MARTIN MARIETTA ENERGY SYSTEMS LIBRARIES

3 4456 0355066 3

ORNL/TM-11749

ornl

OAK RIDGE
NATIONAL
LABORATORY

MARTIN MARIETTA

An Error Detection and Correction Strategy for a Decision Support System

V. Ng

OAK RIDGE NATIONAL LABORATORY
CENTRAL RESEARCH LIBRARY
CIRCULATION SECTION
4500N ROOM 175
LIBRARY LOAN COPY
DO NOT TRANSFER TO ANOTHER PERSON
If you wish someone else to see this report, send in name with report and the library will arrange a loan.
UCN-7969 3 9-77

MANAGED BY
MARTIN MARIETTA ENERGY SYSTEMS, INC.
FOR THE UNITED STATES
DEPARTMENT OF ENERGY

This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from the Office of Scientific and Technical Information, P.O. Box 62, Oak Ridge, TN 37831; prices available from (615) 576-8401, FTS 626-8401.

Available to the public from the National Technical Information Service, U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22161.

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

ORNL/TM-11749

Energy Division

Center for Transportation Analysis

AN ERROR DETECTION AND CORRECTION STRATEGY FOR A DECISION SUPPORT SYSTEM

Vickie Ng

Date Published: June 1991

Research sponsored by the
Military Traffic Management Command
under Interagency Agreement DOE 1405-1405-A2

Prepared by the
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831
Managed by
MARTIN MARIETTA ENERGY SYSTEMS, INC.
for the
U.S. DEPARTMENT OF ENERGY
under contract DE-AC05-84OR21400

MARTIN MARIETTA ENERGY SYSTEMS LIBRARIES

3 4456 0355066 3

TABLE OF CONTENTS

PAGE

ACRONYMS v

ABSTRACT vii

1. INTRODUCTION 1

2. ERROR DETECTION AND CORRECTION STRATEGY 3
   2.1 PROBLEM STATEMENT 3
   2.2 METHODOLOGY 3

3. EXAMPLE 7
   3.1 BACKGROUND OF WHIST-MOD 7
   3.2 EXTERNAL DATA SOURCES IDENTIFIED 8
   3.3 TOTAL DATA QUALITY MANAGEMENT FOR WHIST-MOD 9
   3.4 ERROR DETECTION AND CORRECTION STRATEGY FOR DATA INTEGRATION 10
      3.4.1 Approach 12
      3.4.2 Strategies for Specific Instances 13
         3.4.2.1 Case 1: Matching GBL numbers 13
         3.4.2.1 Case 2: Unmatched GBL numbers 16
      3.4.3 Implementation Method 19

4. SUMMARY 22

REFERENCES 24

ACRONYMS

DBC	Database Computer
DBMS	Database Management System
DoD	Department of Defense
DSS	Decision Support System
FOIA	Freedom of Information Act
GAO	Government Accounting Office
GBL	Government Bill of Lading
LAN	Local Area Network
MTMC	Military Traffic Management Command
MTPP	Personal Property Directorate
PP	Personal Property
PPM&S	Personal Property Moving and Storage
PPSO	Personal Property Shipping Office
QA	Quality Assurance
TOPS	Transportation Operational Personal Property Standard System
WHIST-MOD	Worldwide Household Goods Information System for Transportation Modernization

ABSTRACT

Decision support systems (DSSs) require data generated from external sources. To be effective, a DSS needs a good data quality management program to ensure that every user has quality data on which to base his/her decisions. Hence, data integration is very important in such a system. A major issue of data integration is that given one or more objects that contain members with distinct keys, one must be able to determine with some fair degree of certainty

(1) whether two members from different objects having distinct keys (due to errors) are indeed functionally duplicates, and

(2) whether two members from different objects having the same keys (due to errors) are actually functionally discrete.

In this paper, an error detection and correction strategy based on distance measures with weights is proposed for data integration. A real application example developed for the Worldwide Household Goods Information System for Transportation Modernization (WHIST-MOD) project for the Personal Property Directorate (MTPP) of the Military Traffic Management Command (MTMC) is also presented.

1. INTRODUCTION

Decision Support Systems (DSSs) can be effective only if the level of accuracy and integrity of the necessary data is proportional to the decision support requirements. Data quality requirements differ substantially from one system to another. In order to determine the requirements for data quality, one must consider the costs and benefits of decisions made using the data in the system. The impact the quality of the stored and retrieved data has on the decisions is also an important factor.

