WEB INFORMATION MANAGEMENT SYSTEM: PERSONALIZATION AND GENERALIZATION

Sung Sik Park, Yang Sok Kim, Byeong Ho Kang
School of Computing, University of Tasmania
Hobart, Tasmania 7001

ABSTRACT

Our research focuses on web information management for people who want to monitor and use the World Wide Web (WWW) information, as their information resource. Web information is mainly open to the public and search engines are widely used, although there are complaints about the large amount of irrelevant information. Some web technology research focuses on promptness of changed information and relevance to users. Most web search engines do not fully satisfy these two requirements in general because they try to cover all web sites and users together. The aim of Web monitoring research is to overcome these problems. Web monitoring systems check predefined web pages and prompt users when there are changes in these pages. This research focuses on the management of newly uploaded information in target web sites. In a web monitoring system, the system should cover various different types of web pages (generalization), as well as personal aspects (personalization). This system achieves these tasks by integrating the first and third modules. In addition to web monitoring functions, the second module provides the information management functionality in the user's local storage.

KEYWORDS

Web Information Management, Web Information Sharing, Web Monitoring.

1. INTRODUCTION

WWW has become the most popular information source for people because the Web is now the largest sharable and searchable repository of information. People normally visit some sites regularly. Portals commonly are one of people's favorites and individuals have their own interest. However, people limit the number of sites visited by themselves because the number of regular sites is often beyond what they can handle.

There are two different types of information: updated information and newly uploaded information. Updated information occurs when information is changed information such as values are varied on existing web sites. For example, share prices and exchange rates. The web page itself is not new, however, people update part of the web page regularly. Newly uploaded information in the web site means a new page, which was not available in the past, was added, such as Web based newspaper sites. They upload the news articles regularly and users monitor this page to know what's happening. Though the value update change can give good information in some aspects, such as historical data, however it is not used as the resource in our knowledge management system. In this study, we focus on the newly uploaded information.

There are many approaches to web information management but they usually focus on a specific function, such as monitoring, searching, or sharing. While these approaches can give good directions, in our view these methods must be integrated to support successful web knowledge management. Our approach is to develop an integrated system for web knowledge management. To this end our system integrates the following three core components: web information finding and gathering, web information managing, and web information sharing. Another aim of our system is personalization of web information management because the needs for web information differ from person to person and organization to organization. The difference exits in time, knowledge structure, and sharing intention.

This paper is structured as follows: part 2, will review related research results; in part 3 will explain our system's architecture; parts 4, 5, and 6 will explain web information management functionality; and, part 7 will suggest conclusions from our research.

2. LITERATURE REVIEW

2.1Use of web information

Sellen et al. [3] reported on a diary study of how and why a knowledge worker uses the World Wide Web. They classified six activities: finding, information gathering, browsing, transacting, communicating, and housekeeping. Activities that relate knowledge management are 86% of all activities[1]. They specify desirable technologies as follows: temporary save and display, easy calling; automatic monitoring for "Finding" technologies; better tagging of information; better search tools; web scrapbooks; and, better history functions for "Information Gathering" technologies. However, these technologies only cover information finding or gathering functionality. The knowledge managing and sharing functionality are needed to support the whole knowledge management system.

2.2 Information finding and gathering

2.2.1 Searching

Traditional search engines are not effective in finding newly changed or uploaded information because they gather information based on the pre-defined search schemes, that is, they are not effected by individual web page alterations [15]. Particularly, people have to wait until the search engine includes this updated or new information. For newly uploaded information, the search engine cannot successfully fulfill a user's request until the new information is found and made available. In the case of updated information, the search engine may still locate the web sites, if the user looks for the sites using the unchanged portion of the contents. For example, a web page for share prices. Often the only information changed in these pages are the values of individual company's share prices.

2.2.2 Web Monitoring

Web monitoring seeks to overcome the above problems by limiting the amount of target web pages and detecting changes automatically. Tan et al. [15] gives a back ground study about web page monitoring whether they analysed the behaviour of 105 web pages over a one-month period. Changes were identified based on the following analysis of the HTML code: text analysis, link analysis, image analysis, layout analysis, and the last date updated. Their research results show that: 44.8 % of pages changed during the period; text and hyperlinks are the main attributes changed; the frequency differs from domain to domain; and, 72% of those web pages are changed in a week. These results justify why web monitoring is needed.

