A Hypermedia Database to Manage WWW Documents

Schubert Foo & Ee Peng Lim
Division of Software Systems, School of Applied Science
Nanyang Technological University, Nanyang Avenue, Singapore

Abstract

The surge of interest in the World-Wide-Web (WWW) with its potential commercial payoff has resulted in an explosion of information as organizations join in the bandwagon to publish and do business on the Internet. A related development, Intranet, which basically uses the same Internet technology to building private corporate WWW-based networks, has emerged to provide cost-effective and efficient groupware and information management solutions for organisations. As a result, the HyperText Markup Language (HTML), used for constructing WWW documents, has become a contender for authoring future office documents. With this scenario, this study examines WWW practices and highlights the inadequacy and drawbacks of current publishing on the WWW. This paper will demonstrate the need as well as the advantages in having a hypermedia database system to maintain and manage office or WWW documents and describes the design and prototyping of the D4W3 hypermedia database system.

Keywords: Internet, Intranet, World Wide Web, HyperText Markup Language, hypermedia database, data modelling, query, information retrieval

1. Introduction

Internet, the world's largest network inter-connecting networks across continents, is set to be the IT phenomenon of the nineties. Originally conceived in 1973 by the U.S. Advanced Research Projects Agency (ARPA) as a research program to develop a suite of communication protocols to allow networked computers to transparently across multiple, linked packet networks, it has grown substantially of late to become the world's largest network. This rapid growth stems from the continual support from the US Federal Government, international adoption of its Transmission Control Protocol/Internet Protocol (TCP/IP)[1] communication protocols, introduction of the World Wide Web (WWW, W3, Web) and commercial facilities. The bulk of the system today is made up of private networking facilities in educational and research institutions, businesses and government organizations across the world. It is estimated that up to one billion users will have access to the Internet by the turn of the next century.
Internet exhibits a host of features and facilities whose contents are contributed from many different parties including individuals around the world. The more popular uses of Internet include electronic mailing, USENET Newsgroup (a world-wide distributed discussion system), World Wide Web (a hypermedia system for browsing and retrieving Internet resources), File Transfer Protocol (ftp - a file transfer facility for uploading and downloading files around the world), Gopher (a globally searchable collection of menu-based hierarchical information resources) and Internet Relay Chat (IRC - a global chat line).

The popularity of Internet is derived from a number of contributory factors:

· **Easy accessibility.** Individuals and organizations without computing resources can gain easy access to Internet by acquiring an off-the-shelf computer, a low-cost modem and an account with a commercial Internet service provider (ISP) at competitive rates;

· **Ease of use.** The development of better human-computer interfaces and improvement in the level of computer literacy among many developing nations around the world has resulted in more people finding the Internet technology accessible and easy to use.

· **Global connectivity.** By the virtue of Internet's global inter-connectivity, all that is required is the knowledge of another user's 'Internet address' before communication can take place between two parties from anywhere around the world. Participation in the World Wide Web and USENET Newsgroups guarantees a world-wide audience.

· **Speed of access to Internet facilities.** Access to facilities is almost instantaneous. An electronic mail sent across anywhere round the world reaches its destination within minutes. Downloading a freeware or shareware program from a ftp site and getting it up and running is achievable in a short time. Likewise, browsing and searching for information at WWW sites can also be very fast.

· **Commercialization opportunities.** The advent of the WWW has transformed Internet into an environment which can support commercial interest and to form support business-oriented networks. From an organization's standpoint, the Internet offers a novel and attractive feature (unlike normal mass media tools such as television or newspapers) to allow customers to 'interact' with multimedia information provided by the organization and in some cases, to carry out business transactions on the spot.

· **Scale and distance independence.** Given the technical know-how of the technology, it takes little extra effort and costs to reach out to thousands and even million of users as to reach out to an individual. Additionally, the cost to reach
someone at the other side of the world as it does for someone in the same country is typically the same.

- **Low barriers to entry and equal opportunities.** As a result of the ease of access and an open and extremely democratic environment, equal opportunities exists for all in this business community. Large organizations investing more financial resources into making more impressive WWW pages may not necessarily exhibit a distinct advantage over an individual or a small team who may effectively focus on better communications and support for its niche market customers.

