DESIDOC Bulletin of Information Technology, Vol. 16, No. 4, July 1996, pp. 13-18

Some AI Techniques for Information Retrieval

KK Shukla*

ABSTRACT

Efficient retrieval of information from the vast sea of distributed databases over the Internet has assumed great importance with the convergence of computer and communication technologies. This paper presents some state-of-the-art techniques for online information retrieval over the information superhighway. It further advocates the application of Artificial Intelligence (AI) search techniques to locate and disseminate information in an effective manner. Content based image retrieval can also benefit from AI techniques by utilising evolutionary programming for feature extraction and object recognition.

1. INTRODUCTION

At the beginning of a new project a researcher has to conduct a comprehensive literature survey by looking through bibliographies, contacting experts, reading conference proceedings and university technical reports. In the context of the present day information explosion in virtually all disciplines, this process is painstakingly slow, and hopelessly inefficient.

Today, the use of Artificial Intelligence (AI) based online information retrieval tools has become indispensable. These tools can help accelerate the process of locating the desired topic in the vast sea of knowledge distributed geographically among thousands of servers around the globe. The Internet can be used as the basis for designing tools that enable browsing, searching and organising information distributed throughout a large number of databases. In this paper we discuss state-of-the-art technologies and provide some guidelines for the design of new intelligent search techniques.

* *Dr KK Shukla*
Reader
Department of Computer Engineering
Institute of Technology
Banaras Hindu University, Varanasi-221005

2. WIDE AREA INFORMATION SERVER

The Wide Area Information Server project (WAIS) is aimed at providing uniform, easy to use, location transparent mechanism to access information. As shown in Fig. 1, WAIS clients translate user queries into the WAIS protocol and query the Directory of Services for relevant databases. WAIS then transmits the request to selected set of databases over the communication network.

The database servers maintain inverted indexes on document contents and can execute full text searches on them. On the arrival of a new query the server returns a list of relevant object descriptors that correspond to documents that contain words specified in the user query. The client then displays query results and can receive documents from

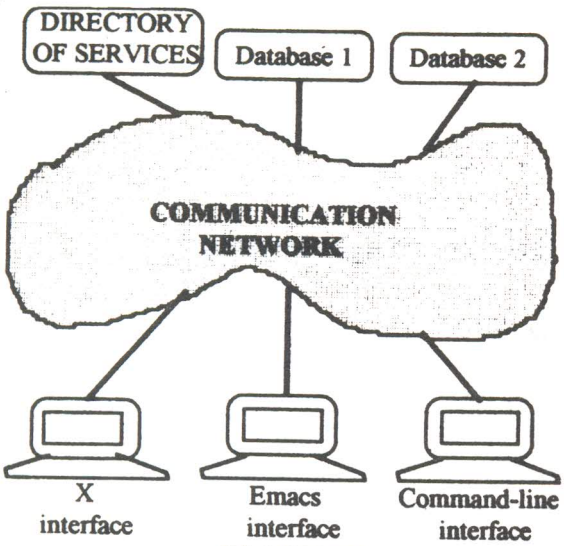

Figure 1: WAIS Architecture

appropriate servers irrespective of their location. Usually the software also displays a numerical score for each hit that relates to the frequency of specified words in the retrieved object's contents. This acts as a feedback to the user and helps him refine his future queries. When a new information provider wants to join WAIS, he must submit his location, description, and other relevant information to the directory server. The WAIS Directory of Services presently contains about 300 registered databases.

3. THE ARCHIE SERVICE

The Archie[2] system is aimed at solving the problem of locating files by attributes. It currently offers Filenames and WHATIS databases. The Filenames database indexes the names of files available from hundreds of Internet file transfer protocol (ftp) sites. Ftp allows users to retrieve files stored on Internet hosts. Anonymous ftp has gained popularity as a medium for distributing software and documents. The WHATIS database contains the names and descriptions of software packages, documents and other information available on the Internet.

