stock market news portal for their stock community using a XPath-based extraction approach.

Our system is inspired by the work done by Myllymaki (2000) and N.V.Thanh et al (2003). We incorporate the web data extraction using standard XML technologies (Myllymaki, 2000) into the portal DotNetNuke () to construct our online news office. In the following section, we look through our system architecture.

## 3. System Rationale

Our system integrates a set of software components for retrieval pre-defined news documents and an e-news office supporting a typical news producing process. Figure 1 presents our architecture. It has the following main components:

**HTML Analyzer:** This interactive tool help users determine which information of a website they want to collect. It will store these information sources as XPath sources. Our heuristic is that the similar presentation parts of a website, especially for online newspaper, will have the same function. This inductive heuristic is constructed by examining identical Vietnamese online newspaper such as VnExpress[1] or Vietnamnet[2]. The purpose of this presentation style is help readers view and browse website easily. This is the key characteristic help our crawler extract desired information from these websites.

**Retrieval Crawler:** Based on preset XPath, the crawler will regularly and automatically retrieve news articles from pre-defined sources. The collected documents will be classified into user-defined category or the users can do it manually. This crawler is an extended version of the crawler implemented by N.V. Thanh et al (2003).

---

[1] http://www.vnexpress.net
[2] http://www.vnn.vn

**E-News Office:** This office support a typical news process based on the news process of Youth Online[3] newspaper. The hierarchical organization of our e-news office includes a chief editor, a deputy editor, a general secretary, a few deputy secretaries, some editors and many reporters. Only the editors and the secretaries have the right to publish information on websites. This news producing is a two-ways transfer. The low-level personnel transfer candidate articles, both manual and retrieval ones, to the higher level personnel. At editor level, personnel can reject, edit and send back to the lower level or approve and send to the higher level. Finally, the approved articles will be published on news website.

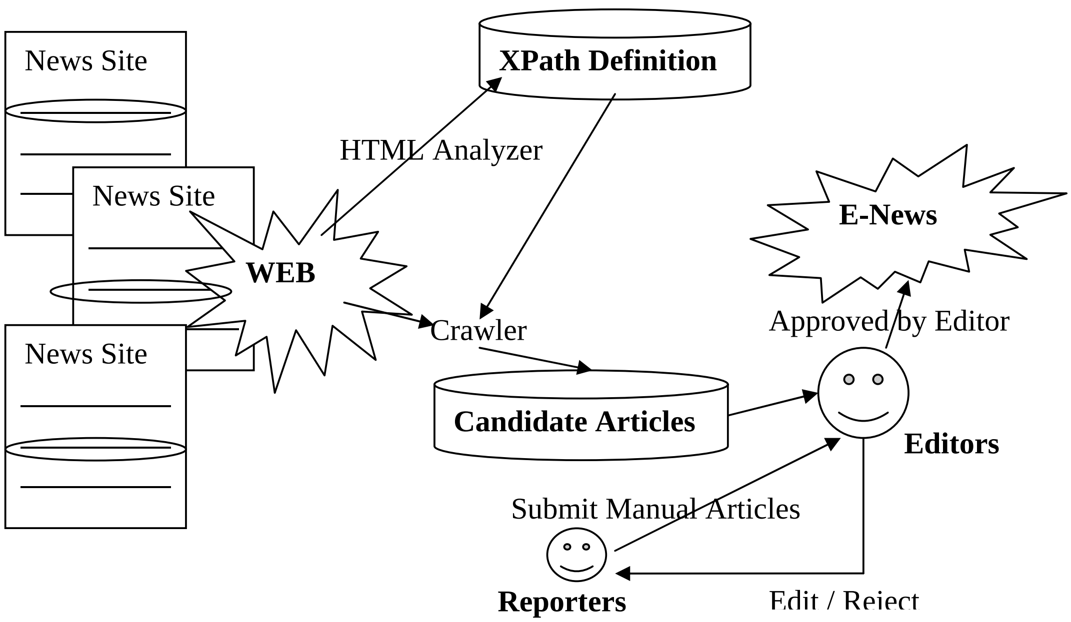


**Fig. 1**: Architecture for retrieval and publishing news articles

## 4. System Deployment

Our system is deployed as a news website for students at the Faculty of Economy, Vietnam National University, Hochiminh City. The teaching staff will define which information sources should be collected for students. Remarking the very low Internet band-width at the faculty, our system becomes essential. Instead of browsing directly many news websites through Internet, now students can view news on our Intranet website. It is evident that our system saves a lot of time for student and band-width for the faculty.

