Development and evaluation of an information retrieval system for user groups and the WWW

MASE Motohiro and YAMADA Seiji
CISS, IGSSE, Tokyo Institute of Technology
E-mail:{mase, yamada}@ymd.dis.titech.ac.jp

Abstract

In this paper, we propose a framework for searching information through both the WWW and a human group. Though the information retrieval using a search engine in the WWW is very useful, we cannot acquire local information owned by a person and not explicitly described in text. A user knows neither where target information is in the WWW nor who knows in a human group. Thus we integrate the information retrieval in the WWW with that in a human group, develop heterogeneous resource information search HERIS as a multi-agent system.

1 Introduction

Information retrieval with the WWW as information resources has spread. The WWW includes huge Web pages and is updating all over the world. Thus the WWW is very useful as information resource. However, it is difficult for a user to investigate the location of information because huge information is ubiquitous disorderly. Generally, we use a search engine for information retrieval in the WWW.

Though the information retrieval with a search engine in the WWW is very useful, we cannot investigate the specific information like configuration of LAN in a laboratory and is personal knowledge like memo of application installation. Such information is not opened to public in the WWW in spite of being useful for other person. We called this information *closed information* contrasted with *open information* like the Web pages in the WWW. Only a part of personal knowledge is opened in the WWW, the most of closed information is not described in a document. To acquire closed information, asking a person who has the information is best. Moreover, results of a search engine often includes non-relevant Web pages, because filtering in accordance with user's intention is not enough. On the other hand, closed information is considered to be filtered by person in advance. Hence searching closed information through a human group is no less important than information retrieval in the WWW. Since judging where target information in the WWW and who knows the target information is very difficult for user, we propose **HERIS**(HEterogeneous Resource Information Search): a framework for searching information through both the WWW and a human group and indicating adequate information resource which should be accessed.

In the matter of sharing personal knowledge on Web pages, sharing URLs in a bookmark file has been studied. Mori and Yamada proposed Bookmark Agent[3]: a multi-agent system consisting of an agent which constructs a profile from a user's bookmark file by analyzing the HTML files. An agent broadcasts a query from user to other agents, and shows the Web pages related to the query to a user. Unfortunately their system deals with only bookmark files, and does not have a mechanism for connection among users.

2 System overview

Fig.1 shows the system overview of HERIS. HERIS is an information retrieval framework that searches both of the WWW and a human group. We build this framework as a multi-agent system consisting of user agents and SE(search engine) agents, and use contract net protocol[1][7] to communication among the agents.

A user agent and a SE agent cover a user and a search engine. Each of the agents manages a user profile and a SE profile. A profile consists of two types of sub-profiles: a knowledge profile and a resource profile.

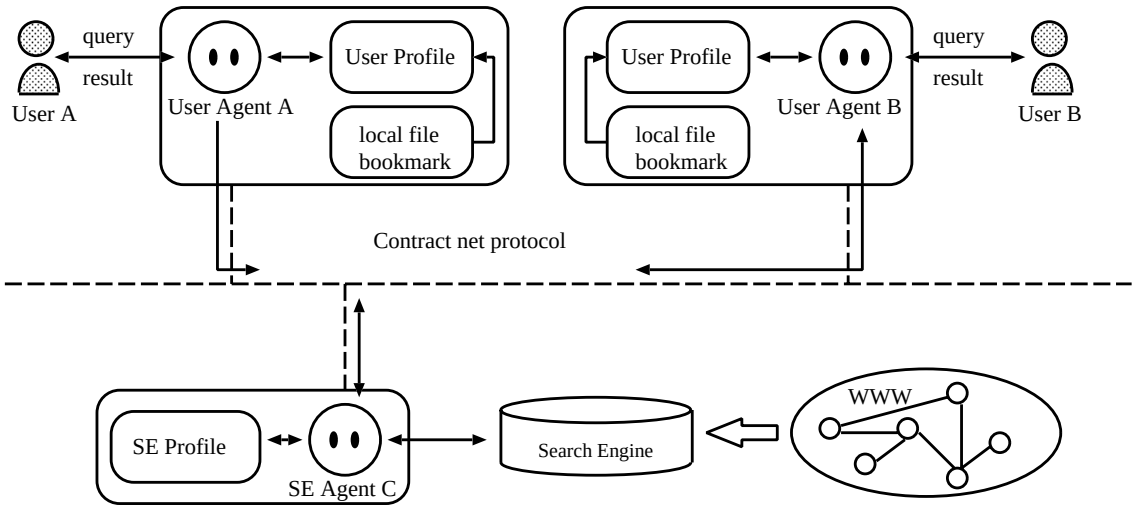

Figure 1: System overview

A knowledge profile contains weighted terms related to knowledge of a user or a search engine. A resource profile contains other properties of an information resource.

