Intelligent Search in Digital Libraries

Nupur Prakash

Abstract

The paper explores the role of an intelligent software agent as a mediator in information retrieval (IR). Software agents are autonomous programs which perform a specific personalized task in a heterogeneous and distributed environment. These autonomous, intelligent agents make decisions on behalf of the user, by narrowing the search domain and decreasing the human computer interaction, phenomenally. With the increasing number of institutes and universities across the country and globe it has become very important to share library services rather than duplicating the library facilities everywhere and incurring huge expenditure. In the present paper, the architecture for a digital library DIGLIB is proposed, which helps a user or a group of users identify and find reading material in a virtual or brick and mortar library of his or her interest. DIGLIB uses a software agent which is a unique combination of filtering and information agent to facilitate intelligent search. The information agent performs a general search on various sites rigorously based on a keyword and phrases. The short listed sites are further subjected to a specific search by the filtering agent. The information agent performs a breadth first search on the links ahead and also provides navigation recommendation. A rule-base is maintained by the filtering agent using premises like broad area of interest, specific area of interest, type of article, book, magazine, journals, periodicals etc. to be searched for further narrowing the search domain and suggesting URNs of digital libraries of user's interest. An attempt has been made to design a digital library interface using an intelligent agent which continually runs in the background and facilitates the task of finding a suitable match between the desired library service and user's requirement.

Keywords : Information Retrieval, Software Agents, Intelligent Search, Distributed Environment

0. Introduction

The Internet can be treated like a library without a card catalog, which can be searched using an intelligent search engine. The information on the Internet is stored in many repositories around the world which are getting interconnected by computer networks.

Research on Digital libraries revolves around exploiting the increasing computational capacity and network bandwidth for transforming a very large amount of distributed complex data into*information* and *knowledge.* For this purpose, lot of work is being done in the following areas [1].

(a) Intelligent user interfaces

- To support intelligent search, retrieval, organization and presentation of information in digital libraries.
- To design and deploy intelligent software agents to support personalized needs, choices and preferences of the library user.
- To allow simple and user-friendly access to the system in a distributed environment.

(b) Access to Content and Digital Collection

- Using efficient data capture techniques involving text, voice, images and graphics.
- Intelligent text processing using natural language processing.
- Gathering, aggregating and establishing relationships among organizing knowledge sources.

(c) System Deployment for digital Libraries

- Involves designing hardware and software systems capable of interpreting user's request by locating, filtering and querying information in a heterogeneous distributed environment.
- Design of new protocols for high bandwidth applications and metadata services over a networked environment.
- Improving quality of service and deployment of payment models for e-subscription of library services.
- Ensuring interoperability, scalability and extensibility of library systems across different institutes and universities
- Adapt and evolve to changing environment

The information available on the Internet is increasing, exponentially. For a given query, the search engine displays a large number of potentially interesting sites, however, only very few sites are of some relevance. Furthermore, the information available on the net is highly heterogeneous in nature. Different data types are used for different websites like textual data, images, sound, video clipping etc., each following a different representation and compression technique. The biggest challenge is to :

- Search for, analyse and integrate data in a heterogenous environment
- Update the user on the availability of data of special interest from time to time
- Adapt and evolve to changing environment

Searching for reading material on unconventional subjects like Marine Engineering., Mining, Buddhist studies , Aeronautical/Aero space Engineering, etc can be very time consuming and difficult. On the contrary, when there is too much information available on overlapping subjects like Information Technology, Computer Science and Engineering, Electronics & Communication etc., the user becomes confused and needs help in taking decision [2].

In DIGLIB the intelligent agent enables the user to narrow down the search domain on libraries that best meet his/her needs.

1. What is an Intelligent Agent ?

With the increasing number of institutes and universities, the library user's are growing day by day and so finding relevant information on the net is becoming difficult. There is a great need for tools to assist users, while searching the digital libraries, which is flooded with enormous amount of information. An intelligent software agent is a program which searches across heterogeneous and geographically distributed information sources using artificial intelligence(AI) techniques, for a relevant piece of information. These agents improve on the searching capability and performance of any search engine by submitting queries to many different engines, simultaneously. Some agents also keep track of user's profile, his personal interest and favourite websites by maintaining his past history [3].
Most intelligent agents :

- Keep track of users special interests
- Expedite the search process by narrowing the search domain
- Integrate information in an intelligent manner in a heterogeneous and dynamic environment
- Provide information on demand, on-the-fly, just-in-time.
- Provide transparent access to different website on the relevant topics

Agent based intelligent search uses a mix of AI techniques like depth first, breadth first and best first search where the user determines the depth and breadth cut off levels [4,5].