1) Target Information

This problem is essentially one of how to extract data from a web page. The most common approach is to use some kind of wrapper technology. Although wrappers provide an effective way to extract data from Web sources, they require a high level of expertise, are difficult to maintain and may not be robust [10] [11] [12]. Another approach is based on Natural Language Processing (NLP) [13]. However, extensive corpus training and a rich grammatical knowledgebase is required [14]. Rahardjo and Yap [14] also suggest approximate matching techniques for monitoring portion change recognition. Foo et al. [20] use Programming By Demonstration (PBD) technique. Lu et al. [4] use a block heading-tree approach to support information specification [4]. Flesca et al. [6] also suggest a method for specific portion change monitoring by using a document tree [6].

[1] Finding 24%, information gathering 35%, browsing 27%, transacting 5%, house keeping 5%, and communication 4%

2) Change Recognition

Liu et al. [5] use a sentinel change detection algorithm to monitor web pages. They chose nine basic sentinel types: any change[2], link change, image change, word change, phrase change, table change, list change, arbitrary text change[3], and keywords. Tan et al. [15] also use four functions to implement a web monitoring system: web page update data monitoring, keyword monitoring, link monitoring, and portion monitoring[4]. Boyapati et al. [21] present site level monitoring tools. They reported a system called ChangeDetector, which uses machine learning techniques for intelligent crawling, page classification and entity extraction to filter detected web site changes and report them to the user.

2.3 Knowledge Managing

Knowledge management is the systemic and organizationally specified process for acquiring, organizing, and communicating employees knowledge, so that other employees may make use of it to be more effective and productive [16]. Knowledge management systems (KMS) are tools to effect the management of knowledge and are manifested in a variety of implementations [17] including document repositories, expertise databases, discussion lists, and context-specific retrieval systems incorporating collaborative filtering technologies. Hahn and Subramani [18] identify issues and challenges related to the utilization of information technology for knowledge management support in three phases of deployment: the setup phase; the on-going utilization and maintenance phase; and, finally, long-term effects of knowledge management support. In the setup phase KMS must balance information overload and potentially useful content. In the maintenance phase KMS must balance additional workload and content accuracy. Lastly, from the long-term effects view KMS must balance exploitation and exploration.

2.4 Knowledge Sharing

The key to using the Internet effectively for competitive intelligence is not just knowing where and how to look for information, but how to share or distribute information. In mid-1996 Internet software companies began to develop applications using what is now known as "push technology" in an attempt to help users cut through the clutter and retrieve only that information they specify as relevant. However, as noted by Buchwitz [8], push is not the panacea of the Web's information overload problems because users tend to weary of their push clients when they become overwhelmed by too much delivered content. This effect can be avoided only through careful monitoring of information specifications, and continual editing and refining of web information management agents. Brandt et al. argue that web push is too narrowly defined as an alternative delivery method for web content and that the technology should be built as a special case of a more powerful and flexible Internet notification service. They view asynchronous notifications as a basic form of communication in distributed environments.

3. PERSONALIZED WEB INFORMATION MANAGEMENT SYSTEM ARCHITECTURE

PWIMS is a stand-alone personalized web information management system. The system consists of three main components as follows (shown in Figure 1):

- Information monitoring agent that gathers information from target web pages
- A knowledge management component that enables users to organise the monitored information
- A knowledge sharing agent that disseminate the obtained information by using personalized push technology, including e-mail notification and Web posting.

[2] Any updates on the page object – watch any change to the modification timestamp of the file, compare MD5 hash values.

[3] Any change to the text fragment specified by a regular expression – identify and detect any change; in the selected fragment.

[4] Tan et al uses follow 5 change categories – text analysis, link analysis, image analysis, layout analysis, and last update date.

