Digital Library Services in Mobile Computing *

Bharat Bhargava
Department of Computer Sciences
Purdue University
W. Lafayette, IN 47907, USA
bb@cs.purdue.edu

Melliyal Annamalai
Department of Computer Sciences
Purdue University
W. Lafayette, IN 47907, USA
man@cs.purdue.edu

Evaggelia Pitoura
Department of Computer Science
University of Ioannina
45110 Ioannina, Greece
epitoura@cc.uoi.gr

Abstract

Digital libraries bring about the integration, management, and communication of gigabytes of multimedia data in a distributed environment. Digital library systems currently envision users as being static when they access information. But it is expected in the near future that tens of millions of users will have access to a digital library through wireless access. Providing digital library services to users whose location is constantly changing, whose network connections are through a wireless medium, and whose computing power is low necessitates modifications to existing digital library systems. In this paper, we identify the issues that arise when users are mobile, classify queries that are specific to mobile users and introduce an architecture that supports flexible and transparent access to digital libraries for mobile users. The main features of the architecture include a layered data representation, support of adaptability, dual broadcast and on demand querying, caching, and mobile-specific user interfaces.

1 Introduction

Digital libraries provide online access to a vast number of distributed text and multimedia information sources in an integrated manner [2]. Digital library data include texts, figures, photographs, sound, video, films, slides, etc. The size of the data and information repositories available is enormous. The data sources are distributed and heterogeneous and digital library services should provide a uniform interface to make the information transparently accessible.

Digital library users include not just users on a fixed local area network (LAN) or a wide area network (WAN), but users with mobile computers and wireless links too. Provision to access digital library services through wireless networks is required by a wide range of applications from personal to research to customized business computing. For instance, archeologists working on remote locations may need access to library data related to their discoveries. Travelers passing a signboard on a highway saying 'next exit 25 miles' would like to know the restaurants within a five mile radius of that exit.

Mobile computing environments are characterized by frequent disconnections, limited computing power, memory, and screen size of mobile hosts and varying modes of connection: fully connected, partly connected, and doze mode for conserving energy. Figure 1 summarizes the characteristics of mobile computing. Many challenging questions arise when one attempts to combine the two evolving technologies of digital libraries and mobile computing. Digital libraries are associated with large, static repositories of information. The queries are complex and involve processing, navigating, searching, and presenting of distributed, heterogeneous repositories of multimedia data. Mobile hosts on the other hand, are associated with realtime computation and processing of small amounts of data. Providing digital library services to mobile users involves modifying the nature of the services to accommodate the constraints placed by the mobile users. The goal of this paper is to clearly identify the implications of the possibility of accessing digital libraries through wireless networks.

2 Impact of Mobility on Digital Library Services

The mobile computing architecture we visualize has two distinct types of hosts: mobile and fixed hosts [7]. The fixed hosts, called *base stations* or *static servers or hosts* are augmented with a wireless interface to communicate with mobile hosts. Each mobile host can directly communicate with one base station,

[*] this research is supported in part by NSF under grant number NCR-9405931, a NASA fellowship, and a Purdue Research Foundation fellowship

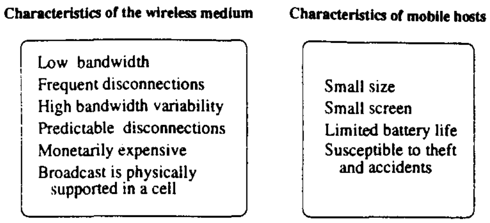
Characteristics of the wireless medium
Low bandwidth
Frequent disconnections
High bandwidth variability
Predictable disconnections
Monetarily expensive
Broadcast is physically supported in a cell
Characteristics of mobile hosts
Small size
Small screen
Limited battery life
Susceptible to theft and accidents