Internet has indeed revolutionize business in that it redefines the methods used in traditional business practice and offers an another important channel for mass communication which is likely to grow even further in future. Depending on the type and nature of the business, Internet provides a platform to carry out a host of potential business applications:

- As a public relations tool to establish a global presence and heighten public interest;

- As an a marketing tool to advertise goods and services, and to open up international markets;

- As a marketplace to sell goods and services;

- As an information kiosk to provide up-to-date business information, to answer frequently-asked-questions (FAQs) and to release time-sensitive information;

- As an alternative support tool to answer customers' queries and solicit feedback;

- As a research or information gathering tool for market surveys, product launches, and even to obtain solutions to problems encountered from around the world;

- As a human resource tool for staffing and recruitment purposes;

- As a support tool to serve mobile employees or telecommuters;

- As a computer supported co-operative work (CSCW) tool to facilitate groupwork and communication in an organization or across organizations, locally and globally.

The WWW in its current form, has already made many of such applications realizable. WWW-based business computing is already seen to be a new competitive business weapon. With the move towards the era of global information technology, regionalization and globalization of businesses, and the ongoing enhancement to Internet's technology (system and data security, encryption techniques, and communication protocols), the
WWW is set to take off further in the commercial dimension. This will be especially so when the pay-offs of organisations investing in this technology are revealed in future.

Another recent development is the deployment of Internet technology (in particular Web technology) into building private corporate Web-based networks. Known as Intranets, these networks are proving to be an exceptionally cost-effective way to distribute reports, track assets, improve employee communications, provide access to diverse corporate databases, distribute and run applications and enhance collaborations at a fraction of a cost of groupware solutions. An Intranet has the same look and feel as the Web. It uses the architecture and standards of Internet and the Web and can be closed off from the public Internet if necessary, through software programs known as firewalls. Organisations, can therefore use it as a closed system or open part or all of it for public access.

2. World Wide Web

2.1 Basic Concepts

The World Wide Web is a wide area hypermedia information retrieval initiative aimed to give universal access to a large universe of documents. It is organized as a set of HyperText Transfer Protocol (HTTP)[17] servers designed specially for rapid distribution of hypermedia documents. Hypermedia is a way of representing and accessing information. It views the information space as a graph whose nodes store information and whose arcs (links) represent semantic relationships between nodes. The links are usually associated with words or image icons within the document that describe the meaning of the links. A set of nodes and the corresponding links makes up a hypergraph (or hyperstructure) which can be represented as a network. This is shown in Fig. 1.

![Hyperstructure of a Hyperdocument](image)

Fig. 1 A Hyperstructure of a Hyperdocument
Node and node contents are independent objects, the association being made through a general referencing mechanism which allows the referencing of an entire document. As hypermedia systems are often used to structure information contained in pre-existing documents, the same document can be referenced by different nodes. Thus, the hypermedia model allows the sharing of node contents. When a hyperlink is activated (usually by double clicking the mouse button), the system extracts its destination node and corresponding loads the document as the current document. Users navigate around the hyperspace by moving from link to link, or by specifying direct links.

The HyperText Markup Language (HTML)[18] is a hypermedia language used to construct WWW documents. It is designed to specify the logical organization and formatting of general text documents, with extensions to include inline images, audio, video clips, fill-in forms and hyperlinks to other HTML documents and other Internet resources (such as files, ftp, USENET, telnet). As a result, HTML is not only applicable to WWW documents alone, but can be applied to the environment of office documents within an organization. It is anticipated HTML will emerge as the main contender for office documentation and that future office documents will be authored using authoring tools developed for the WWW. Hyperlinks will present a powerful means for navigation and to cross reference information. Each WWW resource has a unique address known as the Uniform Resource Locator (URL)[19].

In order to publish and read information on the WWW, three components are required:

(1) Access to a HTTP (Web) server. This is normally part of the package offered when an individual or organization subscribes to an information service provider. For organizations with computing resources, they can set up their own HTTP server and get it connected to the Internet via a commercial network provider;

(2) An HTML authoring tool. This is used to compose the WWW documents (also known as Web pages). Since a HTML document is basically an ASCII text file which contains embedded HTML tags, any text editor can be used for this purpose. However, this is not the normal practice of creating HTML documents since many existing commercial and freeware (with limited but adequate facilities) HTML editors are available for this purpose. Most editors are user-friendly and easy to use. Some uses the WYSIWYG (What-You-See-Is-What-You-Get) paradigm for authoring. Most editors also comes with a browser so that the end result can be reviewed immediately after composition. Corel WebDesigner [2] and HoTMetaL Pro 3.0[12] are examples of MS-DOS based editors. Phoenix[16], Emacs hhm[8] and HoTMetaL 1.0[12] are examples of UNIX based editors (See also Laviolette[5] for a latest compilation of HTML editors and related tools)