There have been many procedures and controls developed for ensuring data quality in centralized environments. However, the current trends in information systems such as distributed systems have increased the difficulty of maintaining appropriate levels of data quality. In particular, in a DSS environment where those who use the data are not necessarily those responsible for collecting and maintaining them, data ownership becomes an issue in maintaining overall data quality.

Despite the difficulties involved, the imperative for ensuring data quality has not diminished. This is because too often computer errors are consequences of data errors. Bad decisions can be made if they are based on incorrect data or incomplete information.

In systems where most of the data are externally generated, error detection and correction techniques should be applied at the time of data integration in order to control data quality before data get entered into the system. As data are merged and integrated to form a DSS, discrepancies on data values might occur because of entry errors or merely because of differences in interpretation. Differences in interpretation or definition can easily be resolved if data acquisition requirements are adequately specified in the first place.[1] On the other hand, entry errors can be hard to identify.

When records are merged from different systems based on a single data value, how can one tell if a "value not found" situation is not a result of an error in either or both systems? Furthermore, how can one be sure that a match of a data value is truly a "correct" match? Section 2 of this paper attempts to address these issues of data integration and discusses a method to detect and possibly correct these kinds of discrepancy errors. As an example, Section 3 presents a DSS in which the proposed error detection and correction method could be applied. Finally, Section 4 contains some concluding remarks on the proposed design.

2. ERROR DETECTION AND CORRECTION STRATEGY

2.1 PROBLEM STATEMENT

Given one or more objects that contain members with distinct keys, can one determine with some fair degree of certainty

(1) whether two members from different objects having distinct keys (due to entry errors) are indeed functionally duplicates, and

(2) whether two members from different objects having the same keys (due to entry errors) are actually functionally discrete.

As mentioned earlier, the level of data quality requirements depends on the functionality of the individual system. For a DSS, the level of aggregation might be a determining factor for the level of data accuracy and integrity. Through the use of data dictionaries and perhaps through the referential integrity checks supported in the database management system (DBMS), data integrity can be enforced. However, if data are aggregated incorrectly because of data inaccuracy that leads to a "false" match or incomplete information, the decision-making process is not fully supported. Furthermore, this DSS will not be used anymore because the system lacks trustworthiness. The primary goal is, to the extent possible, to capture all the necessary data without introducing errors into the system.

2.2 METHODOLOGY

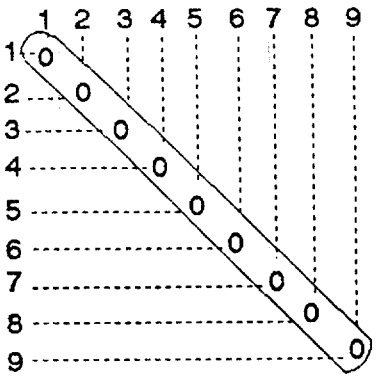
A great deal of research has been done on data quality control.[2] The problem of error detection and correction involves elements of pattern recognition and coding theory. Different techniques have been introduced over the years. Examples are check digits and check equations,[3] diagram and trigrams,[4] bit vector, distance measures, and longest common substrings algorithms.[5] Recently, concepts of fuzzy set theory have been implemented to deal with error correction problems.[6]

The proposed approach is based on distance measures. The idea is to evaluate the degree of dissimilarity in order to determine if a record is to be automatically corrected. Let's assume that there is always a data set that is always more accurate than the others. In this context, "automatically corrected" means data elements with bad values will be updated with values from the more accurate set. If a record is determined to contain so many discrepancies that it cannot be automatically corrected, it will be rejected from the merge and printed on hardcopy. A user can then decide whether it should be added as a new record.

Distance measure is a technique for pattern matching. This technique has also been applied to biological sequences comparisons.[7] Two patterns are considered to be a perfect match if they have a distance measure of 0. Any value other than 0 indicates that the patterns are not the same. The greater the distance measure, the more dissimilar the two patterns are. For example, a comparison of the two patterns '123456789' and '123456789' has a distance measure of 0. However, a comparison of '123456789' and '123356789' has a distance measure of 1. The two patterns '123456789' and '123556759' have a distance measure of 4.

Two patterns to be matched are arranged in a matrix format, with one pattern as the vertical, the other as the horizontal. The antidiagonal elements of this matrix will have an offset of 0 if these two patterns are identical. Distance measure is calculating how much offset a pattern has with this anti-diagonal. A distance offset is calculated by taking the absolute value of the difference between the positions of occurrence in the two patterns. A summation of the distance offsets of all the positions in the pattern will give a distance measure.

