ABSTRACT

This paper describes a parallel computing architecture that supports information processing over the Internet. DaVIME (Data Visualization, Indexing, and Mining Engine) consists of multiple software components that can service various types of information requests. The functions provided by these components include visualization, indexing, searching, and ranking of retrieved information. When a user issues an information service request, DaVIME calls upon the appropriate component to provide the service. Currently, DaVIME includes Document Explorer which analyzes text information in large document collections and provides a suite of visualization tools; DUSIE (Dynamic User-created Searchable Index Engine) which extends the hierarchical indexing schemes currently used in popular Web browsers to let users build content-based searchable indexes; ParaCrawler, a parallel Web search engine under development which uses parallel computing techniques to provide fast, broad-based search results; and Gis2web which allows users to access geographic information system data from the Web. The DaVIME architecture is extensible so that future functionality can be added to the system.

Keywords: Internet, Information Retrieval, Searching, Indexing, Data Mining, Visualization, Parallel Computing.

1. Introduction

The Internet provides great opportunities, yet poses great challenges. Opportunities range from accessing digital libraries from classrooms [9] to building a virtual supercomputer from home [5, 13] to searching for the largest prime number [11]. Many challenges exist, including security, scalability, resource discovery, information extraction, and coordination of parallel computing activities. Research efforts underway to address these problems can be divided into two major categories: computing over the Internet and information engineering over the Internet.

Research efforts in computing over the Internet use the network to access the computing power of a large number of computers to accomplish computations. Issues are studied such as sharing computing power in a heterogeneous environment, distributing tasks to individual computers, coordinating the results, and developing incentive schemes for sharing excess resources. The general model is that the clients needing extra computation or storage access available resources offered by hosts with the help of a set of brokers. The key feature here is that the clients do not access data, only the computing power or storage on the hosts. Examples in this category include Javalin [5] where clients and hosts register with brokers that help coordinate the work, ParaWeb [3] where clients use the Internet and intranets as part of their computing infrastructure in a seamless fashion, and ATLAS [2] where hosts actively take work away from overloaded clients.

The second type of project deals with information engineering over the Internet. Typically, this involves search, retrieval, indexing, categorizing, and visualizing various types of information accessed across the Internet. The Internet’s massive amount of unstructured and often natural language data requires approaches different than used in most database applications. The challenges are to guide users in formulating queries, accurately locate relevant information among hosts, properly index and rank the information, and usefully present the information to users. Harvest [4] is one of the pioneering and successful Internet information access and discovery systems. Harvest servers run on the client and index information in locally stored files. When a service is requested, the server reports the result through a predefined network interface. MetaCrawler [12] uses existing Internet search engines. It forwards user requests to search engines such as Yahoo, Lycos, and others. Results are analyzed, filtered, ranked, and sent back to users.

2. DaVIME

DaVIME (for Data Visualization, Indexing, and Mining Engine) is an architecture that can effectively provide fast, usable information service from the Internet. Our goal is to provide users with a seamless Internet-based
information service using search, categorization, indexing, and visualization techniques in a cooperative computing environment.

DaVIME takes a general information service approach. As with any information retrieval system, it is vitally important that the user formulates proper queries. In practice, a large portion of an application’s efforts can go into properly formulating the problem (asking the right question) rather than optimizing the algorithmic details of a particular data method [6]. We use AI and visualization techniques in query refinement. Parallel search schemes are used to retrieve information. Search results are sorted and ranked, and presented to users through visualization tools. Then users can organize search results to suit their needs.

2.1. System Architecture

Properly servicing users’ information requests is at the core of the research. Our design goal is to have an open, extensible architecture that provides a comprehensive information service for users and allows incremental development for researchers. As shown in Figure 1, the system is divided into four major components, the User Interface Coordinator (UIC), the Data Resource Coordinator (DRC), the Computing Resource Coordinator (CRC), and various Extensible SoftBots (ESBs). The remainder of this section provides an overview of the system architecture and system components.

[UIC] Users request information service through the User Interface Coordinator. This component accepts and refines user input through interaction with other system components and is responsible for display management.

[ESBs] Once a user request is accepted, various softbots act on the requests and provide the user with a visualized response. The ESBs are a collection of various information processing software that can be expanded incrementally. Currently it is designed to include the following components:


Figure 1. Architecture of DaVIME
User-configurable indexing techniques [8] are supplied which extend the hierarchical indexing schemes in widespread use.

