

Using WWW Page Watching Robot as an Information Mediator

1 A C - 3

Saeyor SANTI 石塚 満
 東京大学 工学部 電子情報工学科
 e-mail: santi,ishizuka@miv.t.u-tokyo.ac.jp

1 Introduction

As the huge scale of information grow according to the growth of World Wide Web (WWW), the load of information retrieving is put on users of the WWW. The process is not only time consuming but also vain in case of no change made on the specific site. In other words, WWW users don't even know whether the effort on data retrieving has meaning or not in each time they send access request to the WWW. Since the information put in each site on WWW is stock type or passive. The users have to access to the specific information site in order to get prospective information. The process seems to be easy but it is just routine job. Problems such as time consuming and complexity of site information handling occur when the number of information site keep increasing. Moreover the information in the WWW is supposed to be changed dynamically without any notification to all the users who deal with the information in that site. One way to keep informing the client of each site is to inform directly to the client who register to the site specifically. So long, this fashion have been used with tradition magazine and so on. Today, the world of information is playing its role in different style. The users of WWW must spend a lot of their effort to register to every interesting site in order to have themselves notified when something was changed. That's why the users have one more duty to update the information derived from the specific site themselves with up to date information retrieving from the identical site. We are developing mediator mechanisms in order to shelter the users from massive data flow and serve them with interesting information without troublesome routine works. In this paper, the page watching robot and its relationships with mediator mechanisms are raised to the issue.

2 Virtually Circulated Information on Traditional Stock Type Information System

The problems of stock type information system are mentioned above. We need to circulate the information in WWW space in order to track the dynamic

changes within the information system. The cycle of circulation starts at the source of information. The information flow through the network which individual user apply its own filters or mediator mechanisms. At this state the information will be classified to serve the interests of the user. Unfortunately, since we base ourselves on stock type information system, we need to virtually bring about the flow of information in the network. The circulation can be classified into two kinds, the open loop information flow and close loop information flow. The first kind is straightfor-

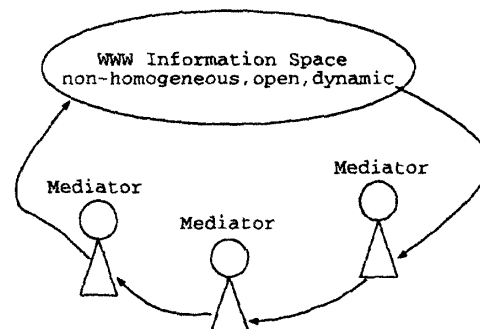


Figure 1: Virtually circulated information on traditional information system

ward, the information flow caused in network will be terminated somewhere out of the information source. This case is the most common case in nowadays situation since the feedback of information flow is still not accepted on the original site. However this kind of circulation is enough to extract the information changes occurred in each registered site. The second type of circulation doesn't have terminal point since the information flow through the network and the source of information. In this paper, the open loop information circulation is taken into account. Even there exist the terminal point of the flow, the current of information flow is caused by the third parties which represent the users of WWW. As a result, the circulation is virtually happened for information in the stock type information sources.

3 WWW Information Mediator

While the circulation of information is brought about on the network, it is the responsibility of the information mediator which the information current of flow through. This research proposed the concept of mediator mechanism applied to the flow of informa-

tion. The mediator is responsible for generating virtual information flow as explained in previous section. Moreover it is also responsible to apply the information current to the specific information filter designed or customized for the user. The definition of information mediator can be discussed in various aspects depends on the scope of its functions. Some may add intellectual capability and behaviors to the definition. Some may include autonomous mechanisms, adaptive learning capability, analytic capability, and even navigation capability to the definition which lead to the agent approaches. However, the mediator's definition is still not specified as de facto standard. This paper raised a case study which utilize WWW page Watching robot as a simple information mediator.

4 WWW Page Watching Robot

The WWW Page Watching Robot was designed to work as an information mediator for stock type information system like the WWW. It is responsible for detecting non-trivial changes of specified pages on various sites. For this experiment, it is presumed that the specified pages are in the set of interesting page from the user's point of view. The robot starts with bringing about the information flow in the network. This process can be achieved by sending a request of WWW page source from the httpd server on destination sites. It manages the page source retrieved in its

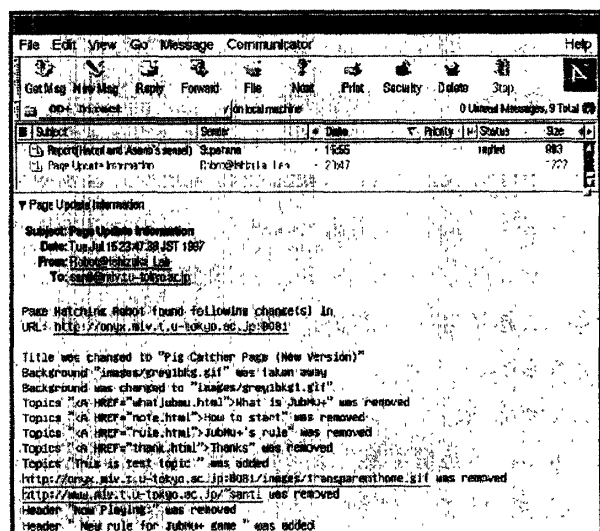


Figure 2: An email that notify the user of changes in specific WWW page

own database and check for the non-trivial changes by comparing with the old copy in the database. Once an interesting change is detected, it will be accumulated in an profile for that page. The process is running through the page comparing step. Finally, the accumulated profile will be sent to inform the user about changes in that page via email.

5 Strategy

There are many programs working like the WWW Page Watching Robot but almost of them is checking only the page specified by user. There are many problems for this superficial method since there are uncountable pages which utilize frame structure out there in the WWW. In those pages, the real information or data are in the difference pages displayed in each frame as shown in figure 3. The content of

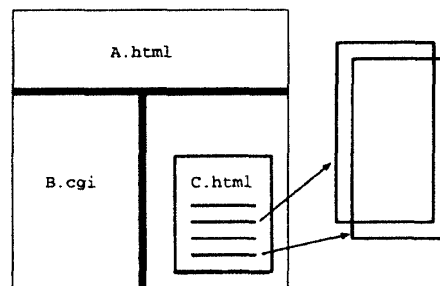


Figure 3: Web page which utilize frame structure

each frame may not conform to just HTML file but also even appears as cgi or other kind of scripts. The robot in this experiment extends the mode of retrieval to the link within each page upto two level. The first level refers the links within the specified page. The second level refers to the links within the first level page which reside on the same site as the first level page. The structure is shown in figure 3. In this case the page that utilize frame structure is the original page, the page in each frame is the first level page, and from the first level page there are only some pages that belong to the same site retrieved. By this extension, the information on the specified page and its relative links is inspected by the robot which results on the higher accuracy of page watching process.

6 Conclusions

The concept of circulating information and discussion on using WWW Page Watching Robot as an information mediator was presented in this paper. Even the circulating information system is rarely implemented nowadays, the virtually circulated information on traditional stock type information system reveals its use for simple purpose like detection of page changes. The experiment results still show the possibility to extend its capability in the future.

References

- 1) Oren Etzioni, Daniel Weld A Softbot-Based Interface to the Internet *Comm. of ACM, July '94.*
- 2) Deniela Rus, Robert Gray, and David Kotz Transportable Information Agents *Department of Computer Science Dartmouth College Hanover, NH 03755*