Archaeological Distribution Map System Using Aggregate Information from Heterogeneous Information Sources

Ayako Hayashi*  
Dept. of Information Science, Kyoto Institute of Technology, Goshokaido-cho, Matsugasaki, Sakyo-ku  
Kyoto 606-8585, Japan

Teruhisa Hochin  
Information and Human Sciences, Kyoto Institute of Technology, Goshokaido-cho, Matsugasaki, Sakyo-ku  
Kyoto 606-8585, Japan  
E-mail: hochin@kit.ac.jp

Hiroki Nomiya  
Information and Human Sciences, Kyoto Institute of Technology, Goshokaido-cho, Matsugasaki, Sakyo-ku  
Kyoto 606-8585, Japan  
E-mail: nomiya@kit.ac.jp

Abstract

This paper proposes archaeological distribution map system sharing archaeological information. Many archaeologists want to analyze the distribution of relics and/or remains. Excavation information, however, cannot be opened until verification and research are completed. Archaeological information sources are heterogeneous and distributed. This paper addresses to these problems through handling aggregate information, clarifying common schema and the structure of log data, and uploading data during a computer in on-line. Execution examples show the possibility of this kind of system.

Keywords: Distributed information sources, Heterogeneous information, Sharing system, Distribution map, Databases, Archeology.

1. Introduction

In recent years, digitization of various information proceeds. Until now, the information written in sheets has come to be digitized. A lot of information about the excavation has been written in paper medium until now. However, in recent years, in order to simplify the work of storage, and analysis of information, archaeologists configure databases on their own. They have been come to manage databases.

Unlike the information on papers, digitized information has many advantages. If the information is digitized, it makes the sharing of exactly the same information even in remote areas possible as well as the handling of information easier. However, information of excavation has not sufficiently been shared. The reasons of insufficient sharing of information are as follows: The first is that the information about the excavations of the ruins is exclusively used. Such information is usually kept closed. As it takes long time to verify and study such information, it may take several years for the
information to become publicly available. The second is that the excavation information is not created to be shared. Each archaeologist has information in different format each other. Thus, unification of the information structure is required. It is, however, a difficult task. Due to these reasons, sharing of excavation information is difficult. There are several studies for treating heterogeneous information.\textsuperscript{1,6} Terakawa \textit{et al.} and Wang \textit{et al.} have proposed the method of allowing such heterogeneous information sources to be uniformly used.\textsuperscript{1,3} The system enables users to use heterogeneous databases. This system has, however, not been put into practical use.

Archaeological data such as remains and relics are in association with the position information.\textsuperscript{7-10} In order to manage such data on a computer, the user must provide map data. Mapping system using the Google Map has been proposed.\textsuperscript{8,9} Users do not have to prepare map information by using this system. In carrying out archaeological considerations, a distribution map is often valuable. It brings easy considerations such as flows of material, and people in the past to archaeologists. For creating a wide distribution map, the information sources covering the wide area are required. As these information sources may be of several archaeologists, they must be gathered and/or shared. Sharing them is, however, difficult as described above.

This paper proposes a system enabling users to share aggregate information without disclosing the excavation information.\textsuperscript{11} The issues resolved are the treatment of closed information, the overcoming limited connection, and the treatment of heterogeneous information sources. By only handling the aggregate information, closed information can be treated. For conquering the limited connection, the data are uploaded during a computer is on-line. The common information structure is trying to be clarified for treating the heterogeneous information sources. Moreover, there are several problems to be conquered in implementing the system.\textsuperscript{12} We clarify the structures of information treated in the server. These are the structures of the common schema, user information, and log data. The system is based on the client-server architecture. Client system sends the aggregate information. Server system creates distribution maps and manages the information. An example of an execution of the implemented system is shown. The system shown in this paper can create distribution maps without uploading detailed information. It is considered that the wide area distribution maps can be created.

The remaining of the paper is as follows: Section 2 briefly surveys related works. Section 3 describes the issues and the approaches. Section 4 describes design and implementation of the proposed system. Section 5 presents the execution example of the system. Lastly, Section 6 gives concluding remarks.

2. Related Works

2.1. Archaeological Distribution Map

Recently, in the field of archeology in Japan, digitization of information is advanced. Study using excavation information is often in association with the position information. The excavation information results in a distribution map. A distribution map is manually created in most cases. This method takes a lot of time and cost, and produces many mistakes.