(3) A Web browser. This is the client software used for navigating and reading the huge amount of Web pages stored on various Web servers on the Internet. Although both forms of text and graphical-based browsers are available, the latter are much more popular as they can display graphics and icons which is

2.2 Case Example: Modelling an University Environment for Publication

The academic environment of Nanyang Technological University (NTU) is used as an example to illustrate how an information profile (i.e. information which is to be presented on the WWW) is modelled to a form suitable for publication on the WWW. This is shown in Fig. 2.

![Fig. 2 Information Profile of Nanyang Technological University on the WWW (Partial)]

By using a hierarchical tree structure, the various components of the organization for which information is to be presented can be defined and cross-linked to each other. The links represent the hyperlinks to connect a piece of information to another. The breaking up of the information profile in a systematic and logical manner will result in a tree structure which is used as the framework for publication. Such a framework can also be applied to any business organization.

At the top of the tree is the organization's home page. In WWW terminology, a home page refers to the document which is intended to be viewed first. It contains introductory information and/or a master menu of documents within the publication. It is generally associated with a particular Web site, person, named collection, or a business
organization. This master menu contains the hyperlinks to link to subsequent home pages or documents.

Using the NTU example, the first level menu contains the main focus of the information and is usually separated into a number of different general categories:

*What's New!* This is a public relations area for the announcement of headlines and announcements (such as new product launches and accolades).

*General Information.* This serves to provide general information about the organization: its mission, objectives, organization structure, financial information, human resource (HR) and other pertinent information. For a business organization, it may contain the company's financial highlights, balance sheets, profit and loss accounts, funds flow statement, excerpts from the company's annual report, human resource philosophy, various opportunities, vacancies, duties, requirements and application procedures.

*Schools and Degrees.* These are the general categories which provides information on the main thrust of the organization. These are the products and services offered by the organization. For large organizations this may be replaced by the various strategic business units (SBUs) with further breakdowns to define its products and services. In the example, each school (which is equivalent to a product division) is organized into a number of divisions with a supporting administrative office; each division comprises lecturers and technicians with their personal resumes. Resumes contain information on the courses taught, research interest and industrial consultancy expertise. Each school offers different undergraduates and postgraduate degree programs (products) together with the associated information of admission requirements, fees, financial assistance, curriculum structure and detailed course descriptions. From the same figure, it should be apparent that hyperlinks are used to link related information together (e.g., resume is linked to courses). Links can have a one-to-one, one-to-many and many-to-many relationships. Links also play the important role of eliminating repetitive information.

*Academic Services.* This plays the supporting role of the organization. In the example, these services include the Library, Computing Center, Center for Continuing Education, Center for Educational Development and so on. In a business organization, this may take on the role of customer support and enquiries. It may be used by customers to request for product servicing, checking for availability of spare parts, and an area to pose other forms of queries.

With the detailed information available for each component, it becomes an easy task to compose the Web documents and to cross-link them together using HTML conventions (i.e., anchors and URLs). As the information is contributed from various sources within the organization, there is generally an overall administrator (or Chief Information Officer) who is in-charge of the overall project with various responsibilities centers set up for actually defining and maintaining the information. Guidelines and rules of thumb are distributed to the various people involved to ensure homogeneity and conformity in presenting information. A schedule of regular updates and maintenance and a system to
handle proposed changes to existing information is also desirable to ensure overall data integrity and accuracy of information of the system.

2.3 Current Problems with WWW Practices

Organizations throughout various industries have jumped onto the band-wagon and started to use Web technology to publish on the WWW. Companies without computing resources will have to rely on commercial vendors to put up and maintain the information on their behalf. However, publishing in the WWW and using it in the present manner can lead to a number of difficulties in the future:

· **Maintainability of publication.** As organizations realize the ease to put information on the WWW, there will be a tendency to put more than less information which can result in the number of Web pages growing rapidly. Ultimately, it will reach a point where maintaining and ensuring the accuracy of information becomes a difficult task.