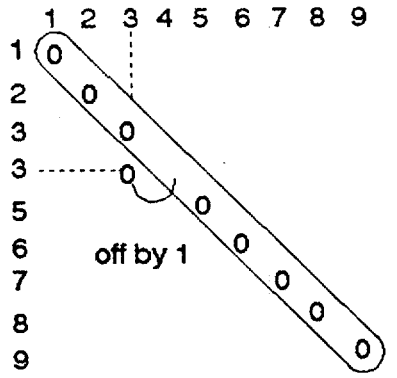
In our approach, the distance measure is computed a little bit differently. For error detection purposes during data integration, knowing how many characters that are not the same is more important than knowing how different the characters are. Hence, our distance measure is a count of all the distance offsets that are non-zero. For example, in our earlier example, patterns '123456789' and '123556759' will have a distance measure of 2 instead of 4. This validation concept is illustrated in Figure 1.



Distance measure = 0

For WHIST-MOD,

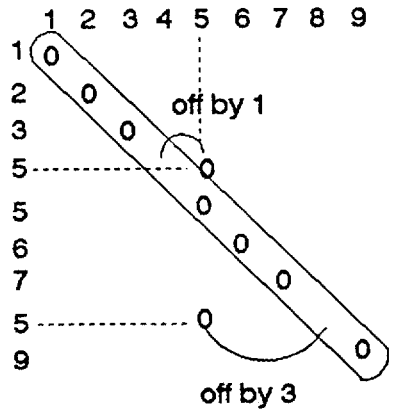
Distance measure = 0



Distance measure = 1

For WHIST-MOD,

Distance measure = 1



Distance measure = 4

For WHIST-MOD,

Distance measure = 2

Figure 1 : Distance Measure Validation Concept

The distance measure tells us how different two objects are. But should a system reject all objects that are not the same? For total accuracy, a system may be designed to integrate only if objects have a perfect match. However, for practical reasons, a system may need some kind of threshold or cutoff limits to determine which erroneous data can be accepted without negative implications. Thresholds are imposed mainly through inspection and frequency distribution analysis of the data set. Thresholds should be chosen in such a way that the acceptance levels for all data in the system are consistent.

Another technique we use in our methodology is a weighting function. We believe that it is not sufficient to determine integration discrepancy based on comparisons of a single data element. Other elements that are essential to the object should also be checked. In addition to the importance of the element, the existence of finite valid values is another criterion for choosing which elements to be examined. For all of the data elements to be examined, weights are assigned based on the significance of the elements to the object and the possible error rate of the elements. Our methodology of determining if integration discrepancies are recoverable is basically to compare values for several weighted data elements using distance measures. In the next section, an example is presented to demonstrate how this methodology is applied.

3. EXAMPLE

3.1 BACKGROUND OF WHIST-MOD

The Personal Property Directorate (MTPP) of MTMC is charged with the overall management of the Personal Property Moving and Storage (PPM&S) program for the Department of Defense (DoD). MTPP's functions are to set policy, interpret and disseminate regulations, evaluate program effectiveness, and implement the rate solicitation and carrier approval process of the PPM&S program. However, MTPP has found that the present automated system -- the Worldwide Household Goods Information System for Transportation (WHIST) -- is inadequate for the growing information needs.

The purpose of the WHIST-MOD project is to provide research and development support for the modernization and enhancement of the current WHIST system. WHIST-MOD is primarily a DSS that is flexible enough to track the effects of evolving policy within the Personal Property (PP) program while offering easy access to a variety of data to novice and experienced users.[8]

The WHIST-MOD system requires a large database for querying and reporting by different WHIST-MOD users having various levels of access to the database. MTPP's information needs are very diverse. MTPP is required to respond to information requests from the military services, Congress, the Government Accounting Office (GAO), the Defense Manpower Data Center, the carrier industry, and individual service members through the Freedom of Information Act (FOIA).[9] Occasionally the military services require data from MTPP to support special studies for budget preparation. The GAO and the Defense Manpower Data Center request actual shipment cost and claims data from MTPP to conduct studies on military personnel movement expenses and entitlement. Other information requests initiated by Congress or the carrier industry demand data on individual shipments, such as shipment pickup and delivery data.

In addition to answering to external organizations, MTPP needs data to perform its own management tasks. It requires data on carrier performance to support the carrier approval process and to monitor the effectiveness of the PP Quality Assurance (QA) program. It requires data on certain types of

shipments to study the impact of policy changes. For example, information on mobile home shipments is currently being collected for a policy evaluation study.