There are two major reasons that creating a distribution map is manually performed.

The first reason is that it is necessary to upload information in order to create a map. Excavation information held by archaeologists is of the study. They do not want to upload them in many cases.

Another reason is that there are some cases that a distribution map is desired to be created on a detailed map. This means that the general-purpose mapping systems can not correspond to such a case.

To solve these problems, some systems have been proposed in order that archaeologists can view the distribution map of the search results easily. For example, in order to facilitate the analysis of archaeologists, remains in an area of the database are displayed by WebGIS.\textsuperscript{10} This makes the creation of a distribution map easy. The application of this system is, however, limited to the fixed area. It cannot treat heterogeneous information sources.

2.2. Heterogeneous Information Management

The Stanford-IBM Manager of Multiple Information Sources (TSIMMIS) system has been proposed for the heterogeneous information integration.\textsuperscript{5} The TSIMMIS system provides a wrapper and a mediator. A wrapper absorbs heterogeneity of structures of information sources. A mediator combines the information to answer...
the query. The tools automatically generating mediators and wrappers are provided.

There are approaches to heterogeneous information integration using XML in recent years. 6 XML is a flexible format. It can be used as an internal representation of both semi-structured data and structured data. Thus, it is expected that heterogeneous information sources can be integrated under the common data model and XML. MediPresto/XM has been proposed. 6 This system treats XML as a common data model. The XML structure conversion has been implemented for the integration of heterogeneous information sources. It is believed that information sources having different structures become possible to be shared by using this method.

3. Issues and Approaches

3.1. Requirement

In order to analyze the excavation information in association with the position information, the distribution indicated is important in archeology. Distribution of a wide area is considered to be useful to think of such path distribution in the past. Therefore, distribution maps of a wide range are required for the archaeological research.

3.2. Issues

3.2.1. Closed information

In order to display the distribution information of excavation in the wide area, it is necessary to use the sources of information possessed by the archaeologists in each region.

Here, verification and research of excavation information take long years. So, it takes a very long time for the information to be publicly available.

3.2.2. Limited connection to information

The second issue is that the source of the information possessed by each archaeologist is not always accessible.

When the information is stored in an archaeologist’s PC, it is not always powered on. When the information is stored in the PC or a server of an institute, the access to it is usually quite limited for the security reason. Few persons can access it. Therefore, in either case, the access to the information is limited.

Fig. 1. An example of the different structures of information sources.

When the information of wide areas must be used, and all of the information is not included in a computer, the system must request the required information to the computers having it. However, if the computer having the information source is off-line, it is impossible to obtain the information required.

3.2.3. Heterogeneity of information

The third issue is that the structures of the information sources of archaeologists are different from each other. A concrete example of the different structures of sources is shown in Fig. 1. Both of the tables have the column “Remains number” as the first column. The names of the second column are different, while this column stores the same information. The third columns are different from each other. The third column of the left table stores the excavation day, while that of the right table stores the information of general classification.

One of the reasons is that archaeologists have stored information according to their needs. The data items and the precision level are different according to the periods treated, their positions, and so on. For example, a piece of a stone tool in the primitive age is precisely examined. Its size, its weight, its shape, and so on are measured and recorded. On the other hand, a piece of a cup in the middle age is not treated in such a way because it is no longer interesting. Only the category of the cup is recorded.

Another reason is the lack of a common form. The form common in all of regions and all of periods becomes a quite trivial one. This may be too common and trivial to be used in the archaeological study. This is due to the diversity of information according to the periods and/or the regions treated as described above. It is said that there is no common form in practice.
3.3. Approaches

The approaches to resolving these issues are as follows:

3.3.1. Treatment of closed information

In order to use closed information, we follow the method of obtaining private information without disclosing it. This method obtains aggregate information rather than information itself.

Based on this method, the information is converted into the form of aggregate values. It is believed that the information for displaying the distribution in a wide area can be available without uploading the detailed information.

3.3.2. Overcoming limited connection

Even when a computer is off-line, in order to provide a service of displaying distribution, the following method is adopted.

The aggregate information is uploaded to the server at the time that the computer storing the source of information is online. The information uploaded is stored in the server to prepare the next search. It is necessary to satisfy the following two assumptions:

- The computer storing the source of information is on-line at least once.
- The user who owns the source of information gives permission of using the information to the system.