These components are integrated to support a complete and personalized web information management system. Users can configure the system by using a configuration manager. System configuration includes internet connection options, knowledge sharing options, information gathering options, knowledge managing options and knowledge sharing options. To start PWIMS users must specify the above options. After initializing PWIMS users can register target web sites for information monitoring, can manage knowledge using a personalized folder structure storage system and can share gathered information using the Knowledge Sharing Agent.

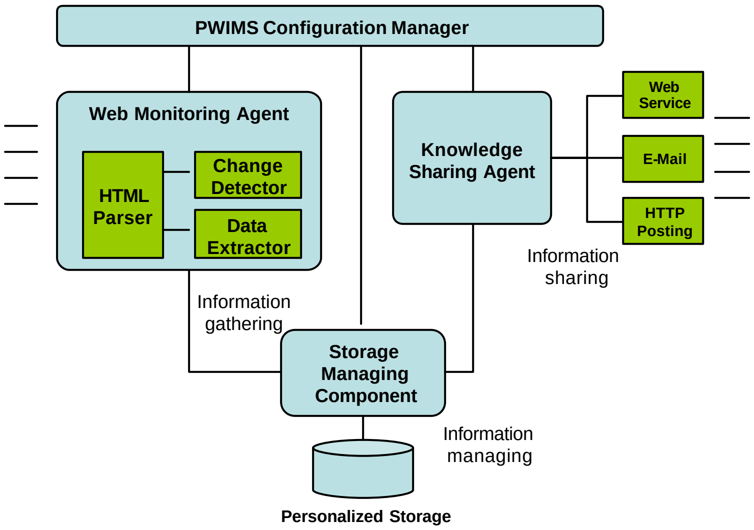
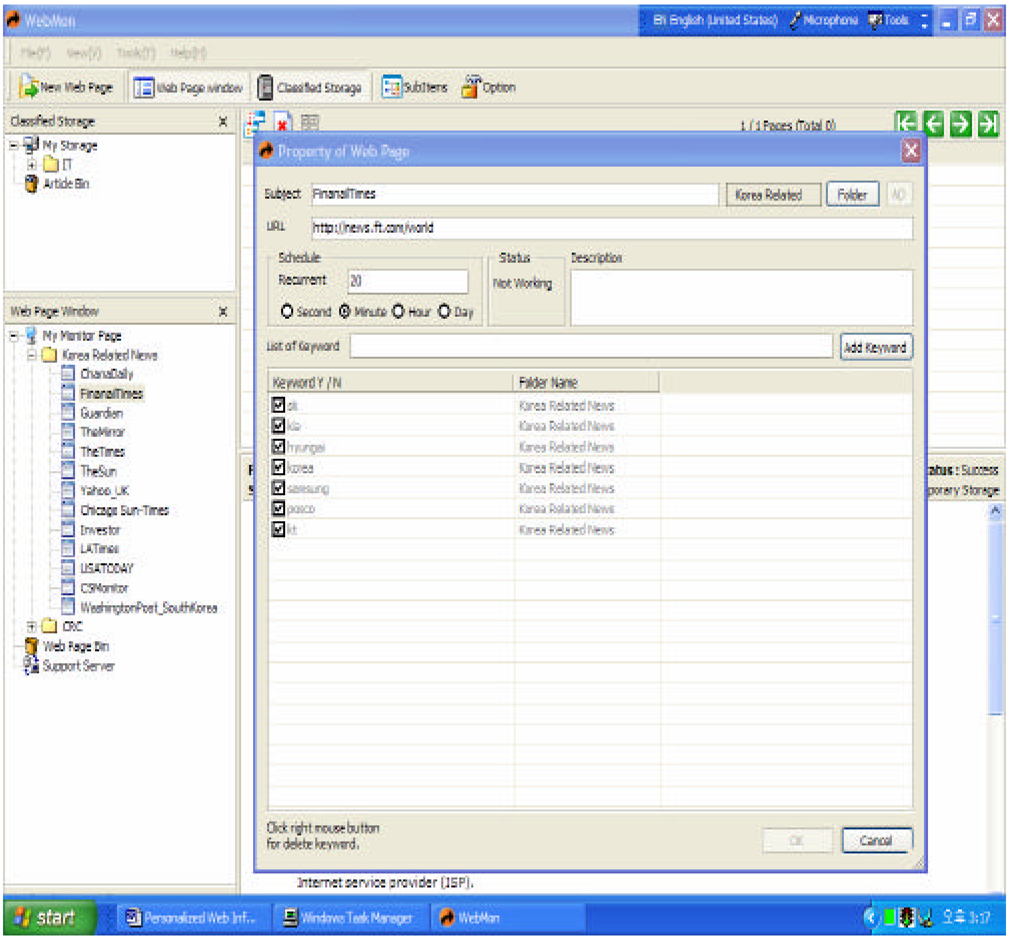