2. Types of Agents

There are various types of agent which perform different tasks. These agents can be categorized based on the nature of work performed by them [6]. Most agents are adaptive and mobile in nature.

2.1 Personal Assistants

These agents are used to monitor the user interaction with any application and provide active assistance to the user in the process of finding and organising information. Agent based information retrieval systems are different from conventional information retrieval systems, as they interact with the user and deal with the dynamic nature of information. The personal assistants prompt the user at different levels and suggest number of options available. These agents learn and adapt, based on explicit user feedback through penalty/reward learning strategy. They can be trained by the user through supervised learning using neural networks [7].

2.2 Filtering Agents

The information available on the internet is enormous. Therefore for a given query, numerous URLs are displayed but very few are actually relevant to the user. Filtering agents can be used to sieve relevant information while blocking the flow of useless and undesired information [8],

The filtering agents can be transplanted by storing the user profile, his areas of interest, likes and dislikes. During search process a filtering agent scans the documents in a database and ranks them in the order of users interest levels. Keyword based search is generally used by these agents which involves matching different combination of keywords. An advanced form of filtering, extracts semantic information of the documents content. An adaptive filtering agent has learning abilities and automatically adapts to the user's interest profile. This type of agent studies the list of favourite items, the documents saved and deleted by the user from time to time and accordingly refines or broadens the filtration procedure.

2.3 Information Agents

These agents perform rigorous search operation for finding relevant information across the web. It scans through on line databases and document libraries looking for information that might be of user's interest. These type of agents help the user in keeping his knowledge up to date by displaying names of new websites containing latest information, related to his area of interest.

Generally, Information agents are mobile in nature as they are able to travel autonomously through the networked environment. Mobile agents can be transported to different sites in the Internet and provide data access in heterogeneous environment. These agents are generally written in interpreted machine independent language like Java so that it can extract information in heterogeneous environments from a remote database server, locally. These agents are very useful in wireless networks as they allow the user to access any kind of information any where and anytime [9].

2.4 Brokering agents or mediators

Brokering agents play variety of roles as mediators in electronics commerce. These agents can perform product brokering through product comparisons and merchant brokering by comparing merchant alternatives [10]. The agent acts as a broker taking requests from a buyer or seller and searches for a set of potential and suitable sellers or buyers using consumer buying behaviour model [11]. Once the potential merchant or a buyer is found, that can satisfy the user's request the broker agent returns the results to the user for transaction or automatically executes the transaction on behalf of the user. The brokering agent some times uses a natural language interface and accepts requests from the user in natural language (e.g." I want to buy a camcorder which is compatible with my VCR and costs below $ 500")

2.5 System Management Agents

These agents manage the operations of a computing system or data communication networks by monitoring device failures, link failures, system overloads etc. These agents are also capable of load balancing by redirecting works to other parts of the system in order to maintain a reasonable level of performance and reliability. These agents are very useful for distributed environment and network management [12]. System management agents are designed to behave in a proactive manner responding not only to specific events but initiating pre-emptive activities within the systems like a perfect system manager/administrator.

3. Architecture of DIGLIB

Although Internet can be treated like a library without a card catalogue, which can be searched using an intelligent search engine, yet, it cannot serve the purpose of a full fledged digital library without a proper architecture. Digital libraries need

- Richer document model where by each object in the library (articles, journals, books, magazines) is available as a digital objects.
- Uniform Resource Names (URNs) to allow persistent, globally unique identifier for each library resource.
- Well defined digital library services like Cataloguing, Searching, Indexing, Retrieving, Browsing, Subscription services, Revenue collection (e.g. collection of membership charges), down loading (copying), payment of royalties and copyright fee, presentation of information in proper form.
- Better facilities for resource discovery in a distributed environment.
- Management of distributed content because the information in stored in many repositories around the world.
- Identity and privacy of users.

In order to achieve the above objectives a digital library architecture, DIGLIB is proposed which provides a user friendly environment, uses software agents to facilitate intelligent and personalized search and collects revenue on line. The proposed architecture is depicted in figure 1.