Users can perform case-insensitive keyword searches to locate the desired information.

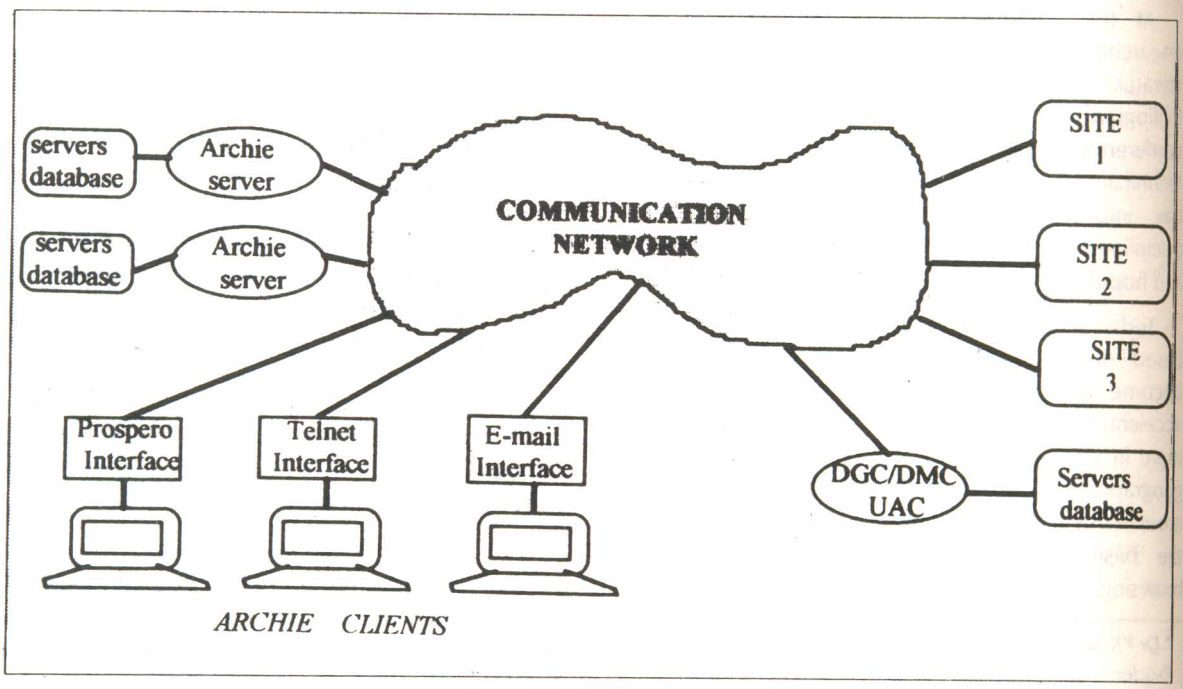

Figure 2: The Archie Architecture

Archie clients can access both databases through the user access component (UAC), while the data gathering component (GAD) relies on the ftp site administrators to discover new ftp archives. Presently the archie consists of about 12 servers around the Internet. Archie's simplicity and the use of existing resources have been the key features responsible for its popularity. After just two years in service archie boasts of 1000 registered ftp archive sites that offer 2,100,000 files and 3,500 software packages. Archie is accessed by users from 47 countries about 50,000 times a day.

4. GOPHER

Gopher[3] was developed at the University of Minnesota as a basic tool for browsing distributed information on the Internet. In this system the information is organised in a hierarchy such that the intermediate nodes are directories, or indexes, and the leaf nodes are documents. Thus the gopher information space is modelled as a directed graph, where cycles in the search process are allowable. The gopher architecture is shown in Fig. 3. It consists of clients and servers communicating through the gopher protocol, which is implemented on the top of TCP-IP (transmission-control protocol - Internet protocol).

