WWW Page Watching Robot

5L－7

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

1 Introduction

Concepts of information society become realistic environment at the moment and even faster than the past. Among the world that filled with overwhelming data flowing in the network, we need a reliable and robust information processing system to classify the significant information according to the interest of individual user. Human may consider this task as a very simple job because we understand natural language. However, we hardly cope with or handle all of incoming information. At this point, we need computer to assist us in this tedious task. So far, software robot would not handle this task with ease because relevant processes or considerations are heuristic. The system plays importance roles on providing user with interesting and important information. At present, the most familiar environment is the ubiquitous World Wide Web (WWW). Unfortunately information in WWW is presented in passive way. For instant, one have to access specific page constantly to check its content for new updates. The system proposed in this project is designed to reduce user's effort on routine works and time by working as page watching robot. The robot is expected to check contents of target pages whether important parts or essential data are updated or not and notify the user if significant changes are detected. The abilities of explained system serve as bases of further sophisticated information agents.

2 Design Principles

In designing our page watching robot, we have considered usability of the robot rather than sophisticated functions for autonomous information agent. This project is planned to launch a prototype of page watching robot which run on the internet environment. The methods used in this project are simple and straightforward. We focus on how to keep the robot running with full of its effort and how to serve multi-user effectively. Our page watching robot embodies the following ideas:

1. **Reasonable:** since the main purpose of the system is to inform the user of meaningful changes in each page that is once read by the user. We have no need to waste the time on reading the same page without meaningful updated data. The robot has to balance the accuracy according to the changes of content in specific page, against the number of nuisance annoucements forwarded to the user. The robot is responsible for deciding how and when to inform the user of significant changes in given page. At this point, some challenging problems may occur if we consider the changes in figures of that page. Anyway, they are beyond our scope here.

2. **Dynamic:** pages in WWW are supposed to be changed either in content structure or even location. The robot should give an appropriate respond to the user, for example, it should inform the user about the obsolete page location or even interpret the new location in case that it is recorded in the old page location. If the page watching robot can detect the new location of that page, it should be able to update its database and inform the user of this change. It should establish a suitable method for dynamically change and update the database of the system corresponding to actual changes of information sources which are distributed around in the network.

3. **Robust:** since the internet access conditions for links or sites are different and sometimes there are network error. These may be caused by link's speed and reliability. The problem may also occur from the temporary shutdown of the destination site so that the robot could not contact with that site. Time out policy must be applied in each connection request in order to prevent unpredictable effects which may occure during the attempt to access destination site. The robot should be able to cope with these problems after it reach the time out. It is responsible for new scheduling or marking that site as invalid location and sending system conditions to the user after multiple fails.

Saeyor SANTI, Hiroshi DOHI, Mitsuru ISHIZUKA
Dept. of Information and Communication Engineering,
Faculty of Engineering, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113, JAPAN

3 Architecture

The system is constructed in client and server style in order to support multi-user system. The diagram of server side and client side are shown in figure 1 and figure 2 respectively. The users have to register their interesting pages together with their email address. We need email address because the page watching robot has to send email to inform each user when their registered pages are updated or changed. This information is kept in database of the system and is accessible from the page watching robot. The information of users and pages is utilized by the scheduler and the announcer too. The scheduler will look up the pages registered for each user then make schedule of checking for the user. It construct a time table for the robot to make sure that each user will be served right on each specific schedule. The page watching robot will be excited by the scheduler together with user identification and specific pages list. It has to check the content of those pages to see whether there are significant changes in them or not. In order to do that, the robot will initialize http connections via the internet to access WWW. Once the robot detect meaningful changes in each page, it will send a signal to excite the announcer. The announcer will get specific user information and event message from the robot and send email to the user via the internet.

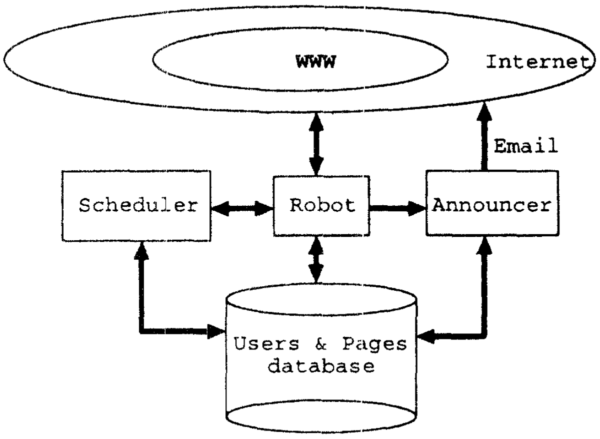

Figure 1: *Architecture of page watching robot on server side*

On the client side, a group of users can access the page that gather and display user's status and register pages. Each user can add or delete their interesting pages to be checked by the page watching robot. Each user must have internet access in order to contact with the data collector page which is located in WWW domain. All the data in this page are directed to the robot via the internet. After the robot get the data, it will update and excite the scheduler to make an appropriate time table again.

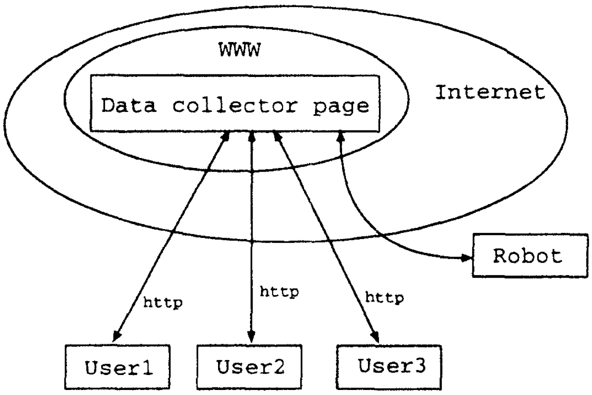

Figure 2: *Diagram of the system on client side*

4 Scheduling

As we see in previous sections, the database and status for users and page locations are supposed to be changed dynamically. The scheduler is responsible for creating the updated time table which is used to excite the page watching robot. Due to the system that run in multi-user environment, we have to synchronize the updating process with the update request sent from the robot. The scheduler has its own clock which is used along with the time table for telling the robot to start checking the given pages. It is also responsible for making the schedule again in case that some pages are not accessible at that time.

5 Conclusions

The concepts for page watching robot is presented in this paper. This robot serves many basic functions of information agent. Those functions do many routine jobs in order to fulfil the purposes of our page watching robot. In anyway, we need more sophisticated algorithms to have the page watching robot run much more profitably.

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*