ParaCrawler is a parallel search engine. The search scheme used is similar to MetaCrawler [12] in that several search engines are simultaneously accessed, but using different indexing and ranking components.

Gis2Web [10] provides facilities for DaVIME users to directly access GIS (Geographic Information System) data through the Web, rather than through standalone GIS software.

The Data Resource Coordinator handles data exchange among ESB components and provides ESB components with internal and external data. When an ESB component needs data from other components or from Internet, DRC checks the data format and retrieves data for the ESB component from appropriate source through the Internet.

The DaVIME architecture allows the ESBs access to external computing resources (CPU cycles and storage) from the Internet or intranets, handled by the Computing Resource Coordinator.

2.2. Document Explorer: Information Retrieval and Visualization Tools for the WWW

Document Explorer [7] is a system for visualization of WWW content structure. Figure 2 is a screen snapshot of the Document Explorer. Visualization, browsing, and query formulation mechanisms are based on documents' semantic content. These mechanisms complement text and link based search by supplying a visual search and query formulation environment using semantic associations among documents. The user can view and interact with visual representations of WWW document relations to traverse this derived document space. The relationships among individual keywords in the documents are also represented visually to support query formulation by direct manipulation of content words in the document set. Finally, the system provides direct manipulation facilities for forming queries to identify starting points in the network of documents or use conventional vector based retrieval mechanisms. A suite of visualization tools for interaction and display are used to supply orientation cues and navigation mechanisms to assist users in browsing the document and keyword networks, as well as in the formulation of queries by direct manipulation of system objects.

Document Explorer's principal visualizations are network displays based on documents' keyword lists. The lists can be provided by automatic content extraction tools, such as Harvest[13] or derived within the system. Keyword lists for each document are used to determine the associations among documents and among terms. The statistical text analyses rely on recovering conceptual information from natural language by considering the frequency and co-occurrence of words. This basic approach has been used in a wide range of contexts and its utility and limitations are well-known.

The visual environment for exploration and direct manipulation uses the same visual representation for query, associative thesaurus, and document content to facilitate query revision by direct manipulation of system objects. The representations underlying the visual displays are minimum cost networks derived from measures of term and document associations. The associations are derived from natural language text for queries, single documents, and associative term thesauri. The document collection is represented as a network of documents based on interdocument similarity.

The spatial representations of the Document Explorer's networks are designed to facilitate users' perception of network structure. Network nodes are positioned in three dimensions using a graph layout algorithm based on a physical spring metaphor in which nodes are treated as connectors and spring length and strength among connectors is derived from network link distances. Nodes are allowed to vary in three dimensions and iteratively positioned at the points which minimize energy in the system of springs. Varying spring length and strength allows layouts which are useful for user interaction and visually reveal clustering and connectivity among nodes.

One of the central challenges in the display of large information spaces is to overcome users' feelings of disorientation, the feeling of being "lost in hyperspace". In the Document Explorer feelings of disorientation are attenuated in part by the layout of network nodes to facilitate perception of global structure coupled with overview diagrams which track the users current viewing position. Additionally, facilities to interactively vary network density provide global orientation while examining local detail. Finally, navigation tools including bookmarks, anchors, and signposts supply mechanisms allowing users to control the course of navigation and facilitate way finding in the large document and term spaces.
2.3. Gis2web: Web Access of Geographic Information System Data

Typical GIS software maintains, manipulates, and displays geographical information in dynamic and graphical ways. The traditional mode of operation is to have the data accessible from within a specific GIS application on a single platform. With the recent wide spread acceptance and availability of World Wide Web technology, it is natural to look toward building linkages between the two distinct worlds of GIS and the WWW in order to make the wealth of information available in the GIS world publicly available through the WWW. Gis2web [10] is such a system intended to bridge the gap between existing GIS software and the WWW. Gis2web accesses distributed data sets in different GIS formats such as Arc/Info, converts the data to formats appropriate for WWW distribution, and provides access through a gis2web server. Gis2web allows distributed GIS to dynamically update their databases and have this information available to the WWW through automatic retrieval by gis2web. Users access GIS data interactively through gis2web over the Internet as if they were using a local GIS. Figure 3 shows the architecture of gis2web.