· **Integrity of Data.** The chances of information duplication or inaccuracy can arise as different people contribute towards publishing. The integrity of data can be compromised as different authors are responsible for their information domain. Such information could be cross-linked to information provided by other authors without their knowledge. Hence, deleting documents will cause integrity problems with related documents. In addition, information updates can be a problem if such information is shared among different information domains. For example, a change in the pricing of goods or services may require corresponding changes to a number of Web pages which are maintained by different people. Any loss of data integrity of the system can have potentially severe consequences, especially if the system is used for direct business transactions.

In order to ensure that these two problems are not present, there is a need for a structure, a change mechanism and manpower resource for the overall management of the system. Such a scenario may not be feasible in the case of organizations with limited financial and manpower resources. Additional costs is also incurred for having such a system in place.

· **Inadequate Search Facilities on the WWW.** There is currently no high level query language for locating, filtering and presenting WWW information. Searching is achieved by using any of the many existing Internet search engines.

Search engines work in many different ways: some search titles of headers of documents, other search the documents themselves, and still other searches other indexes or directories. Some are specially dedicated to the WWW (e.g. Lycos[6], Infoseek Guide[4], and so on), while others can also support searching of other Internet facilities such as USENET, FTP and Gopher (e.g. Magellan[13], Alta Vista[3] and so on). However, searching in this manner usually results in a long
list of matches (with possible duplicate matches if indexing is used) which must be further explored one at a time.

Furthermore, the home page of the document is usually presented so that navigating around the hypermedia space via hyperlinks is still required until the desired information is located. Such form of searching is suitable for Net surfing to look up general information from various sources. However it is inefficient, time consuming and unsuitable for cases when a specific and detailed query is known beforehand.

· Degradation of Performance as Number of Users Increases on WWW. As the number of users grows on the WWW, it will eventually overload the underlying network capabilities and result in a degradation of performance. Such observations are already apparent when accessing the more popular sites where network traffic is extremely high. Depending on the severity of the problem and expectations of users, this could lead to users’ frustration and subsequent loss of interest in using the WWW. For organizations which have invested heavily on doing business on the Internet, it becomes a real threat which they can do nothing or little to solve the problem.

Bearing such factors in mind, this research attempts to provide a solution by proposing the use of an experimental hypermedia system database system for the management and publication of WWW documents for organizations intending to use the Internet for business.

3. D4W3 - A Hypermedia Database System for WWW Documents

D4W3 is a hypermedia database system being developed at NTU. It is designed to address the management of HTML-based documents (including WWW documents) within organisations. In the following, we describe some system requirements that must be satisfied by the D4W3 design.

· Hypertext Editing. As WWW becomes popular and the HTML language becomes a well-recognised hypertext description standard, a wide range of tools has been made available as public domain software. These tools include WWW browsers, editors, parsers and translators. Office documents authored in HTML will allow the inclusion of text, multimedia objects, as well as hyperlinks which provide the means for document navigation. Some of these HTML documents can be made available on the Internet for public viewing while the others may be kept for internal use only.

· Multiple User Support in a Networked Environment. Being a office document management system, D4W3 must be able to support concurrent requests initiated by users working on different clients. All office documents have to be collected and stored at a centralized database which can be accessed remotely. Fig. 3 depicts an office environment in which a number clients'
machines and a central repository of office documents are linked together within a local area network. A gateway can be provided to allow remote access from the outside world. To prevent one user from tampering or accidentally modifying other users' documents, some form of security must be enforced.

![Diagram of Document Management in a Networked Environment](image.png)

**Fig. 3 Document Management in a Networked Environment**

- **Flexible Search and Retrieval Capabilities.** To support content-based queries on the office documents, D4W3 must allow users to specify queries by keywords as well as by some important attributes associated with the documents such as author, subject, title, and so on. In the D4W3 database, every document is assigned a unique document id similar to the URL of the document. The D4W3 query engine can therefore support retrieval by document identifiers (ids).

- **Easy Document Import and Authoring Functions.** Importing a bulk of documents and authoring them are two important considerations in D4W3. The former simplifies the migration of existing office documents into the D4W3 database. The latter facilitates the creation of new documents and updates on documents. Existing documents should be appended to the D4W3 database in batch mode. To benefit from a wide variety of WWW authorware, D4W3 must adopt a flexible design to interface with different WWW authoring tools.