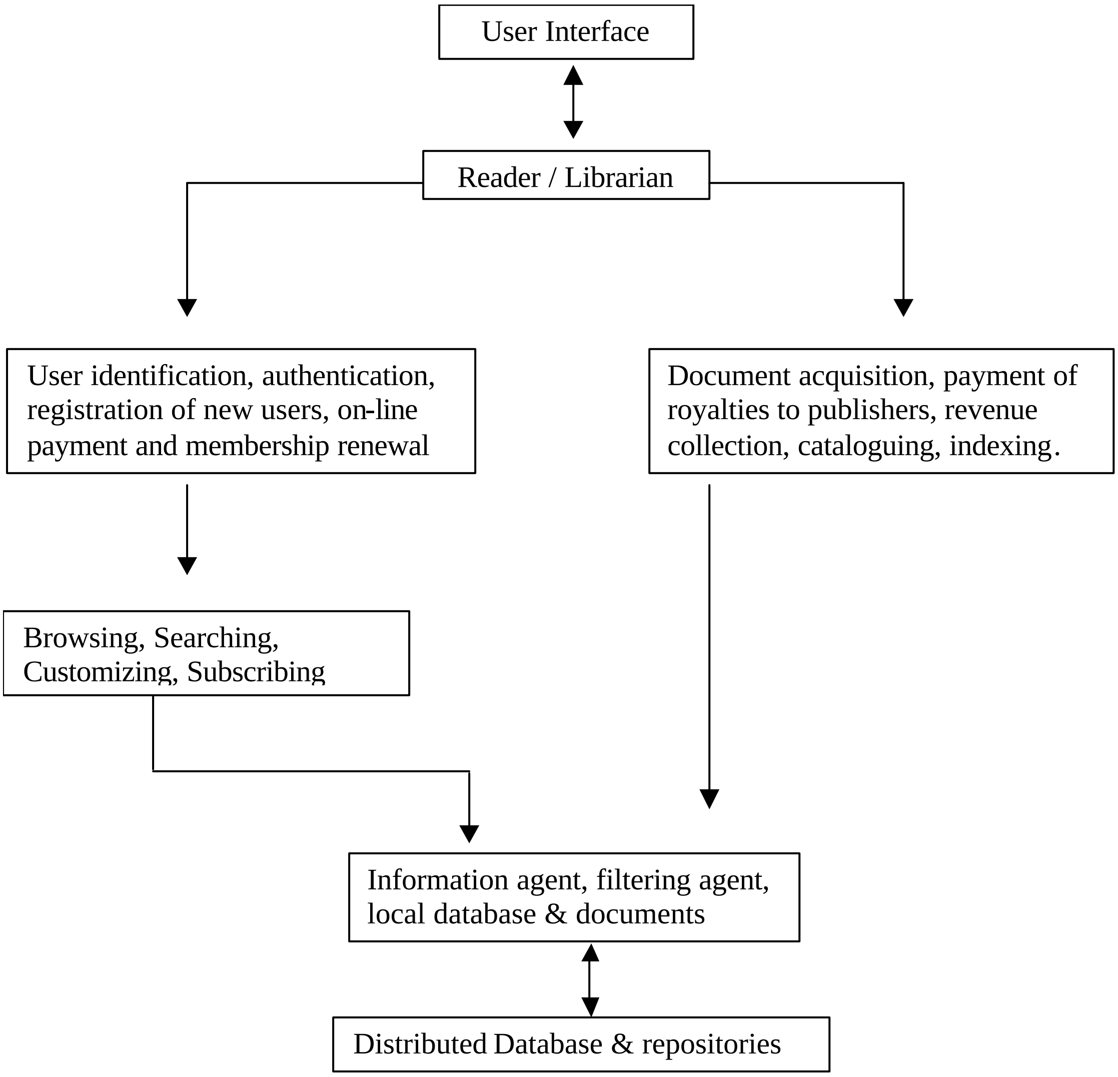
User Interface
Reader / Librarian
User identification, authentication, registration of new users, on-line payment and membership renewal
Document acquisition, payment of royalties to publishers, revenue collection, cataloguing, indexing.
Browsing, Searching, Customizing, Subscribing
Information agent, filtering agent, local database & documents
Distributed Database & repositories

Figure 1. Architecture of DIGLIB

The software agent, designed for finding relevant reading material, is a blend of filtering and information agents. The overall behavior is similar to that of a brokering agent. It searches the entire database available on the net, as well as member institutions of the library and filters all the relevant articles.
DIGLIB is composed of the following modules :

1. **The Graphics User Interface (GUI)** is Java enabled WWW browser to establish the

communication between user and the library.