Figure 1: Mobile Computing Characteristics

the one covering the geographical area in which the mobile host moves. Wireless bandwidth used by a mobile host to communicate with the static host is a scarce resource and the data transmission over the air is currently monetarily expensive [6]. Mobile computers have to conserve energy and their compact design leads to small screen displays and limited storage capabilities. They are susceptible to disconnections due to out of range location, weather conditions, and loss of battery power. We have identified five mobile computing characteristics that have to be addressed by digital library service providers:

1. **Disconnected Operation:**
   In wired digital library access, the success of an operation depends heavily on the network. Mobile computers however are susceptible to frequent network disconnections. An increased number of disconnections can result from site or communication failures, and other disconnections such as those caused by battery limitations or handoffs are predictable. Predictable disconnections include voluntary disconnections. Frequently, users deliberately avoid use of the network to reduce cost, power consumption or because no networking capability is available at their current location. Thus, many users of digital library services will.only occasionally be connected to a network.

2. **Weak Connectivity:**
   Wireless networks deliver much lower bandwidth than wired networks and have higher error rates [4]. While wired network bandwidth of 155 Mpbs for ATM networks has been achieved, wireless communication have only 2 Mpbs for radio communication, 9-14 Kpbs for cellular telephony [4], and 250 Kpbs-2 Mpbs for wireless LANs. Since the bandwidth is divided among users in a geographical region, the deliverable bandwidth per user is even lower [10].

3. **Asymmetric Capabilities of Server and Client:**

   - *Communication:* Most of the static servers have powerful broadcast transmitters while mobile clients have little transmission capability.
   - *Computing power:* The static servers have typically more computing power than the mobile clients. Mobile clients have also less memory and smaller screens.

4. **Varying Client Location:**
   Some user queries might depend on the location of the user. Since the user is mobile, the location of the user might have changed by the time the query is processed.

5. **Variant Connectivity:**
   Mobile systems are characterized by high variation in network bandwidth, that can shift one to four orders of magnitude, depending on whether the host is plugged in or using wireless access and on the type of connection at its current cell.

2.1 System Architecture

The mobile computing characteristics summarized in the previous section influence the architecture of a digital library system. In this section we present an architecture to provide digital library services to mobile users which allows them to store, retrieve and transmit digital library objects. This layered architecture incorporates the special system requirements imposed by the mobile medium. It also serves the particular needs of a digital library system. The hierarchical representation of data accommodates possible heterogeneities and addresses scaling issues. Imprecise queries are central to digital library systems and the proposed adaptability schemes are amenable to multimedia data.

The six layers and their functions are briefly described below. Sections 3 and 4 contain a detailed description of the software at each of the layers.

- **On the Stationary Host:**
  - *Layer 1: Raw Data Storage:* This layer consists of the physical text, image, audio and video data objects.
  - *Layer 2: Metadata Storage:* A hierarchical representation of different versions of the same data objects is proposed as part of the metadata layer. Association of methods for realtime computation of a different version (smaller size) of the data object is also performed at this layer.
  - *Layer 3: Query Services:* Query services can be either *on demand* or *broadcast.* If information is broadcast, then it is the responsibility of the mobile host to recognize the information as relevant and retrieve it. On demand queries can be *realtime* or *delayed* with respect to response time. Realtime queries will retrieve information immediately while delayed queries will allow the mobile host to voluntarily disconnect, meanwhile process the query, and transmit the result when the mobile host reconnects.
  - *Layer 4: Caching:* Results of frequently asked queries will be cached for efficient query processing.