### 3.1 The D4W3 Architecture

A general overview of the architecture of D4W3 is shown in Fig. 4. An HTML-based document application can be constructed by using the following modules:
3.1.1 Client Modules

D4W3 is designed to support hypermedia applications among which three important applications: a **Querying and Browsing Tool**, an **Import Tool** and an **Authoring Tool** must be included. Other hypermedia applications can be added in the future if necessary. The **Querying and Browsing Tool** allows users to formulate content-based searches and to view the hypermedia documents. Using the **Import Tool**, collections of hypermedia documents can be inserted into the D4W3 database in a batch mode. The **Authoring Tool** facilitates the creation of new documents and modification of existing documents.

The **Hypermedia Client Agent** provides a common client interface to the D4W3 database server modules which reside on a server host. Based on the client agent's library of application programming interface (API) functions, different types of hypermedia applications can be developed. A call to any API function is evaluated by sending a request to the D4W3 server and receiving the result from it. The **Communication Processor** on the client machine is responsible for sending and receiving messages to the D4W3 server. Before messages are sent to the server, they are encoded according to the request type and the related parameters. Message received are decoded in the reverse manner.

3.1.2 Server Modules

Like the clients, the D4W3 server requires a **Communication Processor** to interpret incoming messages and to construct outgoing messages. To handle concurrent multiple client-server sessions, a **Server Daemon** is required to listen to incoming server connection requests. For each client-server connection, the daemon creates a **Hypermedia Server** process which handles all queries from the client within the session established by the connection. To support flexible queries, some important attributes of hypermedia documents are extracted and stored in a **Database Server**. For practical
reasons, the database server used by D4W3 is a relational database system. The hypermedia documents are stored as files fully indexed using a **Keyword Indexer**.

### 3.2 Hypermedia Database Modeling and Database Schema

In D4W3, hypermedia documents are represented as a **directed graph (digraph)**. A node of the digraph represents a document which can be either a HTML document or image document. At present, we have restricted D4W3 to handle only text and image data. A **directed link** from one node to another denotes a hyperlink which allows users to navigate from one location in a HTML document to an image document or another location in the same or different HTML document. Given a directed link from node A to node B, we call node A the **source node** and node B the **destination node**. To conform with the HTML language, we do not allowed a directed link to be defined from an image document. On the other hand, zero or more directed link can be defined between two HTML documents or within the same HTML documents.

The overall D4W3 database model is depicted in Fig. 5. The figure shows an **entity-relationship diagram** consisting of entity types represented by boxes, their attributes represented by ovals, and relationship types between entities represented by diamonds. As shown in the database model, every hypermedia **Document** consists of a document id, an URL, a title and a filename. Documents can further classified into either HTML documents or image documents. While the attributes of HTML documents include author, subject, header and date of creation/update, there are only two attributes for image documents: image format (such as gif, tiff, etc.), and description (text describing the content of the image document).

![Fig. 5 Entity-Relationship Modeling of D4W3 Database](image-url)
The **has_link** relationship captures the directed links between HTML documents. It has the attributes: anchor_name, link parameter, text description, and location(offset) in the source document. The **has_image** relationship captures the directed links between HTML documents and image documents. Only the offset of **has_image** is kept by the D4W3 database.

To further organize the hypermedia documents, D4W3 allows users to group a number of related documents together to form a **Publication**. For example, **General Information** (see Fig. 2) is a publication consisting of many different documents. Every publication is owned by some user who is responsible forauthoring and updating documents belonging to the publication. Therefore, we have an **Owner** entity type maintaining an account name and password for every owner.

Since most database systems are not designed to represent and manipulate large text objects, we have chosen to store the original hypermedia documents as ordinary files. However, the filenames of documents are maintained within the D4W3 database.

### Database Relations

From the entity-relationship diagram, we derive the following relations to be stored in the D4W3 database. In our D4W3 implementation, we have chosen an extended relational database management system known as **POSTGRES**[14,15]. POSTGRES allows generalization relationship to be defined between two relations so that a relation can inherit attributes from its superclass relation, and tuples in a relation are also tuples of its superclass relations. Every tuple in POSTGRES is assigned a unique tuple id. We have therefore used the tuple id in place of owner_id and pub_id.

