User-Centered Agents for Structured Information Location

Xindong Wu[1], Daniel Ngu[2], and Sameer S. Pradhan[1]

[1] Department of Mathematical and Computer Sciences
Colorado School of Mines
1500 Illinois Street, Golden, CO 80401, USA
[2] School of Computer Science and Software Engineering
Monash University
900 Dandenong Road, Melbourne, VIC 3145, Australia

Abstract. This paper designs an electronic commerce system that integrates conventional electronic commerce services with contemporary WWW advantages, such as comprehensive coverage and agents for information search and selection. We use a user-centered approach and apply data mining techniques in the design of agents for information search and selection. There are various agents in this electronic commerce system to perform different functions. Among them, SiteHelper is a unique agent in our system compared to existing electronic commerce systems. It acts as a housekeeper for the system and as a helper for the users to find relevant information. In order to assist the users in finding relevant information at the centralized location (with Web links to the global Web), SiteHelper interactively and incrementally learns about each user's areas of interest and aids them accordingly, by deploying data mining techniques with incremental learning facilities as its learning and inference engines.

1 Introduction

Conventional commerce systems have traditionally stressed service, organization, and centralization. Electronic commerce systems have positioned themselves to absorb and take advantage of every new development including the World Wide Web (the Web or WWW for short). The bright colors, hypertext format, graphical user interfaces of the WWW have been widely used in the electronic commerce community to provide the access to multiple remote services. The World Wide Web has embodied flexibility, rapid evolution, and decentralization. Therefore, an electronic commerce system needs to bring together traditional notions of commerce systems with contemporary WWW capabilities.

In this paper, we use a user-centered approach and apply data mining techniques in the design of agents for information location. There are various agents in our electronic commerce system to perform different functions. Among them, the following three agents are important.

J. Liu and Y. Ye (Eds.): E-Commerce Agents, LNAI 2033, pp. 323–331, 2001.

- *Relevance Verification.* Our electronic commerce system has a dictionary to define the scope of documents[1] relevant to the system. The dictionary contains keywords in a hierarchy that describe areas of interests and related detailed topics. When the electronic commerce system is deployed at a different place with a different scope, only the dictionary will need to be changed. The Relevance Agent (RA) in the electronic commerce system verifies the relevance of recommended documents from sellers and Web users, using the dictionary.
- *SiteHelper.* This is a unique agent in our electronic commerce system compared to existing electronic commerce systems. In addition to conventional search engines, the SiteHelper agent acts as a housekeeper for the system and as a helper for the user to find relevant information. In order to assist the users to find relevant information at the centralized system (with Web links to the global Web), SiteHelper interactively and incrementally learns about each user's areas of interest and aids them accordingly. To provide such intelligent capabilities, SiteHelper deploys data mining techniques with incremental learning facilities as its learning and inference engines for incremental exploration of the electronic commerce system.
- *Document Updates.* Given the rapid evolution nature of Web documents, we design a document agent, DA, to check updates of the Web documents linked in the electronic commerce and report new entries from authorized users and the RA. The updates include revisions, insertions and removals.

Our electronic commerce system is centralized with WWW links to documents physically located on the Web, and provides intelligent agents that act as a housekeeper for the system and as a helper for the users to find relevant information. This approach begins from the centralized view of a conventional commerce system, seeks to provide access to the electronic commerce through digital means including the WWW, and maintains the advantages of decentralization, rapid evolution, and flexibility of the WWW.

2 Agent Structure with SiteHelper

The World Wide Web is rapidly becoming an "information flood" as it continues to grow exponentially [3,13]. This causes difficulty for users to find relevant pieces of information on the Web. According to the Internet Domain Survey by [17], the Internet has grown from only 617000 hosts in October 1991 to over 43 million hosts in January 1999 and in excess of 4 million Web servers in April 1999. It has been predicted that the number of connections on the Internet will exceed the number of people of the world by the turn of the millennium [6]. The amount of information on the Web is immense, and with such a speed of growth, the Internet and the Web have become a place of anarchy and chaos [8]. For Web users to use the Web and electronic commerce systems productively, they require

[1] A document in our electronic commerce system is a Web page with information for a product or a set of similar products.

better and smarter software like intelligent agents with AI capabilities to assist them. In the past 40 years, AI has found applications in many different domains, however, the systems built are mostly either very narrow or very brittle. The Web is an ideal environment for AI [9] to provide problem solving techniques, and support for users with methodologies like knowledge representation and data mining.