2. **Mechanism for user authentication** to check access rights of the user and register new users. The identity of each user is established by providing individual name (user-id / password) at the time of registration. Only authorized users are allowed to access resources e.g. e-journals, articles, research papers, PhD thesis etc.

3. **Built in Mechanism for e-subscription** of library services may be provided using on line payment of subscription fee and renewal of membership.

4. **The information agent** which accepts query from the user for searching the library, once the user is successfully logged in to avail library services for which he is authorized. A unified list of all possible URN's is created using keyword matching (i.e. broad area of research e.g. computer engineering) and stored in a database maintained at client side.

5. **The database** where the URNs of all the retrieved sites are stored.

6. **The filtering agent** performs a higher level of abstraction, by maintaining a rule base and set of attributes stored for an individual user to further narrow down the search domain.

In this manner, several processing levels are created between the user and the information available in the libraries to facilitate intelligent search.

4. Implementation

DIGLIB provides a personalized acquisition and cataloguing system based on intelligent software agents. The user's profile and history about reading material referenced recently by the user is maintained at client site, which is updated from time to time. DIGLIB runs on a server which hosts the filtering and information agents. The GUI runs at the client site to control DIGLIB which is built around Java development environment. The user is presented with a GUI based screen once he is successfully registered with DIGLIB. When the user connects to DIGLIB the intelligent agents are activated in the background and start monitoring user's attributes (e.g. his interests, preferences, reading - habits and choices). Once he enters the query to find the reading material relevant to his work the information agent extracts keywords from the query and performs a general search on the distributed database. The links of desired material and URNs are returned to the local database.

- ✍ If the user wants to refine search by way of specifying the narrow area of interest e.g. 'Low power CMOS design' then he /she may further submit his/her request to the filtering agent.
- ✍ The filtering agent checks through the local database using Boolean operators like OR, AND, NOT for search refinement and returns the selected list of items for reader's reference.

The filtering agent also accepts user feedback for refinement of search operation by suggesting URNs and displaying abstracts on demand. The user can check out the contents and rate it according to his/her preference on a (0-10) scale. This rating is used as a feedback to the filtering agent responsible for selecting that document [13]. This improves the overall filtration process making it more accurate. User's selection criteria may keep on changing and the agent keeps track of user's attributes and history file using feedback mechanism.

The agent provides the user with various search strategies by which the user can get to the information required easily and efficiently. The user can perform either general search (i.e. for area of interest (e.g. Information Technology) using information agent or specific search using refine option through filtering agent. The entire module has been implemented using JSP, HTML and Java [14]

5. Comparison of conventional IR vs. Agent based IR

Agent based Information Retrieval (IR) differs from conventional IR in terms of user interaction and dynamic nature of information spread over heterogeneous distributed environment. Conventional IR techniques are designed for relatively static databases, concentrated in a single geographic location using some form of schema or pattern. However, information on the web is generally distributed, unstructured and may often contain graphical (non textual) information. Therefore the search engines and agents generally use :

i. IR Keyword frequency analysis technique for document matching and search. Term frequency times inverse document frequency (TFIDF) is often implemented for arriving heuristically at relevant subject matter [15].

ii. Clustering technique called Latent Semantic indexing [16] which requires more computation than simple keyword frequency analysis but deals more effectively with synonyms and minor variations of keywords.

DIGLIB applies a modified TFIDF analysis to each page by adding the list of weighted keywords to the user profile. The Inverse Document Frequency (IDF) is calculated dynamically over the user's browsing history. Unlike conventional IR systems, agent based IR systems remain continuously active in the background and try to gather relevant information even without the user's explicit command. These systems continuously learn over a period of time and therefore due to dynamic nature of information the same query expressed at different time may have different results.

6. NCSTRL, A working Example

The Networked Computer Science Technical Reference Library (NCSTRL) is a globally distributed digital library with more than 120 participating Universities and Institutions across the globe located in US, Europe and Asia{17f. The research report of various Universities and their lab work is made available along with D-Lib Magazine. The European Research Consortium for Informatics and Mathematics (ERCIM) along with Los Alamos comprising Physics preprints, ACM to etc. are part of this Technical Reference Library.