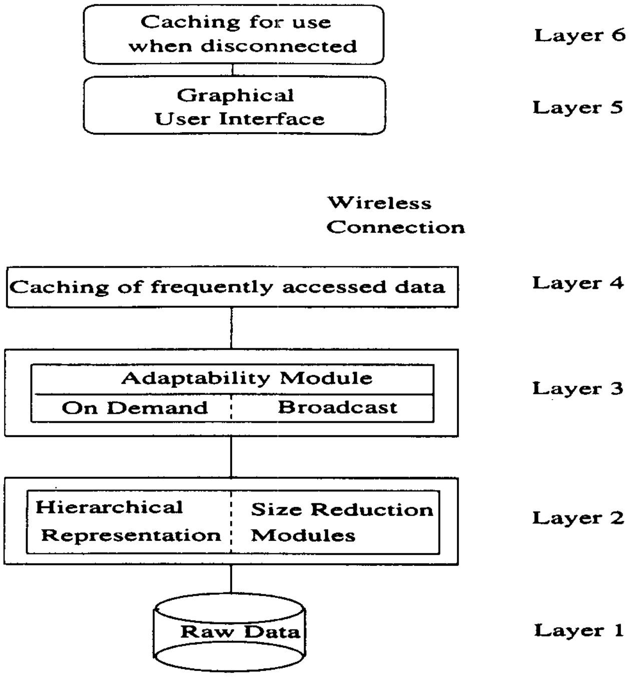

Figure 2: System Architecture

- **On the Mobile Host:**
  - *Layer 5: GUI:* A graphical user interface has to be designed to display a high volume of data effectively on a display screen of small dimensions.
  - *Layer 6: Caching:* Mobile hosts have to cache data. In the event of a disconnection, display of data can continue undisturbed if sufficient data has been prefetched in the cache.

3 Software at the Stationary Servers

3.1 Layer 1: Raw Data Storage

Digital library data consists of documents, alphanumeric data, video/audio data, and image data. The raw data, physical organization of the data, and the digital library databases constitute this layer [3]. This layer is the same as that of a digital library for static users.

3.2 Layer 2: Metadata Storage

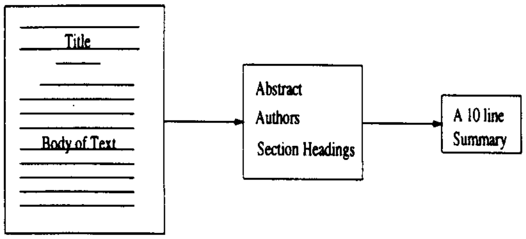

Figure 3: Illustration of Decreasing Order of Detail in Hierarchical Storage of Text

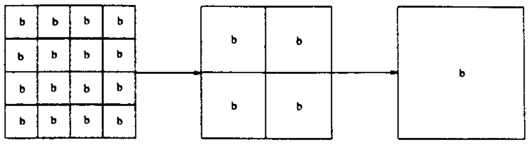

Figure 4: Illustration of Decreasing Order of Detail in Storage of Multiple Resolutions of an Image

Digital libraries use metadata as a form of representing structure, organization and content of data [5]. Metadata typically is of a much smaller size and hence less complex to process, retrieve and transmit in a mobile environment. Weak connectivity constrains the amount of data that can be retrieved by a mobile user. Depending on the available bandwidth and physical limitations of the mobile host, such as its memory or computing power, the size of the data to be retrieved can be computed. The choice of an appropriate representative of data is also guided by the semantics of the application on hand and by the type of the submitted query. A small data object can represent and be substituted for a larger data object. For example, an abstract can represent a body of text, a low resolution image can represent a high resolution image. The emphasis is to provide the user some information quickly instead of a lot of information after a significant amount of time. The representative data objects should be such that they convey as much information as possible. The goal is to increase the ratio of information to data. Two approaches can be adopted to provide a means of retrieving small data objects which represent larger data objects:

1. *Hierarchical Data Representation:* In this scheme a data object has several representations with varying degrees of detail stored with it. The degree of detail varies from being very detailed to having less detail. The system will choose the representation of data to retrieve and transmit so as to maximize information transfer with the available connectivity. As shown in Figure 3, a body of text can be stored as a document containing abstract and section headings and as another document containing two or three lines of the summary of content. An image can be stored in multiple resolutions (Figure 4).
2. *Realtime Computation of a Smaller Data Object:* In this scheme, the data size reduction methods are stored along with the data objects as in an object model. Depending on the response time required by the user, a smaller data object can be computed from a larger data object at the time of retrieval using the methods associated with the data object. Digital library data objects have different kinds of media and the different media have different methods. For

example, lossy compression techniques associated with an image data object will be applied to the image to reduce the size without sacrificing information about the image [1].