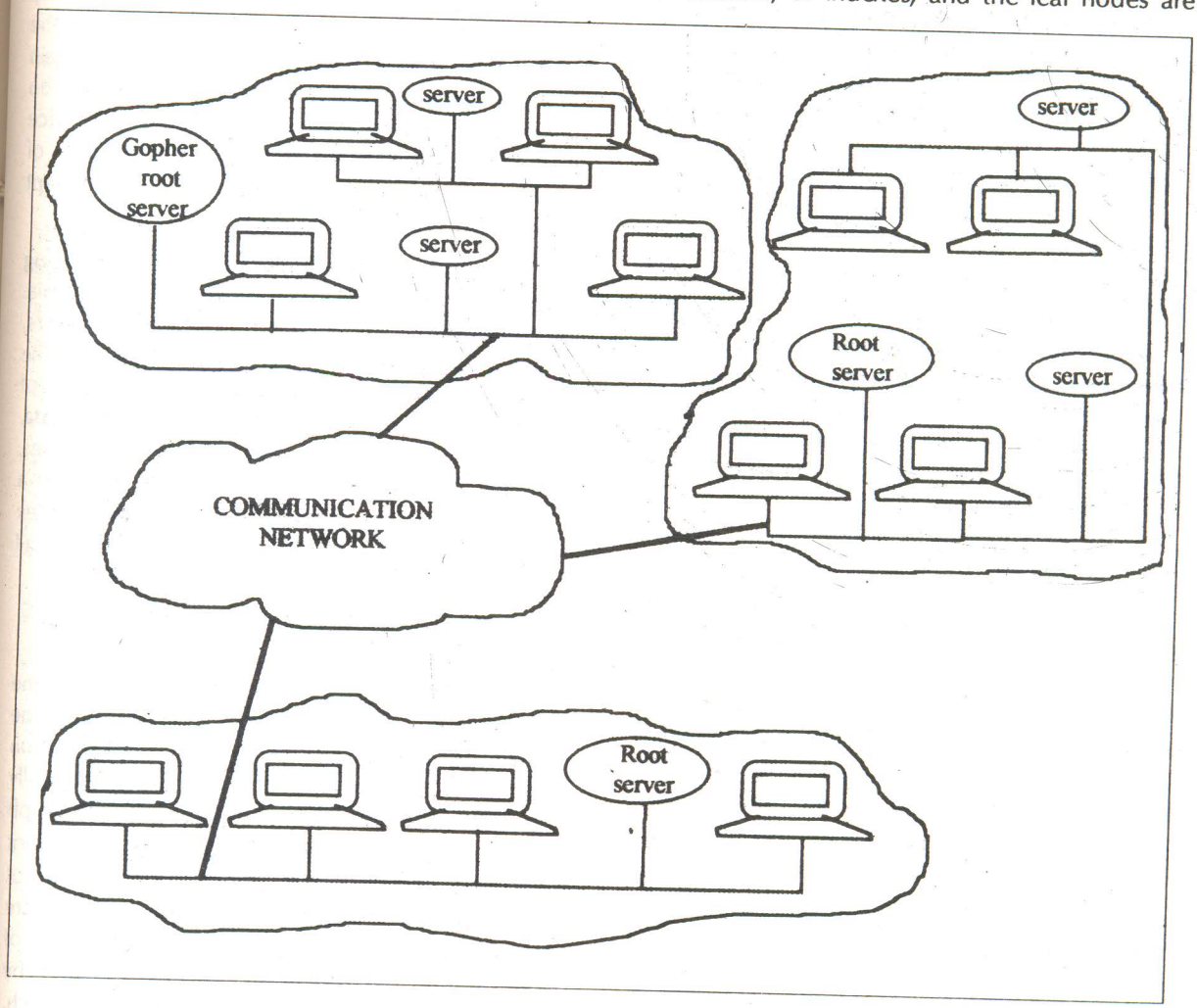

Figure 3: Gopher Architecture.

gopher protocol, which is implemented on the top of TCP-IP (transmission-control protocol - Internet protocol).