Commercial sites like Lycos, AltaVista, and many others are search engines that help Web users find information on the Web. These commercial sites use indexing software agents to index as much of the Web as possible. However, the enormous growth of the Web makes these search engines less favourable to the user because of the large number of pages they return for a single search. Thus it is very time consuming for the user to go through the list of pages just to find the information. To remedy this problem, many researchers are currently investigating the use of robots (or "spiders", "Web wanderers" or "Web worms") that are more efficient than search engines. These robots are software programs that are also known as agents, like WebWatcher, Letizia, CIFI, BargainFinder, Web Learner, Syskill & Webert, MOMspider and many others. Some of theses agents are called intelligent software agents [12] because they have integrated machine learning techniques. The Web page titled "Database of Web Robots Overview" at `http://info.webcrawler.com/mak/projects/robots/active/html/` lists 230 of these robots or agents as of July 20, 2000.

The advantages of the robots are that they can perform useful tasks like statistical analysis, maintenance, mirroring and most important of all; resource discovery. However, there are a number of drawbacks. Robots normally require considerable bandwidth to operate, thus resulting in network overload, bandwidth shortages and increase in maintenance costs. Due to the high demand of robots, network facilities are required to be upgraded - consequently resulting in budget increases. Robots generally operate by accessing external servers or networks to retrieve information, raising ethical issues as to whether people should improve their system just because too many robots are accessing their sites. Koster mentioned in his paper "Robots in the Web: threat or treat?" that a robot visited his site using rapid fire requests and after 170 retrievals from the server, the server crashed [5].

With these drawbacks of Web robots in mind, this paper designs an alternative way to assist the user in finding information in our centralized electronic commerce system (with Web links to the global Web) using incremental machine learning techniques. A software agent named SiteHelper is designed to act as a housekeeper for our electronic commerce system and a helper for the user to find relevant information on the system server. In order to assist the Web user to find relevant information at the centralized system site, SiteHelper interactively and incrementally learns about the user's areas of interest and aids them accordingly. To provide such intelligent capabilities, SiteHelper deploys data mining techniques with incremental learning facilities as its learning and inference engines.

2.1 Agent Architecture

SiteHelper in our electronic commerce system aids the users by learning about the users' interest and preferences through the generation of rules about the users. The rules are refined and improved through two learning processes: interactive incremental learning and silent incremental learning. SiteHelper first learns about a user's areas of interest by analyzing the user's visit records, and then assists the user in retrieving information by providing the user with update information about the electronic commerce.

Interactive incremental learning functions in cycles that interact with the user. SiteHelper prompts the user with a set of keywords in logical expressions and related documents which are likely of the user's interest, and asks for feedback. Considering the feedback SiteHelper makes changes to its search and selection heuristics and improves its performance.

Many Web servers implement a log file system that records user access information to their sites. The log files normally consist of the computer name and location on the Internet, the time of access and accessed Web pages. We implement such a log file system in our electronic commerce system. Silent incremental learning uses the log information as its starting point. SiteHelper extracts a log file for each user. From the log file, SiteHelper learns about the user's interest areas. It extracts a set of keywords in logical expressions about the user's interest areas according to the documents the user has visited in the past.

SiteHelper works differently from search engines and other kinds of agents like WebWatcher and World Wide Web Worm that help the user on the global Web. However, other Web sites can deploy SiteHelper to assist users in finding information in the same way as in our electronic commerce system. This design of SiteHelper avoids the drawbacks of existing robots. In addition, there are other advantages of having SiteHelper at a local Web site (like an electronic commerce). First, through incremental learning of the user's characteristics or interest areas, SiteHelper becomes an assistant to the user in retrieving relevant information. Second, SiteHelper has the potential to reduce user accessing and retrieval time, by displaying a list of changes that have been made since the user's last visit. Finally, SiteHelper can be easily adopted for other Web sites or electronic commerce systems.

Figure 1 shows the design of our electronic commerce system structure with the SiteHelper. The relevance verification agent (RA) and document updates agent (DA) have been mentioned in Section 1, and therefore we will concentrate below on how SiteHelper is designed.

- **Access Log.** Most Web sites allow global user access and have logging facilities in place [11] to record users' access details. The access log of a Web site records all Web transaction/request services by the Web server. The three main elements for each record are: the machine name with its Internet address from which the access is performed, the date/time of access and the Web page being accessed. We implement these facilities in our electronic commerce system.