The tradeoff is that the second approach can produce a data object that will correspond exactly to the bandwidth available instead of a discrete series of data objects that are stored using a hierarchical representation.

3.3 Layer 3: Query Processing

The query processing layer is the most complex layer in the architecture. There are two modules in this layer on the static host, one to handle on demand queries and the other to handle broadcast of information.

3.3.1 On Demand Module

The on demand module on the static host accepts queries from the user and executes them on the static host. Depending on the nature of the query we broadly classify them as follows:

Imprecise Queries: The user does not have a precise idea of what she wants. She specifies the general nature of her request and it is the responsibility of the system to find out what exactly she wants by using a feedback mechanism. The user will know what she wants when she sees it. The hits are ranked based on relevance using standard information retrieval techniques. The user is provided with the most relevant n hits (high recall), where n is bounded by physical constraints such as the currently available bandwidth and the memory of the mobile host. The rest of the hits are cached in the stationary station and transmitted in batches till the user identifies a satisfactory answer.
Example: Retrieving an image based on the memory of seeing it in a book somewhere.

Imprecise queries are highly interactive. The system refines the query at each step based on the input given by the user. The asymmetric computing capabilities of the mobile client and the static server coupled with the weak connectivity make such an interactive process inefficient. To accommodate the disparate capabilities, the architecture should be designed to execute the query on the static server with as little interaction as possible with the mobile client. The mode of operation should be batch rather than interactive. In a batch mode, the mobile client can submit a query, voluntarily disconnect to conserve power, and then reconnect to obtain the result of the query. The architecture should maximize the amount of processing done on the static server and minimize the amount done on the mobile client.

To support the execution of an imprecise and incomplete query in a batch mode, we borrow the concept of a *user's profile* from information retrieval where information about recently asked queries is stored. This gives the system data about the user's interests. Input from the user's profile is incorporated to resolve ambiguities in the query and make it complete. Such an approach is feasible because of the specific domain mobile users' queries will be addressed to. For example, a user might want to find out about a new CD of music [10]. Information from previous queries on music of interest will enable the system to answer the query efficiently. The profile will also include information about the user's career, work, and since the users are mobile, reason for travel and destination. This information can also be utilized to predict the possible location of a mobile user.

Precise Queries: The user has some knowledge of the content and structure of the data. She requires a specific piece of information and knows how to formulate the query. There is no searching involved and thus the computing power requirement is low. Information is retrieved with high precision and low recall and thus the bandwidth required to transfer data is not large.

To better understand the functions and performance of a digital library in a mobile environment, we further classify precise queries for information into two groups depending on the user's needs. Digital libraries serve a wide variety of users and the classification below helps us clearly identify the problems in serving different classes of mobile users. The classification is based on the *response time* - the delay between the time the user requests a data object and the time the data appears on screen - allowed for the result.

- *Realtime Queries:* A mobile user is one whose location is changing. These users typically include travelers - both for pleasure and business, professionals whose nature of the job keeps them on the move like for example, pilots and policeman and users in remote locations without direct access to a static base station. If a result of a query is of no use beyond a certain short response time then that query can be classified as realtime. For example, the driver of a car on a highway might want to find out about restaurants within a five mile radius of the next exit. The information is useful only before she crosses the exit. After that the result is of no use and the query has to be reexecuted with respect to a new exit. Other examples of mobile users requiring realtime response are hospital personnel, police personnel etc.

  *Location*is a parameter which has to be included as part of the query. A query can be interpreted differently based on the value of the location parameter. Location has both physical and temporal coordinates. The following examples illustrate the importance of physical and temporal location coordinates:

  Physical location: Consider the query "retrieve all restaurants close by". Such a query can be interpreted differently based on the location of the user. If the location is on a highway, then the query can be translated to mean "retrieve all restaurants within a 20 - 30 mile radius". On a main street, the query can be translated to mean "retrieve all restaurants within a 9 mile radius". The results of the query are different in each case.