**OWNER**(acct_name,password)
**PUBLICATION**(pub_name, owner_id)
**DOCUMENT**(url, title, pub_id, filename)
**HTML_DOC** is a subclass of **DOCUMENT**
**HTML_DOC**(author, subject,header,date)
**IMAGE_DOC** is a subclass of **DOCUMENT**
**IMAGE_DOC**(format, description)
**HAS_LINK**(src_doc_id,dest_doc_id,anchor_name,parameter, description,offset)
**HAS_IMAGE**(src_doc_id,dest_doc_id,offset)

### 3.3 D4W3 Query Facility

In D4W3, users can query hypermedia documents by specifying search criteria on selected attributes and keywords. The attributes that can be queried include author, title, subject, and header. A complex search criterion can be constructed by a conjunction of multiple simple search criteria. To handle potentially large query result sets for any kind of search queries, D4W3 is designed to keep the result sets temporally at the server so that subsets of result sets can be returned to the client. Fig. 6 illustrates the steps a D4W3 client has to perform in order to search and retrieve documents from the server.
As shown in Fig. 6, a query session has to be established before any query can be submitted to the server. A user first specifies search criteria on the documents’ attributes and keywords. Since every document is assigned a unique URL, the query result of the search request will be represented as a set of document URLs and attribute values. The user can choose to perform one or more retrieval requests in order to obtain subsets of result from the server. Based on the summarized information about the documents which satisfy the search request, the user may wish to browse some of these documents and continue to navigate to other documents. This can be done by performing a breadth-first search on a user-selected document, and retrieving its closed related documents. These are documents that can be reached by direct or indirect links. Finally, the user-selected document and its closed related documents are transferred to the client for browsing.
For example, let A in Fig. 7 be one of those documents which satisfy the user’s search criteria. A has links to B and C which in turns have links to other documents. If the user wish to browse A and 5 of its closely related documents, a breadth-first search on the digraph will return the set of documents: \{A,B,C,D,E,F\}.

3.4 Authoring and Import Facility

In order to benefit and support a wide range of existing authorware which exist currently, a flexible authoring tool utilising the "plug-and-play" paradigm is chosen. In so doing, users can use their own preferred authorware to create and modify HTML documents. Flexibility is achieved as users can switch and upgrade authorware without affecting the functionality of the overall authoring tool.

4. Implementation Issues

In this section, we describe the implementation of D4W3 system. The D4W3 server, authoring tool and import tool have been implemented on the SUN SparcStation running SunOS 4. We have developed a D4W3 query front-end on the IBM PC running Windows 3.1. All software has been developed using C or C++. The client-server communication is implemented on TCP/IP.

To implement the query, authoring and import facilities on the D4W3 clients, we have structured the client-server integration by designing a comprehensive set of application programming interface (API) functions. The same set of functions can be later used to realize new D4W3 application or tools.

4.1 User Interfaces

The two main D4W3 user interfaces that have been developed are the document query front-end and the authoring tool.

D4W3 Query Front-end

As most D4W3 users are expected to be PC users, a graphical query front-end has been developed on the PC client to allow users to perform query and browsing tasks. The essential features of the front-end include:

- **Integrated query and browsing:** Other than querying the documents at the D4W3 server, our query front-end allows documents to be transferred from the server to the PC client and invoked using any user selected HTML browser to read the documents.

- **Search history maintenance:** The search criteria together with their results are kept within a search history maintained by our query front-end. This facilitates users to recall the searches they have previously performed. By modifying previous search criteria, they can easily derive new search criteria. In addition,
only full result sets of searches are stored in the search history. Partial result sets can arise as a result of the user not bringing the complete result set from the server to the client. At such, these result sets are not stored since they cause integrity problems and can lead to confusion in future.

Following the query formulation steps of Fig. 6, a user first starts D4W3 and request for a connection a D4W3 server as shown in Fig. 8. Subsequently, a search criteria is formulated and submitted for query as shown in Fig. 9. A list of search results is subsequently returned to the user in Fig 10. Having identified the desired result set, the user selects the "Retrieve Document" option and specifies the number of levels of related documents to be retrieved from the server to the client. When all the documents are transferred to the client, the user can call up any Web browser to display the HTML document using the "Load Browser" option. A browser setup dialog box showing the default browser (Netscape's Navigator 2.0) is presented to the user (not shown in paper). This can be changed for use with any other Web browser. Fig. 11 shows the retrieved document been displayed on the default browser. Upon exiting from the browser, the application returns to the Query result display screen of Fig. 10.