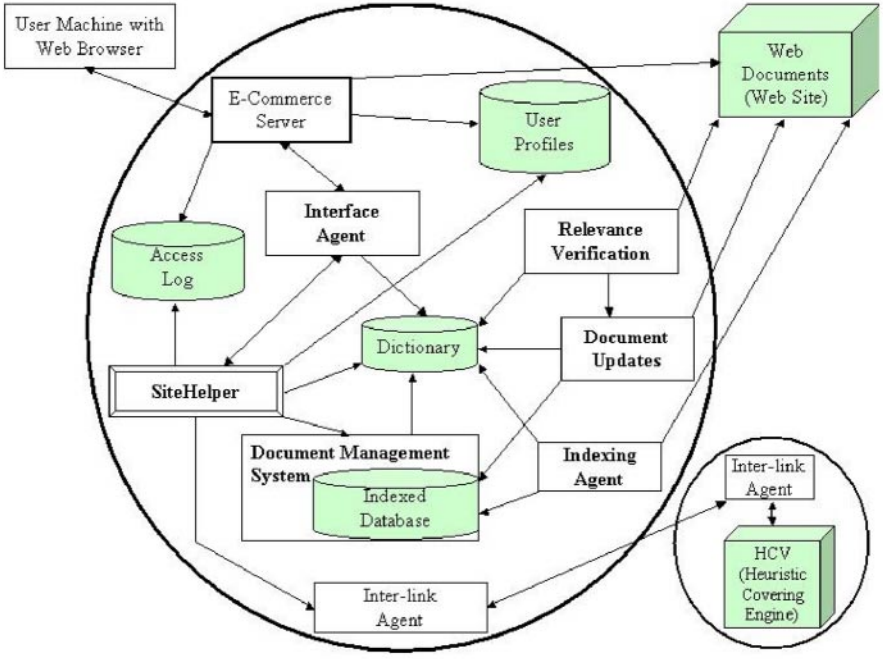

Fig. 1. Agent Structure with SiteHelper

- **Dictionary.** The dictionary is a list of keywords organized in a hierarchy that define the scope of the electronic commerce. It is used by various agents including the relevance verification agent (RA) mentioned in Section 1.
- **SiteHelper,** also referred to as the discovery agent. SiteHelper manages major communications and tasks of the electronic commerce system. In addition to conventional keyword-based search mechanisms, it reads the access log to collect records of documents that a user has accessed, and through the interface agent and the document management system, retrieves all the documents in the indexed database that the user has not visited. It also matches the records with the user profile to compile a set of references to document objects and passes them to the inter-link agent.

Of course, a user might access an electronic commerce for different purposes at different visits. If this is the case, the user can set up the user preferences (see the interface agent below) to stop SiteHelper from providing new and updated documents at some visits.

- **Inter-link Agent.** The inter-link agent forwards the system indexes from the discovery agent to the HCV engine for rule generation. Our electronic commerce system is designed across two different platforms, Windows NT and Unix. To establish the communication between the two, the inter-link agent is designed to run on both platforms.

- **HCV** [14,15]. The HCV induction engine is the "brain" of the discovery agent. It takes two input sets of documents; one set the user has seen, and the other the user hasn't visited. It generates rules in the form of conjunctions of keywords in the dictionary to identify the user's areas of interest, and forwards the rules to the user profiles.
- **Electronic Commerce Server.** This is the Web server for the electronic commerce system. It transmits information using the Hypertext Transfer Protocol (HTTP), and serves requests from external browsers for documents in the electronic commerce. It is the entry point for a user to use our electronic commerce system, with the user name and password to identify the user. Access details of the user are stored in the user profile.
- **Document Management System and Indexed Database.** The indexed database has an entry/index for each document in the electronic commerce system, and is managed by the document management system. A document here can possibly be a compound document that consists of many other simpler documents. Each entry in the database has a pointer to the corresponding document, with (a) a set of keywords from the dictionary to index the document, (b) a date and time to show when the document was last modified, and (c) a format indicator. The document entries can be nested to allow for compound documents to be searched.
- **Indexing Agent.** This agent traverses the electronic commerce weisite, indexes all relevant documents according to the dictionary and stores the results in the indexed database. It incrementally refreshes the indexes and automatically updates indexing when document updates are forwarded by the document updates agent mentioned in Section 1.
- **Interface Agent.** The interface is what the user sees. It presents a friendly interface that allows the user to interact with the system. The agent provides the user with the following functions: the bookmarking of interesting documents in the electronic commerce, navigation in the system, evaluation of retrieved documents, the setting up of user preferences, and a help system with hyperlinks using the semantic links between the keywords in the dictionary. We also have a facility on the electronic commerce interface for authorized sellers to submit relevant materials directly in electronic format.
- **User Profiles.** A user profile consists of the user's account details, areas of interest, access history, and the rules generated by the discovery agent.
- **User Machine with Web Browser.** A Web user can access our electronic commerce through a Web browser that supports the ActiveX technology, for example, the Microsoft Internet Explorer and Netscape Navigator with the ActiveX plugs-in.