Temporal value: Consider a user wishing to travel to the airport submitting a query "what is the taxi availability at the location I am in ?". The departure time of the flight is a factor too. Beyond that time, availability of taxis is useless.

Our architecture will accept as input physical location coordinates and temporal value associated with the query and execute the query based on these parameters.

- *Delayed response Queries:* Some queries by mobile users are not time critical. The user can submit the query to one base station and choose to pick up the result at the next base station. Such queries may involve searching and hence realtime response is not possible. The software will execute the query and store it along with the user id and the user can access it when she wants to. Any query not retrieved by the user in a certain time period will be removed from the storage space of the server.

  Example: 1) Weather at a certain town for the next day. 2) Results for certain tests by an archeologist at a remote location. 3) Information needed by an executive for the day's meeting to be available before she reaches her office.

3.3.2 Broadcast Module

The static server has a higher transmission capability than the mobile hosts. Commonly accessed information such as traffic and weather information can be broadcast so that the mobile hosts can access broadcast information instead of querying for it. The broadcast module decides the information and frequency depending on the need and pattern of access. Library policy changes, rate changes are other types of information that can be broadcast. The module also keeps track of updates. Information can be multicast also by a static server to specific groups of users: newsletters, stock market information etc. The concept of broadcast/multicasting information is similar to radio transmission but in a mobile computing environment the user has the advantage of being able to respond and seek further specific information.

3.3.3 Adaptability Module

We have seen that wireless connections can have low bandwidth, varying bandwidth availability and periods of disconnection. To present a smooth transparent data display to the user, the query processing module should adapt to the current bandwidth availability. The goal in transmission is to maximize information content and minimize the size of the data. Multimedia data objects are especially amenable to preserving information content while reducing the size of the image. We discussed different resolution images above. We now discuss adaptability of image and video transmission to varying bandwidth availability.

Image Transmission: Images can be stored, manipulated and viewed in multiple resolutions. They are amenable to losing data without losing the semantics of the image [1]. Images are rich in semantic content and careful manipulation will result in an image so that no visible information is lost. Lossy compression techniques result in a smaller size and lower quality image. They exploit the fact that human eyes are more sensitive to luminance than chrominance. They reduce the bits used for chrominance and retain the bits used for luminance so that it is hard for the human eye to perceive the difference between a compressed and uncompressed image. Depending on the semantics of the image, we can achieve up to 100:1 [8] compression ratio. An image can have different levels of compression. The adaptability module has the responsibility of deciding what level of compression to use for the current bandwidth. The mobile user receives images at the same rate irrespective of the varying bandwidth.

Video Transmission: Like images, video clippings can be viewed, stored, manipulated in multiple resolutions, color schemes, sizes etc. Changing these parameters will result in a video file of a smaller size and hence lower the response time [9]:

- Color Depth: A greyscale video clipping requires lesser bits than the corresponding color video clipping
- Frame Size: Reducing the video frame size by half results in a 50% reduction in bits needed to code the video frame
- Frame Resolution: Resolution can be lowered according to the resolution available on the user's machine
- Codec Scheme: Different coding schemes have different compression ratios. Typically, schemes with high compression ratios require more time to compress but the smaller compressed frames can be transmitted more quickly. There is a tradeoff between compression time and communication time.

3.4 Layer 4: Caching (Static Host):

Frequently accessed data is cached by this layer for efficient access. Its functions are very similar to a caching layer in a digital library for static users. An additional builtin functionality is that data which is frequently accessed by *many* users is communicated to the broadcast module. The cache is also used to store answers to imprecise queries waiting to be transmitted to mobile hosts.

4 Software at the Mobile Host

The software at the mobile hosts includes counterparts of the modules at the static servers. The two principal layers are the caching layer and the graphical user interface layer.