Currently, the root of the gopher's hierarchy resides on the host (rawBits.micro.umn.edu) at the University of Minnesota. This is the default directory received by the gopher clients when they are first invoked, although clients can also be configured with other entry points into the hierarchy. The gopher root server has the knowledge of all top level services and advertises their existence to users. For example, a R&D organisation, or a university running gopher server may register a central top-level server with the root server. Then this central gopher server can link to other departmental servers, which in turn, can link to lower level servers.

A user selects an object on the basis of its user-visible name, and the client retrieves by constructing a handle from the server's host name, its port number, and the object's path name. gopher clients can also retrieve objects from WAIS, archie, and ftp servers.

5. WORLD WIDE WEB

The World Wide Web[4] (WWW) was developed at CERN in Switzerland, and combines hypertext with information discovery technology. In this system the information is organised into a large distributed hypertext document. Here the nodes are either full text objects, directory objects (called cover pages), or indexes.

The architecture is based on a client-server model with the WWW clients offering the users hypertext like interface. The native protocol for WWW is hypertext transfer protocol (HTTP). In addition to this, WWW clients can understand ftp, and the network news transfer protocol (NNTP). Ftp can be used to access file archives on the Internet, where file directories are browsed as hypertext objects. Nntp is used to access news groups and news articles. These news articles usually contain references to other articles or news groups, which are represented as hypertext links.

Currently, besides WAIS and gopher servers, there are about 24 WWW servers accessible to WWW clients. The WWW servers on info.cern.ch has logged access from approximately 6,000 hosts that use their own WWW clients, or connect to a publicly available client.

Besides the information discovery services described above, two other services—X.500 and Indie are popular. The X.500 resulted from the standardisation efforts in the field of directory services by CCITT and the ISO. Here, the difference is that instead of the usual domain naming system, (which maps the host name on to corresponding Internet addresses, and vice versa) X.500 entries consist of a set of attribute-value pairs. Hence, X.500 can accept attribute based queries.

Indie is an acronym for Distributed Indexing developed at University of Southern California. It consists of a replicated directory of services, and a collection of broker databases. The brokers automatically cluster references of related information by indexing their own data, as well as data stored in other broker databases, and other information search tools. Thus, in a fashion similar to archie, this clustering of index information lets users efficiently search all participating databases.

6. INTELLIGENT SEARCH

As stated in the previous sections, the information space on the Internet can be modelled as a directed graph. Information discovery process can then be computationally viewed as graph search. Conventional graph search techniques employed by the current systems suffer from the symptoms of combinatorial explosion. We believe that there is a great scope for improving the efficiency of the information discovery on the Internet by using heuristic functions to guide the search. Ideally, we would like to know, at any given node which operator would get us closest to the goal node where we can find the desired information. If we knew this magic information throughout the search, for every node in the information space, we would find the optimal path in the optimal way. Here, we can use the standard AI technique of assigning a numeric estimate to each state that represents its

distance-to-goal. This will enable the search routine to rank the alternative paths in order of their closeness to goal. A heuristic evaluation function $h(x)$ achieves this by assigning to each state x an estimate of the minimum cost to reach a goal state. It must be realised, however, that the performance of heuristic search varies directly with the accuracy of the heuristic evaluation function.

With this model, the information discovery process can be guided by the A* algorithm[5] listed below:

A * SEARCH ALGORITHM

INPUT: Two nodes in the distributed information space, START and GOAL.

TO COMPUTE: Minimum cost sequence that transforms START into GOAL.

OUTPUT: Operator sequence, if one exists, otherwise NO.

1. **a.** Form queues of action records

   OPEN = (START, nil, nil, 0)

   b. CLOSED = empty

2. If OPEN is empty return NO, terminate

3. Otherwise, remove the first action record from OPEN, call it RECORD_X; put RECORD_X at the front of CLOSED.

   a. If state variable of RECORD-X is equal to GOAL, return the plan used, terminate

   b. Otherwise, expand STATE_X, giving a list of action records called SUCCESSORS; for each record in SUCCESSORS, insert it into OPEN in increasing $g(x)+h(x)$ order

   c. Go to step 2.