3 Related Work

Assisting Web users by identifying their areas of interest has attracted the attention of quite a few recent research efforts. Two recent research projects reported in [2] and [16] are along this line. Three other projects, WebWatcher [1], Letizia [7] and Web Learner [10] also share similar ideas.

[2] develop a system that helps a Web user discover new sites that are of the user's interest. The system presents the user every day with a selection of Web pages that it thinks the user would find interesting. The user evaluates these Web pages and provides feedback for the system. The user's areas of interest are represented in the form of (keyword, weight) pairs, and each Web page is represented as a vector of weights for the keywords in a vector space[2]. From the user's feedback, the system knows more about the user's areas of interest in order to better serve the users on the following day. If the user's feedback on a particular Web page is positive, the weights for relevant keywords of the Web page are increased, otherwise decreased. Balabonovic & Shoham's system adds learning facilities to existing search engines, and as a global Web search agent does not avoid the general problems associated with search engines and Web robots. In addition, compared to SiteHelper, the (keyword, weight) pairs used in this system cannot represent logical relations between different keywords. This type of logical expressions is the starting point for knowledge representation and data mining with SiteHelper.

[16] investigate a way to record and learn user access patterns in the area of designing on-line catalogues for electronic commerce. This approach identifies and categorizes user access patterns using unsupervised clustering techniques. User access logs are used to discover clusters of users that access similar pages. When a user comes, the system first identifies the user's pattern, and then dynamically reorganizes itself to suit the user by putting similar pages together. An (item, weight) vector, similar to the (keyword, weight) vector used to represent each Web page in [2], is used in [16] to represent a user's access pattern. The system views each Web page as an item, and the weight of a user on the item is the number of times the user has accessed the Web page. This system does not use semantic information (such as areas of interest) to model user interests, but just actual visits. Also, it does not aim to provide users with newly created or updated Web pages when they visit the same Web site again. This is a significant difference in design between this system and our SiteHelper.

WebWatcher [1] is an agent that helps the user in an interactive mode by suggesting pages relevant to the current page the user is browsing. It learns by observing the user's feedback to the suggested pages, and it can guide the user to find a particular target page. A user can specify their areas by providing a set of keywords when they enter WebWatcher, mark a page as interesting after reading it, and leave the system at any time by telling whether the search process was successful or not. WebWatcher creates and keeps a log file for each user and from the user's areas of interest and the "interesting" pages they have visited, it highlights hyperlinks on the current page and adds new hyperlinks to the current page. WebWatcher is basically a search engine, and therefore does not avoid the general problems associated with search engines and Web robots. Although it has been extended to act as a tour guide [4], it does not support incremental exploration of all relevant, newly created and updated pages at a local site.

[2] The vector space approach is one of the most promising paradigms and the best-known technique in information retrieval.

Letizia [7] learns the areas that are of interest to a user, by recording the user's browsing behaviour. It performs some tasks at idle times (when the user is reading a document and is not browsing). These tasks include looking for more documents that are related to the user's interest or might be relevant to future requests. Different from WebWatcher, Letizia is a user interface that has no predefined search goals, but it assumes persistence of interest, i.e., when the user indicates interest by following a hyperlink or performing a search with a keyword, their interest in the keyword topic rarely ends with the returning of the search results. There are no specific learning facilities in Letizia, (but just a set of heuristics like the persistence of interest plus a best-first search), and therefore it does not perform incremental learning as SiteHelper does.

Web Learner [10] is similar to SiteHelper in that it learns about what a user is interested in and decides what new Web pages might interest the user. However, Web Learner generates keywords (called a feature vector) automatically from pages on the global Web, and does not provide facilities for incremental learning.