During the deployment, we find that our system work well for almost common Vietnamese online newspaper. However, for several news-oriented websites which have *"ill-formed"* HTML or missing syntax, sometimes our system can not retrieve desired document. Therefore instead of a rigid XPath definition, the future system must have the ability that can *"repain"* the broken syntax and produce the well-formed HTML.

[3] http://www.tuoitre.com.vn

In addition, the most human dependent of our current system is the manual XPath definition and topic categorization. If the news source change it format, although this is rare, the users must re-define XPath sources. Moreover, human categorization makes the collected news is not instant for readers. Both the ability of intelligent XPAth definition and automatic topic categorization will be our future works.

## 5. Conclusion and Future Works

In this paper, we introduce our online news office that supports a typical news producing process and other websites news extraction. Our approach is based on standard XML Technologies like XML and XPath. This system is successfully deployed as a news website for student at the Faculty of Faculty of Economy, Vietnam National University, Hochiminh City. Our experience shows that it will be very useful for small companies and organizations, especially in Vietnam low-bandwidth condition. We will aggressively investigate this approach in following tasks.

First, unlike many foreign systems, our system does not support an automatic topic categorization. This is due to the lack of successful Vietnamese text categorization approaches. Trying to apply English text categorization approaches into Vietnamese is unfeasible because of the ambiguity of Vietnamese word segmentation. Our research group currently carry out an approximate Vietnamese word segmentation only for text categorization. We believe that this will help a lot for further Vietnamese online news text categorization.

Second, our current approach based on manual XPath-defined, in the future, we will focus on automatically identifying XPath of desired news given some sample pages of this source. Regarding the most important parts of a news article are the title, the abstract, the content and, if possible, a picture, we can compare and map correlative parts of the sample pages. With enough sample pages, we can automatically define XPath for this source.

## 6. Acknowledgements

We would like to thank Mr. Nguyen Viet Hoang at University of Illinois – Urbana Charmpaign and Mr. Tran Doan Thanh at Kookmin University for helpful ideas and discussions. This work is supported in part by University of Natural Sciences, Vietnam National University 2004-2005 Research Grant.

## References

[1] DotNetNuke Inc., *DotNetNuke Portal*, http://www.dotnetnuke.com.

[2] H.M.N. Hai and N.D.Hiep, in: *Building an e-news office supporting news collection from other websites*, Bachelor thesis, University of Natural Sciences, Vietnam National University, 2004.

[3] N.T. Long and N.P. Binh, in: *ICPS - Information Collecting and Parsing System,* Vietnam Intelligence Contest Document, 2002.

[4] N. Maria, M.J.Silva in: *Theme-based retrieval of Web News*, Proceedings of the Third International Workshop on the Web and Databases, (WebDB2000), 2000.

[5] J. Myllymaki, in: *Effective Web Data Extraction with Standard XML Technologies*, Proceedings of the Tenth International World Wide Web Conference (WWW10), 2001.

[6] N. V. Thanh, T. D. Thanh, N. V. Hoang, in: *TTCKOL – An online stock market community*, Vietnam Intelligence Contest Document, www.ttckol.com, 2003.

[7] W3C, *Extensible Markup Language (XML)*, W3C Recommendation, February 1998 http://www.w3.org/TR/REC-xml.

[8] W3C, *XML Path Language (XPath)*, W3C Recommendation, November 1999. http://www.w3.org/TR/xpath.html.

[9] Z. Zheng, *International News Connection: A Real-time Online News Filtering and Classification System*, ACM SIGIR'01 Workshop on Mathematical/Formal Methods in IR, 2001.