Gis2web periodically extracts geographic information from the GIS databases accessible to the system and stores the information on a local disk. The type and amount of information to extract is based on access statistics. The information most frequently referenced is extracted and stored locally. Gis2web periodically collects this information from the GIS server and sends it to the WWW server for processing. These will most likely be separate servers connected by the Internet. Gis2web parses data received and converts it to formats used by the WWW servers, which in turn provide information at users' requests.
2.4. ParaCrawler: A Machine Learning Approach towards Web Searching

ParaCrawler is a search engine under development that has combined search, indexing, and ranking functionalities. Using a search mechanism similar to that of Metacrawler[12], ParaCrawler takes a user request then re-sends it to a number of well-known search services on the Internet in parallel by starting one process for each target search engine. Each process contacts a particular search engine and sends the user request to it in the proper format. Once a process receives the results of a query, the process then uses machine learning approaches to locate the most relevant search results and presents them to the user.

From the machine learning perspective, one needs a sufficiently large, unbiased samples of labeled examples to identify a target concept. After receiving input from the user, the algorithm interactively refines its search and finds a new concept for the user. This is the process of relevance feedback used by a number of search engines and widely known in information retrieval.

2.5. DUSIE: Interactive Content-based Web Structuring

The Dynamic User created Searchable Index Engine (DUSIE) [8] is a user-created dynamic open indexing system, which can connect to other information resources and allows individual annotation, provides the sort of functionality suggested above for augmenting Web structure. The goal of the user-created index is to reduce cognitive overhead by allowing users to define an index using their own terms, based on their understanding. Each index term contains semantic information relevant to the user. The indexing system keeps track of nodes and links to conceptual materials, thus providing a navigational tool through conceptual space. With its additional ability to cross-reference, the index is often more relevant to the user than an index created by the author. See Figure 4 for an illustration of DUSIE.

In the simplest terms DUSIE provides users facilities to construct an index for Web documents. Users define a list of index terms, make annotations to the terms, and link related concepts. In the context of the educational setting in which DUSIE was developed, a primary function of the indexing process is to provide a set of terms for each document that reflects the user's conceptual domain and can be used efficiently and effectively to retrieve information from the document. More generally, the system's functionality supplies mechanisms to create 1) new nodes through the annotation facility, 2) links to points in existing Web documents for the new nodes, and 3) alternative links for existing Web nodes through the addition of terms linked to points in existing documents.

DUSIE provides content retrieval in a manner similar to two-level architecture for hypertext documents through a top level index information and a bottom level containing content nodes and links. In the DUSIE implementation the hyperindex is a set of indices linked together. When an index term describing the required information is found, the objects from the underlying hypertext doc-
ument base are retrieved for examination. Retrieving information in this manner is retrieval through navigation. After selecting an index term, the information associated with the hypertext base link is displayed in the browser. The annotation can also be displayed in the browser. Information may also be retrieved using the related concepts entries, facilitating the retrieval of semantically related materials. The user can move directly to the desired information as well as access related information without having to remember locations, links or other structural details. As a reflection of the users conceptual domain, the index terms might be considered metanodes which are linked to other metanodes via terms in the related concepts term set. In this way the index system expresses local relations among clusters of nodes.

![Figure 4. DUSIE screen.](image)

4. CONCLUSION

DaVIME presents a unified view of a information service to users. The potential application is an information superstore where customers can stop and pick the information they want. Customers make a one-stop-shop for various categories of information. When a user issues an information service request, DaVIME calls upon appropriate software components (agents) to provide the service requested. Depending on the type of the service requested, DaVIME may call one or more softbot into action to satisfy the user request. DaVIME is an open, extensible architecture that allows researchers and developers to add software components (agents) incrementally. Currently DaVIME includes the following components. Document Explorer analyzes text information in large document collections and supplies the user with a suite of visualization tools. DUSIE extends the hierarchical indexing schemes currently used in popular browsers to allow users to build content-based, searchable indexes. ParaCrawler is a parallel Web search engine that incorporates users' document rankings as part of a search algorithm designed to rapidly converge on user search concepts. Gis2web allows users to access GIS data from the Web. The DaVIME architecture is open and extensible. Future components can be easily added to it to enhance the capability of the system.

5. ACKNOWLEDGEMENTS

This project has been supported in part by a NASA grant NAG 9-842.
6. REFERENCES