Our localized Web agent SiteHelper starts with the same idea of assisting Web users by learning and identifying their areas of interest. However, SiteHelper works with a centralized electronic commerce server which contains indexes to Web pages on the Web by using a keyword dictionary local to the electronic commerce. Further, based on the indexing of the Web pages on and linked to the electronic commerce server, SiteHelper supports interactive and incremental learning. The rules with logical conditions in SiteHelper are more powerful than the (keyword, weight) pairs used in some existing systems in representing users' areas of interest.

SiteHelper is different from existing search engines and robots on the World Wide Web in that it does not traverse the global Web, but acts as a housekeeper for a centralized electronic commerce server and as a helper for the user who visits the electronic commerce to find relevant information, with particular attention to the newly developed and modified documents in the electronic commerce.

4 Conclusions

As the Internet and World Wide Web continue to grow, more and more Web sites and electronic commerce systems are being set up. There have been many national efforts on electronic commerce systems in various countries. Our agent structure with SiteHelper is unique compared to existing electronic commerce efforts, and provides new capacities for electronic commerce systems to serve existing and new user communities.

The electronic commerce structure in Section 2 can be plugged in other electronic commerce systems by revising its keyword dictionary. In addition, the idea of having a localized agent to help the user find relevant information can be applied to many other domains. When a particular user visits an electronic commerce site (whether a specialized or a general system), the site allows the user to search for particular documents (or products). It can then log the user's searches as well as their browsing behaviour. From the log information, Site-

Helper can be used to learn about the user's areas of interest, and at the user's following visit, SiteHelper may prompt the user to look at those new/updated materials that match their areas of interest.

References

1. Armstrong, R., Freitag, D., Joachims, T. and Mitchell, T.: WebWatcher: A Learning Apprentice for the World Wide Web. In: *AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*. March 1995.
2. Balabanovic, M. and Shoham, Y.: Learning Information Retrieval Agents: Experiments with Automated Web Browsing. In: *On-line Working Notes of the AAAI Spring Symposium Series on Information Gathering from Distributed, Heterogeneous Environments*. 1995.
3. Berners-Lee, T., Gailliau, R., Luotonen, A., Nielsen, H. F., and Secret, A.: The World-Wide Web. *Communication of the ACM* **37**(August 1994).
4. Joachims, T., Freitag, D. and Mitchell, T.: WebWatcher: A Tour Guide for the World Wide Web. In: *Proceedings of the 15th International Conference on Artificial Intelligence*. Nagoya, Japan, August 23-29, 1997. 770-775.
5. Koster, M.: Robots in the Web: Threat or Treat? *ConneXions*, **9**(April 1995).
6. Lawrence, A.: Agents of the Net. *New Scientist*, 15 July 1995, 34-37.
7. Lieberman, H.: Letizia: An Agent That Assists Web Browsing. In: *Proceedings of the 1995 International Joint Conference on Artificial Intelligence*. Montreal, Canada, August 1995.
8. Murray, J.: Anarchy and Chaos on the Net. *IEEE Computer*. May 1995.
9. O'Leary, D.E.: The Internet, Intranets, and the Al Renaissance. *IEEE Computer*, January 1997, 71-78.
10. Pazzani, M., Nguyen, L. and Mantik, S.: Learning from Hotlists and Coldists: Towards a WWW Information Filtering and Seeking Agent. In: *Proceedings of IEEE 1995 Intl. Conference on Tools with Al*. 1995.
11. Pitkow, J. E. and Bharat, K.A.: WebViz: A Tool for WWW Access Log Analysis. In: *Proceedings of the First International World-Wide Web Conference*. Geneva, Switzerland, May 1994.
12. Riecken, D.: Intelligent Agents. *Communication of the ACM* **37**(July 1994).
13. Wiggins, R.W.: Webolution: The Evolution of the Revolutionary World-Wide Web. *Internet World*. April 1995, 35-38.
14. Wu, X.: *Knowledge Acquisition from Databases*. Ablex Publishing Corp., USA, 1995.
15. Wu, X.: Rule Induction with Extension Matrices. *Journal of the American Society for Information Science* **49**(1998), 5: 435-454.
16. Yan, T.W., Jacobsen, M., Garcia-Molina, H. and Dayal, U.: From User Access Patterns to Dynamic Hypertext Linking. In: *Proceeding of the Fifth International World Wide Web Conference*. Paris, France, May 1996.
17. Zakon, R.H.: Internet Timeline v4.1. `http://info.isoc.org/guest/zakon/Internet/History/HIT.html`