A user gives a query HERIS, a user agent that received a query from a user becomes a manager agent and all other agents become contractor agent. A manager agent broadcasts a task announcement message including the received query to contractor agents. Next, contractor agents compute similarity between the query in the task announcement message and their profiles, and reply bid messages including the similarity and resource profiles. After investigating the bid messages, a manager agent selects adequate contractors that win a bid, and send the query as a client message to the selected contractor agents. As a result, a manager agent receives a result message including retrieval information from the selected contractor agents, and integrates the messages. Finally, a user is shown a list of adequate Web pages and person who can answer the query with audio/video applications, and selects information resource from the list.

3 Construction of profile

3.1 User profile

A user profile consists of a knowledge profile and a resource profile.

Knowledge profile A knowledge profile of a user profile stands for what a user know. We consider that user's information consists of two types of the information. The first type is described document like personal memorandum on application install, LaTeXsource file of a paper and so on. The second type is owned by a person and is not explicitly described in documents. We think that the Web pages that a user looks at well usually is the second type of information. Thus HERIS builds a knowledge profile by investigating user's local files and Web pages stored in a bookmark file that a user selects. The following shows the procedure of constructing a knowledge profile.

1. HERIS shows windows for a user to select the local files and Web pages. By selecting files registered into a profile from a local file and Web pages, a user can restrict to reference from other users.

   If local files selected by a user include LaTeXsource files, they are translated into HTML files using latex2html command. The set of these preprocessing local file is D_l. HERIS actually fetches the selected Web page on-line. The set of HTML files is D_h. HERIS sets D_l and D_h to a document set D in all.

Table 1: Weights of HTML tags

tag	weight	tag	weight	tag	weight
<title>	10	<meta>	10	<h1>	10
<h2>	8	<h3>	6	<h4>	4
<h5>	2	<h6>	1	<big>	2
	2	non tag	1		

2. HERIS uses TFIDF[6] and the weighting using the structure of HTML tag for extraction of term and weight from D. In extracting terms and frequency from D, if a document is a HTML file, frequency is weighted according as tag. A set of tags and weights is shown in Table.1. Terms and weights are computed by applying TFIDF.

3. HERIS sums the weight for each term in all documents. All set of term and weight is a knowledge profile of a user profile. Finally, HERIS shows a profile editor window. A user can add/delete terms and edit the weights in this window.

Resource profile A resource profile consists of three elements, presence, cognitive load and social relation. Presence shows whether user is at his/her desk. Cognitive load shows user's busyness. Social relation is computed with ontology among users in their group.

3.2 Search Engine profile

A search engine profile consists of a knowledge profile and a resource profile. A knowledge profile is built by the method used in MetaWeaver[4]. A resource profile has an element, network load. Sending ping command to a search engine and measuring response time investigate network load.

4 Selecting adequate information resources

4.1 The computation of the similarity between the query and profile

We applied a vector-space model[5] to calculation of a similarity between a query and a profile. A similarity is computed as a cosine coefficient of document vector of a query and a profile. When a similarity exceeds a threshold, a bid message is transmitted to a manager agent.

4.2 Selecting the contractor agent

A manager agent selects adequate contractor agents as follows. A manager computes evaluations for contractor agents with weighted linear sum of a similarity and elements of a resource profile, chooses higher rank of contractor agents with high evaluation and requests them to send their result messages.

After receiving the result messages, a manager agent integrates the messages and indicates a result like Fig.2. A list of users who can answer the query is displayed on the left-hand side of this window, and a hit list of Web pages is displayed on the right-hand side. When a user selects a person from the list, if a selected person accepts a user's request, a user can talk with the person with audio/video applications. Fig3 shows the communication established between the users.

5 Experimental evaluation

We conducted two types of experiments for evaluating HERIS. For the following experiments, we used eight user agents and a single SE agent for Google. We implemented HERIS using Ruby and GTK+, and used vat and vic[2] for audio/video communication between users.