More than forty independent servers are used as Resource Providers across the globe. Multi-formatted documents can be accessed in distributed repositories. Documents are accessible through their URNs. NCSTRL supports a modular system based on a standard open architecture. The number of participating institutions are growing rapidly.

7. Conclusion

The agent has to deal with uncertain, incomplete and vague information in an efficient manner, while making intelligent decisions on the fly. However, the success of DIGLIB depends on the net-enabling of various libraries across the country/globe. At times information retrieval becomes very complex due to ambiguous content, complex goals, changing environment and non-standardized storage format.

More features can be incorporated in DIGLIB by understanding the needs of a user, his qualification, experience and his areas of interest. Further improvement is needed for effective interaction between user and information stored in a semi-structured form using multimedia data. A voice and natural language interface may further enhance the usability of the system.

8. References

1. H. Chen, J. Yen, and C.C. Yang, International Activities: Development of Asian Digital Libraries, IEEE Computer, Special Issue on Digital Libraries, vol.32, No.2, February, 1999,p.48-49.
2. Jason T.L. Wang, Chia-yo Chang, Fast Retrieval of Electronic Documents in Digital libraries, In Proc. Seventh International Conference on Tools with Artificial Intelligence (Herudon, Virginia, USA, 1995), 208-215.
3. W. Brenner, et al., Intelligent Software Agents (Springer Publications, Germany, 1998).
4. Munoz, G., et al.: Virtual Reality and agents in a digital library, Second European Conference on research and advanced technology for Digital Libraries (Crete, Greece, Sept.,1998), LNCS 1513 Springer, 681-682.
5. C.C. Yang, J. Yen, and H. Chem, "Intelligent Internet Searching Agent Based on Hybrid Simulated Annealing," Decision Support Systems, vol.28, no.3, May, 2000,p.269-277.
6. M.Dhamyanthi, M. Ponnavaikko, Intelligent Agents and Automated Data Mining, Proc. Seventh International Conference on Advances in Computing and Communications ADCOMP-99, Roorkee, India, Dec. 1999,163-166.
7. Henry Lieberman, Personal Assistants for the Web, in Matthias Klusch (Ed), Intelligent Information Agents (Springer Publications, Germany, 1999).
8. P. Baclace, Competitive agents for information filtering, Communications of the ACM, Vol. 35(12), 1992.
9. M.Barbuceanu, and M.S. Fox, The information agent: An infrastructure for collaboration in the integrated enterprise, Proc. Meeting on Cooperating Knowledge based Systems, Kelee, UK, l994.
10. M.Jeusfeld and M.P. Papazoglou, Information Brokering, M.P.Papazoglou(Ed), Information Systems Interoperability, (Somerset, England,1998) 256-302.
11. J.Engel, and R. Blackwell, Consumer Behavior, 4 th edition (College Publishing,1982). [12] M. Huhns and M.P. Singh.(Eds.), Readings in Agents (Morgan Kaufmann, San Fransisco, 1998).
12. Moukas Alexandras et al., Amalthaea and Histos: Multiagent Systems For WWW Sites And Reputation Recommendation, in Matthias Klusch (ed), Intelligent Information Agents, (Springer Publications, Germany, 1999).
13. Janice Winsor and Brian Freeman, Jumping Java Script (The SunSoft Press, Java Series, 1997).
14. G. Salton, Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer (Addison Wesley, 1989).
15. G.Salton and M.J.Mcgill, Introduction to Modern Information Retrieval (McGraw Hill CS Series, NY, 1983).
16. Payette Sandra, Digital Library Architecture: A Service based Approach, http://www2.cs.cornell.edu/payette/presentation s/DL-architecture.ppt

About Author

Prof. (Mrs.) Nupur Prakash is working as Principal in the School of Information Technology,GGS Indraprastha University. She received her B.E & M.E (Computer Science & Technology) from UOR, Roorkee in 1981 and 1986 respectively. She completed her Ph.D (Engineering & Technology) from Punjab University, Chandigarh in 1998 in the areas of Neural Networks and Natural Language Processing. Before joining GGS Indraprastha University, she has been serving the Department of Computer Science & Engineering at Punjab Engineering College, Chandigarh. Her major areas of interest are Computer Graphics, Computer Communication Networks, Operating Systems etc. Sha has published 18 papers in various national and international level conference proceedings and journals. She is also member of Governing Board of INFLIBNET