Figure 1. PWIMS System Architecture

4. WEB MONITORING AGENT

4.1 Selective Pull

To support personalized web monitoring, we use a selective pull method. Selective pull means the user subscribes to certain types of requirement specifications. The service, once set in motion, is automated and will change only if the user requests it. In our system there are three main required specifications: target web pages, schedule, and keyword. Firstly, users specify target web pages to be monitored. This is not necessarily the homepage of the target web site. Rather, it must be the web pages that includes the required information. Secondly, users must specify the monitoring schedule, which depends on the user's monitoring intention and the target web site's modification cycle. However, users usually don't know the exact a time of these changes, therefore, our system automatically finds the target web page's modification cycle, using a modification time detection module, and suggests an appropriate monitoring schedule. Thirdly, users must specify target information by using keywords. Because our system supports hierarchical folder structure when we organize the monitoring subscription, users can systemically organize target information. Users can create a folder that is an abstract concept of special knowledge. We use the term "abstract" because the folder represent many keywords, which are used to justify why he/she created the folders. These keywords automatically are included by the system when user adds a target web page below this folder. Users can make an exception for the site by exclude some keywords from the parent folder or adding new keywords. Figure 2, shows the target web page specification process.



Subject FinanalTimes

URL http://news.ft.com/world

Schedule

Recurrent 20

○ Second ● Minute ○ Hour ○ Day

Status Not Working

Description

List of Keyword

Keyword Y / N	Folder Name
☑ sk	Korea Related News
☑ kia	Korea Related News
☑ hyungai	Korea Related News
☑ korea	Korea Related News
☑ samsung	Korea Related News
☑ posco	Korea Related News
☑ kt	Korea Related News

Click right mouse button
for delete keyword.

Internet service provider (ISP).

Figure 2. Personalize Web Monitoring

To effectively support keywords we use the stopping word algorithm. Due to our focus on change through the creation of new information, our keyword finding activities occur in the linked contents, not in the monitored web pages. To support keyword searching we use a binary search algorithm.

4.2 Data Extraction and Change Detection

Web pages usually mix text with images, advertising banners, lots of menu items and target information, therefore the monitoring system must be able to support target data extraction. As noted above there are various methods for data extraction. Though these researches suggest methods for web pages monitoring, they only focus on changes using value update [4].

However, when we monitor the newly created information, sometimes the linked contents are more important than the current target web pages, because the information that user's want to get is usually located in the linked content, such as newspaper sites and bulletin notice boards. Usually the web pages where information changes take place are normally expressed in headlines on the front pages.

Many researchers separate data extraction from the detection of web page changes. However, our system integrates the two functions because, we focus on newly uploaded information. This means the change can be detected by observing hyperlink changes in the target web pages and the required information is usually located in the linked content. To perform data extraction and change detection we use hyperlink comparison as follows:

- Once PWIMS is initialized by the user, it collects HTML source (HSold) and detects hyperlinks and makes a list of hyperlink's information (HLold) base on absolute original URL.