3.3.3. Treatment of heterogeneous information sources

In order to share the information sources having different structures, mapping of the information sources is necessary.

3.4. Toward System Implementation

There are several problems to be conquered in implementing the system.

3.4.1. Clarification of the structure of information on the server

In order to integrate the information sources, it is necessary to determine the underlying information structures. If a minimal structure is not provided, the structure cannot be decided. The implementation of the system is impossible. It is believed that the underlying information structures are obtained from the attributes of relic, location information, and so on. Since it is necessary to cover the attributes frequently used by archeologists for displaying a broad distribution, it is necessary to investigate the attributes.

3.4.2. Dealing with duplicate data

A problem occurs when there are multiple sources of information in the same area. If they are exactly the same information, one of them must be selected. However, the contents of the information about the same thing may be different. For example, there are two copies, and only one of them is updated. In this case, a mechanism to select one from the two is required. Since it is unknown which information is correct to the system, a selection method must be introduced. A method is to select the newest one.

3.4.3. Decomposition of query

If there is a demand to display the distribution in a particular region, it is necessary to send a query to each resource stored in the computer of the community included in the area.

In the proposed system, the information sources included in the target range are identified. The proposed system builds a query to send it to the computer storing the information source. The client program receiving it aggregates information in the computer, and sends it to the server. The server creates distribution information based on the aggregate information gathered from the clients.

An example of the required query decomposition is shown in Fig. 2. The area within the dotted line is the target area of the retrieval. The information sources A, B, and C store the information related to the retrieval. The original query is decomposed to three queries, each of which is for each of the information sources. The results from the information sources are summed up to create the final result to the query.

3.4.4. Keeping integrity of information

The information source may be updated because it may be under the verification. The system, however, may not be able to obtain such information because the computer is off-line. The information cannot be obtained until the computer comes online again. Therefore, the consistency of the information may be violated. As accessing the information source in off-line is impossible, a fundamental solution of the problem is...
impossible. Our plan is informing the last updated date and time of the information source to the user who made the request.

4. Design and Implementation

4.1. Common Schema

A common schema that we have designed is shown in Fig. 3. We decided it by referring to the multiple archaeological literatures.13-15

The common schema includes seven tables as shown in Fig. 3. The table “Ruins” includes the information of an archaeological site. The table “Excavation” includes the information on excavation. The table “Remains” includes the information of an archaeological architectural structure. The table “District” manages the blocks of excavation. The table “Relic” has the information of an archaeological artifact. The table “Common information” has the information shared by the tables described above. The information of media data is included in the table “Media data.”

By mapping individual schema to common schema, users’ information can be stored in a common structure.

4.2. System Overview

The proposed system consists of server-side system and client-side system. System overview and the flow of the processing are shown in Fig. 4. Fig. 4 (a) shows the flow of processing for registering information sources
that a client possesses. Fig. 4 (b) shows the flow of processing when a client requests a distribution map. In Fig. 4(b), Client 1 gives a request of creating a distribution map to the server. The server gathers the information required for creating the distribution map. Please note the server requests the information from Client 1, which is the requester of the distribution map. After the information is gathered, a legend of a distribution map is generated. Next, the distribution map is created and displayed. Finally, the data used are stored in the server for future requests.

4.3. Client-Side System

Client-side system has the three functions: registration of user information, searching, and request of distribution maps.

4.3.1. Registration of user information

Server-side system selects the source of information to be referred to during the search. It uses the user information for the selection. Client-side system performs registration of user information. The user information is shown in Table 1. Maximum and minimum longitude and latitude represent the area which a user manages. The column “schema” holds the schema information of user’s tables. The column “minimum_grid” stores the minimum length of a grid, where grids form sections drawn on a map. The column “upload_condition” specifies the type of data uploaded: “detailed,” “aggregate,” or “impossible,” which will be described in 4.4.1. The column “hold_state” specifies the type of data stored in the server: “detailed,” “aggregate,” or “none.” The column “last_update” has the date of last update.