In the above search algorithm, the heuristic function $h(x)$ uses problem dependent information to guess the whereabouts of the goal node. There are three broad categories to search heuristics:

a. Heuristics that reduce search effort and guarantee optimal solution;

b. Heuristics that greatly reduce search effort but do not guarantee finding the optimal solution;

c. Heuristics that minimise a combination averaged over the problem instances likely to be encountered.

The online information discovery process on distributed database can benefit from the intelligent search described above if the heuristic function $h(x)$ can be obtained by collecting statistics on Internet access patterns for a particular class of users.

7. CONTENT BASED IMAGE RETRIEVAL

With the proliferation of the Internet and intranets, an ever increasing number of information access services are now receiving image retrieval demands from their users. Images are being generated by a number of sources, such as, defence and civilian satellites, military reconnaissance and surveillance flights, fingerprinting biomedical imaging, etc.

An efficient image retrieval system is required to effectively disseminate information from such image repositories. The content based image retrieval (CBIR) technology models images as a set of attributes extracted from the database. The user can then specify his frequency based on these attributes. Current research efforts are directed at an integrated feature extraction/object recognition subsystem to overcome the limitations of the attribute based retrieval. This is another point at which artificial intelligence techniques become relevant. In addition to using generic query classes like colour, texture, shape, spatial constraints etc., AI techniques involving evolutionary programming can be used to extract objects using a vertically integrated vision model.

The evolutionary program is a biologically motivated AI search technique that works with a pool of candidate solutions, and uses selection, crossover, and mutation operations to find the optimum solution satisfying user query. The evolutionary programming approach requires the formulation of the information

discovery as an optimisation problem. In image retrieval, the principal parameters of interest are the position co-ordinates r_i (i.e. x and y), orientation Θ, magnification γ, and the object classification label m. A vector G_i of the following form represents a candidate solution to the image retrieval problem:

$$G_i = [r_i \ \Theta_i \ \gamma_i \ m_i]$$

The overall merit of a candidate solution is represented by the objective function

$$\mu (G_i \mid s)$$

and depends upon two quality criteria—the *coverage merit* and the *merit of fit*. The evolutionary program is a biologically motivated AI research technique that works with a pool of candidate solutions, and uses selection, crossover, and mutation operations to find the optimum solution satisfying a user query.

8. CONCLUSION

Information discovery services have become indispensable to scientific and engineering progress. Currently, many useful retrieval tools are available for locating information in databases distributed geographically throughout the globe. AI oriented techniques can fully utilise Internet connectivity to effectively locate and disseminate information. Content based image retrieval can benefit from evolutionary programming techniques to quickly sift through large databases of images, and respond to user queries in acceptable time. There is a lot of scope in application of similar AI techniques to design the information discovery systems. It is imperative that future systems will use these methods to offer efficient and effective retrieval services.

REFERENCES

1. Kahle, B. and Medlar, A. An information system for corporate users: WAIS, *Connections—The Interoperability Report*, 1991, **5**(11), 2-9.
2. Emtage A. *et al.* Archie: An electronic directory service for Internet, Proc. 1992 Usenix Conf., 93-110.
3. McCahill, M. The Internet Gopher Protocol, *Connexions—The Interoperability Report*, 1992, **6**(7), 10-14.
4. Berners-Lee T. *et al.*, World Wide Web—the information universe. *Electronic Networking: Research Applications, and Policy* 1992, **1**(2), 52-8.
5. Nilsson J Nils. Principles of Artificial Intelligence. Narosa, 1994.
6. Michalewicz, Z. Genetic Algorithms + Data Structures = Evolution Programs, Springer-Verlag, 1992.