- At the scheduled time, which is specified by the user, the PWIMS agent gets the new HTML source (HSnew) from the target pages (the same as that used in the initial stage).
- The PWIMS compares hyperlink's Information (HLold with HL new) to make a decision whether the links have changed or not using change decision criteria (shown in Table 1). The PWIMS gathers the linked information (TI) for keywords searching in the HTML source.
- If the keywords exist in the HTML source of article, the system marks green icon to the new articles for notification and to be known user easily. (Figure 3 Monitoring report screen shot)

Table 1. Change Decision Criteria

Title of Objects	URL	Change Decision
Unchanged	Unchanged	No
Unchanged	**Changed**	**Yes**
Changed	Unchanged	**Yes**
Changed	**Changed**	**Yes**

* If the title or subject of objects changes without the URL changing, the title added old hyperlink's information. Because this means that the target URL web pages have been modified by the information writer.

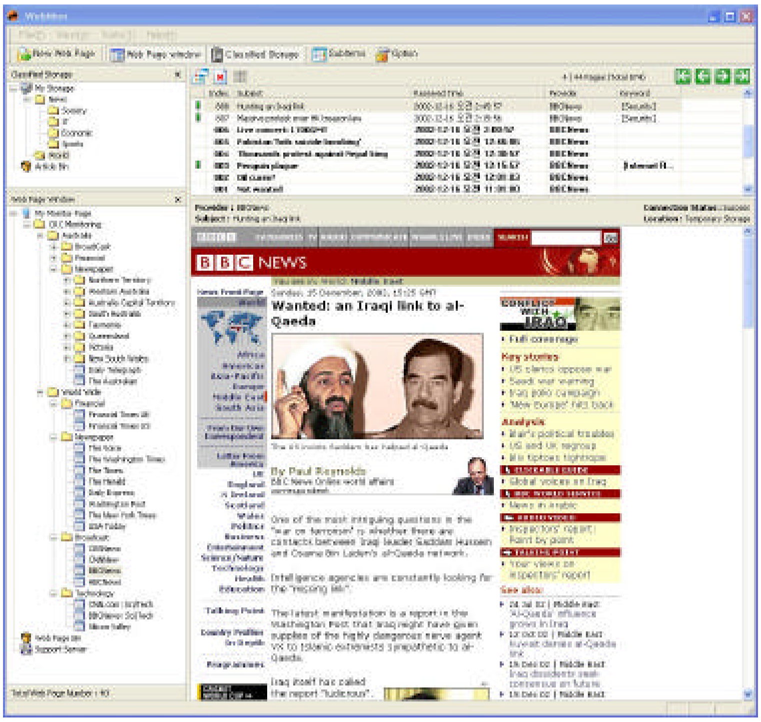


Figure 3. Monitoring report screen shot

5. KNOWLEDGE MANAGING COMPONENT

The knowledge managing component mediates between the knowledge monitoring agent and knowledge sharing agent. To support knowledge management our system gives the user the following functions:

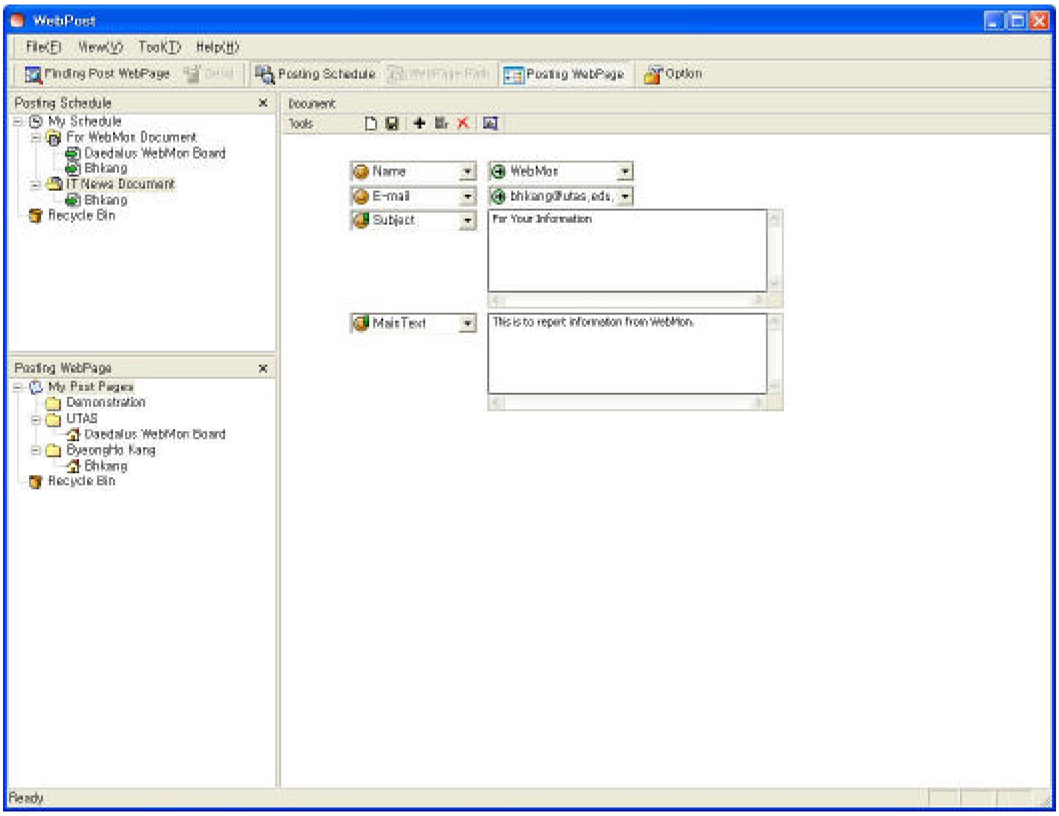
- Local storage which is a folder structure and can store the acquired information
- Folder control that is able to manage the information, such as create, copy and move.

By using these functions users can manage web information more systematically and in a user friendly manner: can use web information whenever they want to whether the system is online or not; and, can allow users to share his/her information with others.

6. KNOWLEDGE SHARING

In our system, push technology is used to share information with others. There are a number of different models for information push on the Internet: broadcast, selective or automated pull, distributed push/pull, interactive push, and customized pull [8]. There are two kinds of methods in our system that users can use for knowledge sharing. Firstly, users can send the gained information by using e-mail. Users can register the receiver's e-mail address to monitored web sites on the monitoring folder structure. Users can select either, for the e-mails to be sent automatically to recipients when the new information is detected or sent upon the user's requested. Secondly, users can post information on the special target web pages (HTTP posting). The HTTP posting procedure takes places as follows:

- Register the posting target site URL to the posting agent. This process specifies the posting target web pages.
- Define the data format that will receive the information from posting agent because the posting target web pages differ from each web site.
- Define the data format that the posting agent. This process defines what information will be posted to the posting target web pages.
- Define the sending option. User can post the information on the fly or not.



Name: WebMon

E-mail: bhkang@utas.edu.

Subject: For Your Information

MainText: This is to report information from WebMon.

Figure 4. Web posting screen shot

7. CONCLUSION AND FUTURE WORK

Our research focuses on information monitoring of newly uploaded web pages and uses this information as the knowledge resource. Particularly, not only should the system be general enough to share the same control and resources, but it also should be personalized to satisfy the various demands from the user. We integrated web monitoring, knowledge management, and knowledge sharing functions in our system. As shown above these functions must be closely interrelated with each other. Further research will focus on the interoperation of these three functions. This system introduces the information management system with the virtual folder interface and it enables the user to easily manipulate the multiple contexts.

Personalization is another requirement for the web information management. The system has three functions for personalization. In monitoring, we use a hierarchical folder system for grouping the web sites from the same domain. The user can operate the monitoring or define keywords by a group while it is possible to manipulate by an individual site. Also, the system provides the exceptional options for the individual site for keywords and scheduling. In the case of keyword maintenance, an individual can have their own keyword sets and exclude default keywords from parents as an exception.

There are many issues to be studied further and this system is a milestone to demonstrate the integration of web page maintenance system as the information resources with personalized web monitoring functions. It requires an extensive empirical study to prove the system's performance. It is now used in several organizations as their information monitoring and management system. It is too early to claim the general success of the system but it appears quite useful and the users have been able to manage the system without the help of computer engineers.