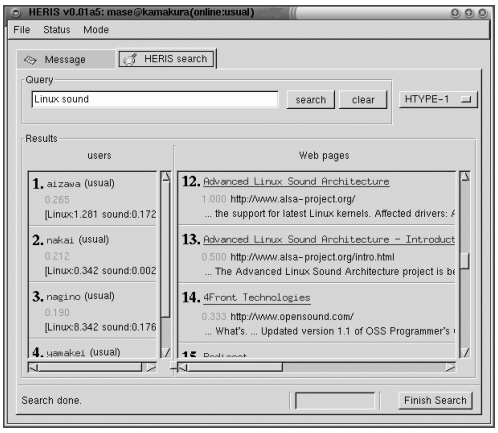

Figure 2: Indication of retrieved results

Figure 3: Communication among users

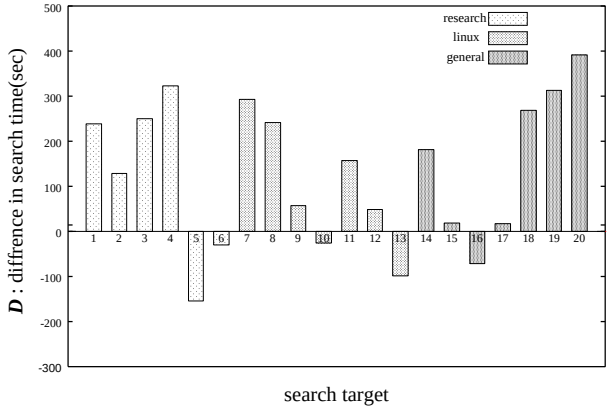

Figure 4: Difference in search time

5.1 Comparison with a search engine

We compared HERIS with a search engine Google. We used two targets for each domain, "research", "linux" and "general", and investigated the difference in search time until relevant information was found. The results are shown in Fig.4. The equation for computing D(difference in search time) is D = (search time of a search engine) - (search time of HERIS). The search time of HERIS is less than that of a search engine, if D is positive value. We see from Fig.4 that it was fifteen searches that the search time of HERIS is less than that of a search engine among twenty searches, and a feature by domains is not seen. We consider that the difference in search time is due to the difference in features of human and search engine as the information resource. A result of a search engine often includes many non-relevant Web pages because it is difficult for users to describe intention in adequate query. Thus finding relevant page takes much time.

However when asking a person directly, a user can tell his/her intention in a conversation with a person who may have the target information. Therefore a user can find the target information easily. Moreover a user may receive advice about more adequate query filtered by a person.

5.2 The feature of search with HERIS

The feature of HERIS is to search the target information through both the WWW and persons. We investigated situations in which person would be accessed. Also we recorded the following logs for every search with HERIS. The recorded logs are "query", "time accessed to the information resource", "order accessed

to the information resource" and "comment about acquired information and situation". From analyzing the recorded logs, three types of search process were found. At first, without browsing Web pages at all, a user accessed a person first. Next, although a user browsed Web pages, since the target information was not found or a part of the information was found, a user accessed a person at last. Finally a user accessed a person while browsing Web pages, acquired an adequate query from a person, and searched with a new query again.

As an absorbing instance, some persons searched with a query of the higher concept, without a direct query related to target information. For instance, when a user searched what is sound driver in VineLinux, a user used a query that was not "VineLinux, sound driver" but "VineLinux", and asked a person indicated by system. In this case, user may ask a person who does not have target information. However, if a person asked by a user has target information, a user may be able to retrieve the information that is not registered in a knowledge profile and not described in a document. The same query in search with a search engine causes useless search results against user's intention. This technique of search is characteristic in HERIS searching for a human group.

6 Conclusion

We proposed HERIS: an information retrieval framework that searches open information and closed information through both the WWW and a human group. We developed the information retrieval system as a multi-agent system consisting of user agents and SE agents. Profiles for each person and each search engine are built, and the system selectively searched information resources consisting the WWW and a human group. As a result, a user acquires the results of hit lists from the WWW and direct audio/video connections to a person who can answer the query. It was found through experimental evaluation that the information retrieval using a human group as information resources is valid and a search with HERIS has a character different from conventional search engines.

References

[1] R. Davis and R. G. Smith. Negotiation as a metaphor for distributed problem solving. AI Memo 624, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, May 1981.

[2] S. McCanne and V. Jacobson. vic: A flexible framework for packet video. In *The Third ACM International Multimedia Conference and Exhibition (MULTIMEDIA '95)*, pages 511–522, New York, November 1996. ACM Press.

[3] M.Mori and S.Yamada. Bookmark-agent: Information sharing of urls. In *The 8th International World Wide Web Conference(WWW-8)*, pages 70–71, 1999.

[4] M. Mori and S. Yamada. Adjusting to specialties of search engines using metaweaver. In *WebNet 2000 World Conference on the WWW and Internet*, pages 408–412, 2000.

[5] G. Salton. *Automatic Text Processing*. Addison-Wesley, 1989.

[6] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Tokio, 1983.

[7] R. G. Smith. The contract net protocol: High-level communication and control in a distributed problem solver. In *IEEE Transaction on Computers*, number 12 in C-29, pages 1104–1113, 1980.