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>User name</td>
</tr>
<tr>
<td>range_x_min</td>
<td>Minimum east longitude</td>
</tr>
<tr>
<td>range_x_max</td>
<td>Maximum east longitude</td>
</tr>
<tr>
<td>range_y_min</td>
<td>Minimum north latitude</td>
</tr>
<tr>
<td>range_y_max</td>
<td>Maximum north latitude</td>
</tr>
<tr>
<td>schema</td>
<td>Schema</td>
</tr>
<tr>
<td>minimum_grid</td>
<td>Minimum grid</td>
</tr>
<tr>
<td>upload_condition</td>
<td>Upload conditions</td>
</tr>
<tr>
<td>hold_state</td>
<td>Holding state</td>
</tr>
<tr>
<td>last_update</td>
<td>Last updated date and time</td>
</tr>
</tbody>
</table>

4.3.2. Search

Client system performs a database search locally. Then, the search results are classified according to the position information and the user-specified minimum grid size. The client sends the total value of each section to the server. In addition, since the server sends the query based on the common schema, it is necessary to perform query transformation for the individual schema in the client system.

4.3.3. Request of distribution maps

A user specifies the retrieval condition on a graphical user interface (GUI). Based on the mapping of the common schema to the individual schema, input fields suitable to the user can be created and shown on GUI. These are believed to be useful to the user.

4.4. Server-Side System

The server-side system has the following four functions: information management, information gathering, legend generation, and creation of a distribution map.

4.4.1. Information management

The server-side system has to manage the user information and the information used for distribution maps. These are stored in a database. The system refers to the user information as a selection criterion in selecting an information source required for a search request.

The server holds the data uploaded from a client. Since the disclosure level is different from each client, we define the following levels for making the data open:

- Detailed data: Raw data
- Aggregate (log) data: Numbers of objects satisfying the search condition with the position information having a certain size in the source.

The data structure of log data is shown in Table 2. The name, and minimum and maximum longitude and latitude are the same as those of the user’s information shown in Table 1. The columns “site_type,” “site_period,” “remain_type,” “remain_period,” “relic_keyword,” “ex_year,” “layer,” and “ex_order” are for storing the retrieval condition. For example, the retrieval condition is “site_period = ‘Middle_Ages’,” the column “site_period” contains the character string “Middle_Ages.” As relics usually have qualifiers such as “large” and “white,” the retrieval condition on relics...
is represented with comma separated words. The character string “large, white, dish” is an example of this retrieval condition. This means a large white dish. The columns “ex_year” and “ex_order” are for an excavation year and an excavation order, respectively. The column “type” is of the type of retrieval target: “site,” “remain,” and “relic.” The column “agg” is for aggregate value. The column “last_update” has the date of last update.

Table 2. Structure of log Data

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>User name</td>
</tr>
<tr>
<td>range_x_min</td>
<td>Minimum longitude</td>
</tr>
<tr>
<td>range_x_max</td>
<td>Maximum longitude</td>
</tr>
<tr>
<td>range_y_min</td>
<td>Minimum latitude</td>
</tr>
<tr>
<td>range_y_max</td>
<td>Maximum latitude</td>
</tr>
<tr>
<td>site_type</td>
<td>Type of the site</td>
</tr>
<tr>
<td>site_period</td>
<td>Period of the site</td>
</tr>
<tr>
<td>remain_type</td>
<td>Type of the remain</td>
</tr>
<tr>
<td>remain_period</td>
<td>Period of the remain</td>
</tr>
<tr>
<td>relic_keyword</td>
<td>Keyword of the relic</td>
</tr>
<tr>
<td>ex_year</td>
<td>Excavation year</td>
</tr>
<tr>
<td>layer</td>
<td>Layer</td>
</tr>
<tr>
<td>ex_order</td>
<td>Excavation order</td>
</tr>
<tr>
<td>type</td>
<td>Type</td>
</tr>
<tr>
<td>agg</td>
<td>Aggregate value</td>
</tr>
<tr>
<td>last_update</td>
<td>Last reference date and time</td>
</tr>
</tbody>
</table>

4.4.2. Information gathering

When a client sends a request to the server, the server creates a distribution map based on the retrieval condition. When a request is received from a client, the server selects an information source with the information of the corresponding range, and requests the total value of the section of a constant size.

Fig. 5 shows the flow of the selection of information source. The system divides the area of a distribution map into sections, and selects one of the sources for each section. When the data holding status of a section is “detailed data,” the system sends a query to the server’s database, where complete data exist. When the data holding status is “log data,” the system checks whether the log data stored can be used. If the log data whose retrieval condition is identical to the current one exist, the system uses that log data. If not, the system tries to get information from the client having the information. When the client is in online, the system sends a query to the client for obtaining the target information. When it is not in online, the system notifies the user that the information cannot be obtained, and the distribution map is not perfect.