REFERENCES

[1] Glover, E. J., Lawrence, S., Gordon, M. D., Birmingham W.P., Giles, C. L., 2001. Improving Web searching with user preferences: Web Search – Your Way. *Communication of the ACM.* December 2001 / Vol. 44, No. 12.

[2] Balabanovic, M., Shoham, Y., 1997. Combining Content-Based and Collaborative Recommendation, *Communication of the ACM.*.March.

[3] Sellen, A. J., Murphy, R., Shaw, K. L., 2002. How Knowledge Workers Use the Web. *CHI 2002.*April 20 – 25. Minneapolis, Minnesota, USA.

[4] Lu, B., Hui S. C., Zhang, Y., 2002. Personalized Information Monitoring Over the Web. *First International Conference on Information Technology & Application (ICITA 2002).* 25-28 Novembwe, Bathurst, Australia.

[5] Liu, L., Tang, W., Buttler, D., Pu, C., 2000. Information Monitoring on the Web: A Scalable Solution. *International Conference on Information and Knowledge Management (CIKM)* .7-10 November. Washington D.C., ACM Press. Pp 512 – 519.

[6] Flesca, S., Furfaro, F., Masciari, E., 2001. Monitoring Web Information Changes. *International Conference on Information Technology: Coding and Computing ITCC.* April 2-4.

[7] Mladenic, D.,1999. Text-Learning and related Intelligent Agent. *Application of Intelligent Information Retrieval*, July – August.

[8] Buchwitz, L., 1997. Monitoring Competitive Intelligence using Internet Push Technology. *Competitive Intelligence Review.*

[9] Brandt, S., Kristensen, A., Web Push as an Internet Notification Service. *Hewlett-Packard Laboratories*, Bristol, UK Available on http://keryxsoft.hpl.hp.com/doc/ins.html.

[10] Ashish, N., Knoblock, C.A., 1997. Semi-automatic wrapper generation for internet information sources. In Proc. *Second IFCIS Conf. on Cooperative Information Systems.*

[11] Muslea, I., Minton, S., Knoblock, C.A., 1999. Hierarchical Wrapper Induction for Semistructured Information. *In Proc. of Intl. Conf. on Autonomous Agents*.

[12] Sahuguet, A. and Azavant, F., 1999. WysiWyg Web Wrapper Factory (W4F). *Proc. of WWW Conference.*

[13] Soderland, S., 1997. Learning to Extract Text-based Information from the WWW. *In Proc. of the Third Intl Conf. on Knowledge Discovery and Data Mining - KDD-97.*

[14] Rahardjo, B., Yap, R.H.C, 2001. Automatic Information Extraction from Web Pages. *SIGIR'01.* September 9-12, New Orleans, Louisiana, USA.

[15] Ong, H. L., Tan, A.H., Ng, J., Pan, H., Li, Q.X., 2001. FOCI : Flexible Organizer for Competitive Intelligence. *CIKM'OI.* November 5-10, Atlanta, Georgia, USA.

[16] Alavi, M., and Leidner, D. E., Knowledge Management Systems: Issues, Challenges, and Benefits, *Communications of the AIS* (1:Article 7).

[17] Davenport, T. H., De Long, D. W., and Beers, M. C., 1998. Successful Knowledge Management Projects. *Sloan Management Review.* Winter, pp. 43-57.

[18] Hahn, J., Subramani, M. R., 2000. A Framework of Knowledge Management Systems: *Issues and Challenges for Theory and Practice.*

[19] Liu, L., Pu, C., Tang, W., 2000. WebCQ– Detecting and Delivering Information Changes on the Web. *CIKM 2000,* McLean, VA USA.

[20] Tan, B., Foo. S, Hui, S.C., 2001. Monitoring Web Information using PBD Technique, *Proc. 2nd International Conference on internet Computing(IC'2001).*June 25-28, Lasvegas, USA,.

[21] Boyapati, V., Chevrier, K., Finkel, A., Glance, N., Pierce, T., Stockton, R., Whitmer, C., 2002. ChangeDetecter: *A Site level Monitoring for WWW.*