Fig 8. D4W3 client opening screen with request for server connection
<table>
<thead>
<tr>
<th>Type</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>courses</td>
</tr>
<tr>
<td>Subject</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 9 Search screen for query formulation

![Search Result](image)

Fig. 10 Query result display

<table>
<thead>
<tr>
<th>Total number of records found:</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of records retrieved:</td>
<td>3</td>
</tr>
<tr>
<td>URL to be retrieved:</td>
<td></td>
</tr>
<tr>
<td>Search Result List</td>
<td></td>
</tr>
</tbody>
</table>

- Author = 'lilian'  
  Title = 'P129-95 EYP Report'  
  Subject = 'p129-95fyp/report'  
  URL = 'http://www.ntu.ac.sg/abstract.html'

- Author = 'lilian'  
  Title = 'Roadmap.html'  
  Subject = 'report|typ'  
  URL = 'http://www.ntu.ac.sg/lilian/report.html'

17
4.2 Keyword Indexing and Searching using GLIMPSE

While the D4W3 server keeps extracted document attributes in POSTGRES relations, the documents themselves are stored as UNIX files external to POSTGRES. To support keyword search on these documents, a keyword index has to be established. GLIMPSE (GLobal IMPlicit SEarch), a search tool developed by Manber and Wu at the University of Arizona, has been used to provide the indexing features required by D4W3[7]. The novelty of GLIMPSE is that it uses a very small index but allows very flexible full-text retrieval including Boolean queries, approximate matching and even searching for regular expressions. It’s index builder, glimpseindex, is used to index HTML documents for the D4W3 server.

In D4W3, all HTML documents are stored as files in a designated directory. To allow us to identify which documents in the POSTGRES satisfy a user’s keyword search criteria, we use the POSTGRES tuple ids as the filenames for the HTML files. With the keyword index constructed by glimpse, GLIMPSE allows us to perform keyword searches and obtain the filenames of HTML documents containing the keywords. This is done by the following UNIX command line.

```
glimpse -c -I -l -y keywords > result_file
```
4.3 System Features and Advantages

Using D4W3 as a means for WWW publication and management offers a distinct number of features and advantages over traditional WWW publication:

- Integrated system for the management of HTML documents from authoring to publication.

- Allows users to directly login into the system to carry out query and retrieval operations. It offers an enhanced query and retrieval facilities not normally found in WWW search engines. Users can prescribe the amount of search results to be brought back to the client as well as the amount of information for browsing. As the central depository of information is smaller, stand-alone and specific to the organization only, it results in superior efficiency of searches.

- System can be configured to act as a mini Web site to store an exact copy of the organization's publication on the WWW thereby allowing remote users to login and visit the local site and use the normal hyperlink navigation to search for information.

- Quality of service (e.g. speed of access) is enhanced as the organization provides the direct computing resources for which they have full control upon.

- Easy to set up the access control structure to control ownership of documents and well as access to the information domain within the information profile. Such a structure provides a framework to control the information resources of the organization.

- Data integrity is preserved in the editing and updating of documents. It eliminates the chances of deleting documents (linked to other documents) by mistake. The locking mechanism of the database system ensures document integrity during updates.

- Integrity of system is ensured as the database is responsible for the management of overall system. Each time information is updated on the system, it can be triggered to automatically update the same information in the actual Web and local sites.

- System is flexible, configurable and upgradable as the users select their own HTML authorware and browsers.

- Duplicate information is minimized as the query engine provides a convenient facility for checking prior to publication.
5. Conclusions

This research has demonstrated the need as well as the advantages in having a hypermedia database system to maintain and manage office or WWW documents in the HTML format. A prototype, D4W3, had been implemented to demonstrate the research idea and it allows a structured and controlled means to publish, maintain and query HTML documents which is currently lacking today. The design is generic and employs a plug-and-play paradigm so that users can choose their own HTML editors and browsers for authoring and browsing. The database search and retrieval engine is interfaced with a set of API functions. The same set of functions can be later used to realize new D4W3 application or tools, or it can be replaced for use with other database systems.

References


[16] University of Chicago, *Pheonix HTML Editor* [URL: http://http.bsd.uchicago.edu/]

