A METHOD FOR MAINTAINING ROUTING DATA IN AUTOMATED RECORD KEEPING SYSTEMS*

Dorothy E. Denning

Computer Science Department
Purdue University
W. Lafayette, Indiana 47907

An economical method for maintaining routing records in an automated record keeping system is described. Routing records are stored separately from data records on secondary storage devices. A "time stamp" is stored in each routing record, and all routing records associated with the same data record are linked by a "time link". The method is shown to have better performance and lower impact on the remainder of the system than several alternative strategies.

Introduction

While reading the Report of the Privacy Protection Study Commission[1], I began wondering how one might implement the requirements for routing data. The Privacy Act of 1974 mandates the use of "routing data" in all record keeping systems maintained by Federal agencies. Routing data, also called "tracking data" or "traffic data", provides an account of the recipients of data stored in a record associated with some individual. The purpose of the routing data is to provide the individual with an accounting of the uses and disclosures of his record on request, and to facilitate the propagation of corrections to the recipients and sources of the information. Section 3(c) of the Privacy Act states that the routing data must include "the date, nature, and purpose of each disclosure ... and the name and address of the person or agency to whom the disclosure is made," and that the routing data must be maintained for five years.

The technological implications of the requirement for routing data are obviously enormous. Existing record keeping systems might require costly restructuring to accommodate routing data. the amount of storage needed to save routing data for five years and the time required to search it for data pertaining to a particular individual could be immense. Finally, the maintainance of the routing data could severely impact the performance of the other operations of the system.

In this note, I would like to share with the reader the thoughts I had while thinking of different strategies for storing and retrieving routing data. The strategy I like best is based on "time stamps". The routing records are maintained separately from the data records on secondary storage devices. Each routing record has associated with it a time stamp, and all records associated with the same data record are linked by a "time link". The only information that is recorded with a data record is the time of its most recent disclosure.

I shall first outline a simple model of a record keeping system and state the minimum requirements I believe any routing system must meet. I shall then describe and evaluate the time-linked stragety. Finally, I shall compare the strategy with several alternative approaches.

Model and Requirements

An automated record keeping system consists of a collection of data records, a set of operations for retrieving and maintaining the data records, and a routing subsystem. Each data record contains information about some individual identified by the record, and no individual is represented by more than one record. The routing subsystem consists of routing records together with operations for retrieving and maintaining them. A routing record is created whenever a data record is disclosed and contains the associated individual's identifier, the date, the nature of the disclosure, and the recipient. A routing record may also be created whenever a data record is updated; in this case it gives the nature and source of the update. Three assumptions are made: 1) storage and retrieval of data records is performed much more frequently than storage and retrieval of routing records, 2) routing records are created much more frequently than they are retrieved, and 3) the number of routing records created could be inordinate.

The requirements of a record keeping system fall into three areas: 1) the initial integration of the routing sybsystem into the record keeping system, 2) the performance of the record keeping operations which retrieve and maintain the data records, and 3) the performance of the routing subsystem. It should be possible to integrate a routing subsystem into an existing system without reorganizing the entire data base.

*Work reported herein was supported in part by NSF Grant MCS77-04835.

It should also be possible to reorganize the data records without reorganizing all of the routing records. The routing subsystem should not severely degrade the performance of the record keeping operations. Since storage and retrieval of data records is likely to be performed much more frequently than storage and retrieval of routing records, it is preferable to optimize the former at the expense of the latter rather than vice-versa.

The performance requirements of the routing subsystem fall into four areas: a) creation of new routing records, b) retrieval of some or all of the routing records corresponding to a given data record, c) purging of obsolete routing records, and d) reorganization of routing records. Since creation is likely to be performed much more frequently than the other operations and is performed in conjunction with storage and retrieval of data records, the performance of this operation is critical. On the other hand, retrieval is probably performed infrequently, so that some loss of efficiency is tolerated here. However, it still must be possible to retrieve a routing record in some "reasonable" period of time (searching 1,000 tapes for 1 record is not reasonable!). It should be reasonably easy to purge obsolete records from the system. Finally, it must be possible to reorganize some or all of the routing records; e.g., to transfer them to less expensive hardware.