4.4.3. Legend generation

In a distribution map, the color of a section is changed according to the aggregate value of the section. The system automatically generates a legend of colors for the corresponding numerical value. The legend is calculated by using the number of sections and the aggregate values collected.

Two generation examples are shown in Fig. 6. These are examples of the divisions of three and four. In the division of three, a range of values are divided to zero, the values less than thirty, and those larger than or equal to thirty. The colors corresponding to the aggregate values are painted to the sections. For the division of four, the legend is created as in the similar way.
4.4.4. Creation of a distribution map

The distribution map is created by using the Google Static Maps API. By sending URL parameters, it is possible to obtain the map image. Following the legend created as described above, the color of a section is determined. By using the path parameter, we can draw figures and specify colors.

5. Execution Example

We show an execution example of the system implemented. We currently constructed only the Japanese version.

Fig. 7 is a selection window of the client system at the startup. The user moves to other windows. When the left (right, respectively) button is pressed, the registration (search) window shown in Fig. 8 (Fig. 10) appears.

In the registration window shown in Fig. 8, a user uploads the user information as described in 4.3.1. The first four fields are for the maximum north latitude, the minimum one, the maximum east longitude, and the minimum one. The next field is for storing schema. The names of columns of a table separated by commas are specified. The next is for the upload condition. The last field is for specifying the size of the minimum grid. When the right bottom button is pressed after specifying data, the window for confirming users’ data shown in Fig. 9 appears. The data specified can be confirmed in this window.

Fig. 10 shows the search window for specifying the condition of making a distribution map. The first four fields are the same as those of the user information: fields for the maximum north latitude, the minimum one, the maximum east longitude, and the minimum one. These four fields specify the whole area of a distribution map. The next four fields are for specifying the retrieval condition. The bottom three buttons are for selecting the target of retrieval: site, remain, and relic. Fig. 11 shows the search window under the selection of an item of retrieval condition. These items are “site_type,” “site_period,” “remain_type,” “remain_period,” “relic_keyword,” “ex_year,” “layer,” and “ex_order” as shown in 4.4.1. A value or string is specified in the right field. It is tested whether a value of an item is equal to the value specified in the right field. A pair of an item and a value constitutes a predicate. Four predicates can be specified in the window. These predicates are connected with logical ANDs. That is, when all of the predicates specified are satisfied, the data value becomes a candidate of the retrieval result. When the upper button of the two bottom ones is pressed, the distribution map is created.

An example of a distribution map created is shown in Fig. 12. The four by four sections appear on a map picture. The legend and the precise numbers of sections are also shown.

![Fig. 7. Selection window.](image)

![Fig. 8. Registration window.](image)
6. Concluding Remarks

This paper proposed the information sharing system for archaeology. Archaeological information sources are heterogeneous and distributed. This is because archaeologists independently store their excavation information. They have the information of excavations carried out by themselves. Excavation information cannot be opened until verification and research are completed. On the other hand, many archaeologists want to analyze the distribution of relics and/or remains.

The issues resolved are the treatment of closed information, the overcoming limited connection, and the treatment of heterogeneous information sources. By only handling the aggregate information, closed information can be treated. For conquering the limited connection, the data are uploaded during a computer in on-line. The common information structure was trying to be clarified for treating the heterogeneous information sources. We also clarified the structures of information treated in the server. These are the structures of user information and log data as well as the common schema. We designed and implemented the system. The system implemented is based on the client-server architecture. Client-side system sends the aggregate information. Server-side system creates a distribution map and manages the information. An example of an execution of the implemented system was shown. The system can create distribution maps without uploading detailed information. It is considered that the wide area distribution maps can be created.

Although we have clarified the data structures of the information on the server, there remain three issues in implementing the system: dealing with duplicate data, decomposition of query, and keeping integrity of information. Addressing to these issues is in future work.
We have not implemented the schema mapping capability yet. It supports the mapping between the client (local) schema and the common schema. A user specifies the mappings of data items of the client schema and those of the common schema. Introducing GUI in specifying the mappings may make the specification easy. This schema mapping capability is very important because it is inevitable in the information exchange between the client and the server. Its implementation is also in future work.

Acknowledgement

We give special thanks to Mr. Kengo Miyahara, Kyoto City Archaeological Research Institute, for his support and cooperation to this work as well as deep discussion with us.

References