4.1 Layer 5: Graphical User Interface

The mobile hosts have limited computing power, memory and screen size as a result of the requirement to keep their size and weight small [10]. The limited display size of about 10 square inches imposes a restriction on the amount of data that can be displayed to the user at a given time. The graphical user interface has to utilize the available space efficiently to integrate and present data from different sources as a coherent piece of information. Relationships between data objects which are unknown or not clearly visible to the user should be identified and emphasized.

4.2 Layer 6: Caching (Mobile Host):

Caching is used by digital library systems to reduce the response time for static users. The role of caching is even more important when data is accessed by mobile users. One important difference when compared to conventional computing environments is that in a mobile computing environment, disconnections can be voluntary. Our architecture will support caching to have a continuous data display during the voluntary disconnection periods. If the period of disconnection is known, then the amount of data that can be displayed in that time period is computed and is prefetched and stored in the cache.

We illustrate the function of the cache in the mobile host by using an example of retrieving digital library video data. The video-on-demand architecture [9] has been adapted to a mobile computing environment. The static server sends chunks of the video file to the user; the user displays the data that has been received while the rest of the video data is being transmitted simultaneously. If there is a voluntary disconnection sufficient video data is fetched before disconnection. This video data is stored in the cache and displayed during the disconnection period.

5 Summary

Providing digital library services to mobile users raises new challenges in the design of digital library systems. In this paper we have discussed the issues that arise and propose solutions that provide mobile users with a flexibility of use and access equal to that of static users. We have presented the design of a layered architecture which addresses the problems we identified. Our metadata layer stores data in multiple levels of detail to support adaptability with respect to available bandwidth in the query processing layer. We classify the different types of queries specific to mobile users. The query processing layer includes both a module for processing queries on demand and a broadcast module for frequently accessed information. This layer also contains an adaptability module which maximizes information transferred without increasing response time. The mobile host has a graphical user interface designed with the constraints of the limited screen size and a caching layer for smooth, transparent operation under disconnected mode. This architecture is part of a system under development in Raidlab at Purdue University in a research effort to address the impact of communications on digital library services [3].

References

[1] M. Annamalai and B. Bhargava. An evaluation of transmitting compressed images in a wide area network. Technical Report CSD 95-064, Department of Computer Sciences, Purdue University, October 1995.

[2] B. Bhargava and M. Annamalai. Communication costs in digital library databases. In *Lecture Notes in Computer Science Series (LNCS) 978, Database and Expert Systems Applications (DEXA '95)*, pages 1–13. Springer-Verlag, September 1995.

[3] B. Bhargava, M. Annamalai, S. Goel, S. Li, E. Pitoura, Y. Zhang, and A. Zhang. DL-Raid: Environment for supporting digital library services. In *Lecture Notes in Computer Science Series (LNCS) 916, Digital Libraries, Current Issues*, pages 281–300. Springer-Verlag, April 1995.

[4] G.H. Forman and J. Zahorjan. The challenges of mobile computing. *IEEE Computer*, 27(6), April 1994.

[5] J. Griffiths and K.K. Kertis. Access to large digital libraries of scientific information across networks. In *Proceedings of the the First Annual Conference on the Theory and Practice of Digital Libraries*, June 1994.

[6] D. Hayden. The new age of wireless. *Mobile Office*, May 1992.

[7] T. Imielinski and B. R. Badrinath. Mobile wireless computing: Challenges in data management. *Communications of ACM*, 37(10), October 1994.

[8] D. C. Kay and J. R. Levine. *Graphics File Formats, second edition.* Windcrest/McGraw-Hill, 1995.

[9] S. Li, S. Goel, and B. Bhargava. VC Collaborator: A mechanism for video conferencing support. In *SPIE Photonics East '95 Symposium - First International Symposium on Photonics Technologies and Systems for Voice, Video, and Data Communications, SPIE Proceedings Vol. 2617*, Philadelphia U.S.A, October 1995.

[10] E. Pitoura and B. Bhargava. Building information systems for mobile environments. In *Proceedings of the the Third International Conference on Information and Knowledge Management*, November 1994.