Different design strategies for routing subsystems will be evaluated in terms of their impact on the system in these areas. The discussion will not consider those aspects of the system that are independent of the strategy used; e.g., when and what routing information should be recorded or the exact format of the routing records.

Time-Linked Strategy

Description

The basic idea is very simple. The routing records are stored separately from the data records on one or more secondary storage devices (e.g., disks, tapes, etc.). Included in each record is a TIME field which gives the time (including date) when the record was created (i.e., a time stamp), an ID field which identifies the individual of the corresponding data record, and a TLINK field which gives the time of the last routing record for this individual. Since a routing record is required to contain the date and identification anyway, the only additional field required is the TLINK field. A TLINK field is also included in each data record giving the time of the most recent routing record. Thus all of the routing records corresponding to a particular data record are linked on a chain beginning with the TLINK field in the data record and following the TLINK fields in the routing records.

The value in the TIME field of a routing record need not be unique; if two routing records have the same value for TIME, they can be distinguished by ID. The precision used for time values would depend on the amount of routing activity on the system and could be at the second, minute, or even day level.

Since the TLINK field of a record gives the time rather than the physical location of its successor, a method is needed to map the time into a location. This is done with a Routing Index Table (RIT) which gives a list of (time, location) pairs: (t_1,loc_1), ... , (t_n,loc_n). Each (t_i,loc_i) pair gives the location loc_i of all routing records created between time t_i and t_{i+1} (for i=n, between t_i and the current time). The location loc_i is the name of a routing file; e.g., a logical file name, tape identifier, disk identifier, etc. The routing records stored at loc_i may be ordered chronologically, though this is not essential. Figure 1 illustrates the use of time-links to chain together two routing records associated with an individual identified as "X".

The four basic routing operations: creation, retrieval, purging, and reorganization are performed as follows. To create a new routing record for an individual X, the current time t is placed in the TIME field, the TLINK field in the data record for X is copied into the TLINK field of the routing record and is then replaced by t, and the routing record is added to the end of the current routing file. When the routing file reaches capacity, a new file is initialized and an entry placed in the RIT.

To retrieve one or more of the routing records associated with an individual X, the chain of TLINK values is followed beginning with the data record for X. To retrieve a record created at time t, the RIT is first searched for a pair (t_i,loc_i) such that $t_i \leq t \leq t_{i+1}$ (for $1 \leq i < n$) or $t_i \leq t$ (for $i = n$). Then the routing file at loc_i is searched for the routing record with TIME = t and ID = X.

To purge obsolete routing records created before time t_i $(1 < i \leq n)$, it is sufficient to destroy the routing files at locations loc_1,..., loc_{i-1} and remove the pairs (t_1,loc_1), ..., (t_{i-1},loc_{i-1}) from the RIT.

To reorganize the routing records onto different storage devices, it is sufficient to update the RIT. The routing records require no modification.

One final observation: the time stamps are much like logical or "virtual" addresses and the RIT like a virtual memory mapping table. The primary difference is that the time-addresses do not contain offsets. The reason offsets are not included is to facilitate reorganization of the routing records. Furthermore, offsets are of little value for retrieving a record from a sequential file.

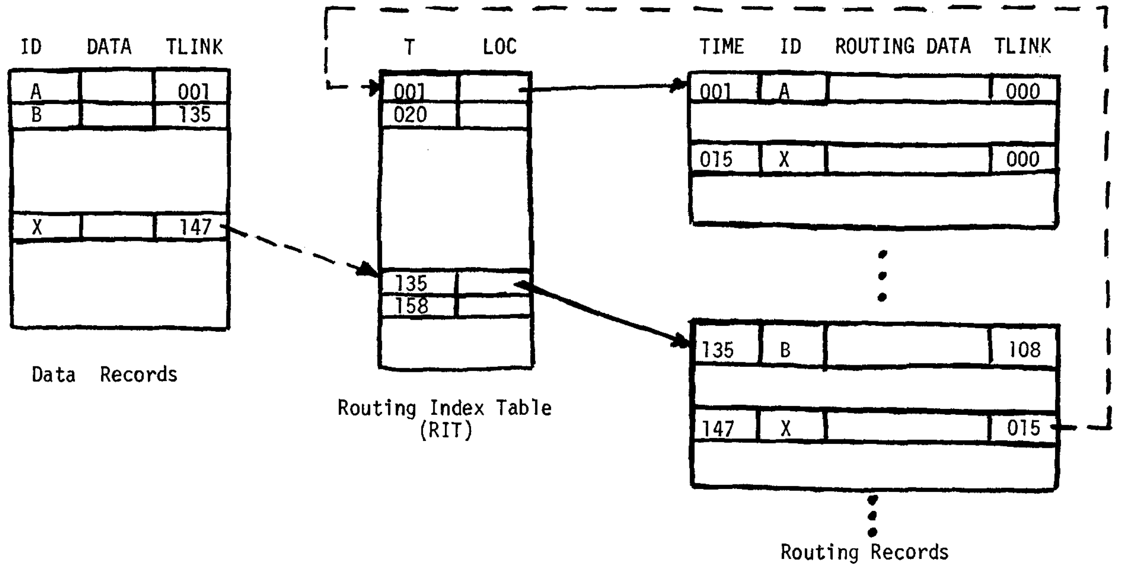

Figure 1. Routing Subsystem Based on Time-Links. (Solid arrows represent address-links; dashed arrows represent time-links.)

Evaluation

The impact of the routing subsystem on integration, performance of the record keeping operations, and performance of the routing subsystem is quite low. The cost of integrating the subsystem into a new or existing record keeping system is minimal: only one field must be added to the data records and the RIT initialized. If the addition of the TLINK field to the data records presented a problem, a separate table of (ID, TLINK) pairs could be created.

The routing subsystem should not degrade the performance of the record keeping operations. The additional storage requirements for the data records and the RIT are but a small fraction of the total data record storage, and the operations which retrieve and maintain the data records are not affected by the presence of the routing system.

The performance of the routing subsystem should be quite acceptable. The most frequently performed operation, creation, requires only a few simple steps. Purging is likewise simple.

The effort required to reorganize is dominated almost exclusively by the operation of physically moving the records; little effort is required to update the RIT. Because of the low cost, it would be possible, for example, to allocate space for new routing records on a high speed storage device; when the device is full, the records can be migrated to a slower speed device.

The effort required to retrieve a single routing record is dominated by the time to search the RIT for the appropriate routing file and then search the routing file for the record. If the RIT is chronologically ordered, it can be searched in $O(\log_2 n)$ time using a binary search. For example, if the routing data is stored in 100 files, then the correct file can be determined in just 7 probes into the table. The time required to locate the correct record within a routing file depends primarily on the type of device on which the file resides and to a lesser extent on the amount of precision used for the TIME field. If the file is chronologically ordered on a random access device, a binary search can again be used to retrieve the record in logarithmic time. Otherwise, if the file resides on a sequential device, a linear search must be made. If enough precision is used for the TIME field to make time values unique, then no additional searching is required; otherwise, several records may require inspection to locate the one(s) with the proper ID. The conclusion is that although the time required to locate routing records is much greater than the time required to create and purge records, it is acceptable given that the operation is not performed frequently. Five routing records, stored in a routing library containing 1,000 tapes can be located in no more time than it takes to read 5 tapes.

To summarize, the total impact of a time-linked routing subsystem on a record keeping system is very low. The cost of integrating it into

	Strategy			
Requirement	Time Linked	Address Linked	Integrated	Dump
1. Integration	0	0	●	0
2. Performance of Record Keeping Operations	0	0	●	0
3. Performance of Routing Subsystem				
a. Creation	0	0	◒	0
b. Retrieval	◒	◒	0	●
c. Reorganization	0	●	●	0
d. Purging	0	0	●	0
OVERALL	0	◒	●	●

Table I. Relative Evaluation of Strategies for Implementing a Routing Subsystem (0 - good, satisfies the requirements; ◒ - fair, some deficiencies; ● - poor, probably unacceptable except under special circumstances).

the system and the overhead imposed on the record keeping operations is low. The cost of maintaining the routing records is low for critical operations and acceptable for retrieval.

Comparison with Other Strategies

The time-linked strategy will be compared with three other strategies: 1) an address-linked strategy, 2) an integrated strategy under which the routing data is integrated into the data records, and 3) a dump strategy under which the routing records are simply dumped into files without linking. The results are summarized in Table I.

Address-Linked Strategy

The address-linked strategy is similar to the time-linked strategy, but the routing records are linked by physical location rather than by time. The location gives the starting address of a device and, for direct access devices, an offset within the device. It is important to note that the location is a physical rather than logical one and includes an offset; otherwise it would be equivalent to the time-linked method.

The cost of integrating an address-linked strategy into a record keeping system is small as for the time-linked strategy. A field is required in the data records to point to the location of the first (most recent) routing record, though no RIT table is required. The impact of the strategy on the record keeping operations is likewise minimal.

Retrieving routing records under address-linking will be slightly faster than under time-linking since no mapping of logical to physical addresses is required. Records stored on direct access devices can also be retrieved faster since searching is not required. However, the time to retrieve records on sequential devices is the same for both strategies. Thus, if the bulk of the routing records are stored, for example, on tape, there will be no significant difference in the retrieval time for the strategies, since the retrieval time will be dominated by the tape search.

The address-linked strategy has one disadvantage; reorganizing the routing records requires updating of all links -- a nontrivial operation. The necessity to update the links rules out the possibility of creating new routing records on high-speed devices and later migrating them to slower-speed devices without incurring considerable overhead. Also, purging obsolete records requires some sort of time stamp to be associated with the routing files anyway. For these reasons, the time-linked strategy is given a higher overall rating than the address-linked one.

Integrated Stragegy

An integrated strategy stores the routing records in the central data base along with the data records. This approach has the advantage of making retrieval of routing records very fast. However, in all other aspects it is inferior to either of the linked approaches. Integrating the routing subsystem into the record keeping system could be extremely difficult and costly. The performance of the record keeping operations could be severely degraded due to the increased size of the central data base. The presence of unused routing records is likely to interfere with data retrieval and maintainance operations.

The task of performing critical routing operations is also more costly than under the linked strategies. Severe overflow problems arise if a newly created routing record does not fit into the amount of space allocated in the data record. Purging of obsolete routing records requires a search of the entire data base and possible reorganization of the remaining routing records. Thus there appears to be no significant advantage to an integrated approach.

It is also worthwhile to observe that the same problems arise with any scheme under which the routing records are organized by individual rather than by time, even if the records are maintained outside of the central data base. Unless routing records are frequently retrieved, the benefits of low retrieval costs are lost.

Dump Strategy

The third strategy considered here is that of simply dumping all routing records into files without linking them to their corresponding data records. This strategy has several obvious advantages. No changes are required to the data records and the process of creating, reorganizing, and purging routing data is minimal. However, it does have the serious disadvantage that the cost of retrieving a routing record could be prohibitively expensive. The method has merit only for systems which generate a sufficiently small number of routing records that exhaustive search is not out of the question.

Conclusions and Future Research

The conclusion is that routing records are best organized by some linking strategy and that logical addresses based on time are preferable to physical addresses. However, there are still several questions worth investigating. For example,

Cost

What is the cost of employing the time-linked strategy)? It would be useful to develop a model in which the cost could be estimated in terms of the quantity of routing data, the frequency of retrieval, the types of storage devices used, etc.

Encryption

Does encryption of the routing data affect the choice of strategy and how should the data be encrypted? Preliminary investigation suggests that the time-linked strategy would fit nicely into an environment where data and routing records are routinely encrypted. For example, the TIME field of each routing record could be left in plaintext and the remainder of the record encrypted. Purging and reorganization of the routing records could then be performed without decrypting the records. Retrieval would require decrypting only the routing record(s) with the desired time stamp. The encryption key could be the same for all routing records or could be unique to the individual associated with the record.

Multiple Record Disclosure

How should the disclosure (update) of several data records at once be handled? Rather than creating a separate routing record for each data record, it might be preferable to create a single routing record to represent the entire disclosure. If the data records are logically related into a group, it might be possible to associate an additional time-linked chain with the group.

Acknowledgements

I am grateful to Peter Denning and James P. Anderson for their comments and encouragement.

References

1. The Report of the Privacy Protection Study Commission, Appendix 5, Technology and Privacy, July 1977, U.S. Government Printing Office, Stock No. 052-003-00425-9.