3.2 EXTERNAL DATA SOURCES IDENTIFIED

A personal property shipping office (PPSO) is responsible for processing the moving and storage requests of service members under the policies established by MTPP. The PPSO implements the MTPP program. With the installation of the Transportation Operational Personal Property Standard System (TOPS) in each PPSO, high-quality shipment description data will be available electronically. Shipment and shipment QA data are entered into the TOPS database at each PPSO, and at the end of processing each day these data are transmitted via modem to the TOPS central machine, known as the TOPS Switcher machine. The TOPS Switcher machine transfers to WHIST-MOD a copy of the data received from all PPSOs. The WHIST-MOD system then extracts the data elements required by WHIST-MOD, converts them into appropriate WHIST-MOD database format if necessary, and performs validation checks to ensure good data quality.

Although TOPS can provide all pertinent shipment-descriptive and QA data to WHIST-MOD, it does not store any financial or accounting information. As mentioned in the previous section, WHIST-MOD requires the actual shipment cost and claims data. Such data have to be obtained from the military Finance Centers and Claims Centers on computer magnetic tapes, which are created from data manually entered into one of the MTPP machines.

Another data source identified for WHIST-MOD is the carrier industry. The carriers are invited to bid for moving and storing personal property of military service members. Several methods can be used to transport DoD-sponsored personal property shipments. Rates for each method are submitted by the shipping carriers approved by MTPP. The carriers enter their bids for shipping charges on computer magnetic tapes or paper forms. WHIST-MOD has to compile the rates data for the rate solicitation function of MTPP. It then broadcasts the rates and other traffic management information to all PPSOs via the TOPS Switcher machine.

3.3 TOTAL DATA QUALITY MANAGEMENT FOR WHIST-MOD

WHIST-MOD's primary function is decision support. The WHIST-MOD users need to have a database with good quality data on which their queries are based. Since most of the WHIST-MOD data originate from different sources, the WHIST-MOD system has no control over how accurate the data are. One way to ensure that only good data are contained in the WHIST-MOD database is to perform a thorough validation at the time of acquisition.

Types of errors that are common to all data sources are errors in content and errors in interpretation. Errors in content can result either from human errors or from errors in data transmission. An example of errors in content is transposed numbers. Errors in interpretation are less obvious. They are often undetected except by users who are very familiar with the data and their definitions. Errors in interpretation basically are results of poor data definitions.

WHIST-MOD employs a variety of methods to ensure total data quality. It is believed that all data should be validated and corrected before their inclusion in the database. Data that are entered directly into the WHIST-MOD database by MTPP staff will be validated upon entry. All external data will be validated and either corrected or filtered out at the time of acquisition. In other words, TOPS shipment data from the PPSOs and the financial data from the military services' Claims and Finance Centers will be checked for data accuracy and integrity. Moreover, data from different sources are examined for possible discrepancies upon integration. Exception reports will be generated whenever the uncovered discrepancies cannot be automatically corrected.

Column, row, and table level validations are applied to the data to ensure data accuracy and integrity. The following are the general types of data validations WHIST-MOD, or any other DSS, should perform:

(1) Format check (column level)
- data type check
- length requirement check
- data format check

(2) Valid values check (column level)
- List of values check
- Min/Max value check

(3) Elements relationship check (column level)

(4) Record completeness check (row level)
- Data value dependent mandatory columns check
[i.e., certain columns are mandatory (must exist) depending on the value of some particular data elements]

(5) Referential integrity check (table/row level)
- Table-dependent referential integrity check
[i.e., existence of a row (R1) in a table (T2) depends on the existence of the row (R1) in another table (T1)]

- Data-value-dependent referential integrity check
[i.e., existence of a row/rows (R1,...) in a table (T2) depends on the value of certain data element (E1) in another table (T1)]

(6) Discrepancy check (column level)
- Data discrepancy check between different input sources
[i.e., value of a data element can only have one value at a given time, independent of its input source]

Since it is almost impossible for a system to automatically recognize errors in interpretation, WHIST-MOD is going to use a range of tools to maintain data quality. Audits and standard reports will be produced periodically. These reports can then be scanned for problems in usage or interpretation by individuals who are familiar with the data. Another tool WHIST-MOD uses is a system dictionary. The system dictionary contains the exact definition and usage of the data and gives users guidelines by which data should be entered.

3.4 ERROR DETECTION AND CORRECTION STRATEGY FOR DATA INTEGRATION

Two of the WHIST-MOD data sources are the TOPS system and the Finance Center/Claims Center. Shipment-descriptive data from the TOPS system are transmitted electronically to WHIST-MOD. The shipment accounting data come on magnetic tapes that are created from data manually entered by data entry clerks. There might be discrepancies when the shipment cost/claims data from the Finance Centers and Claims Centers are merged with the shipment-descriptive data from TOPS. One

important task of data administration is to ensure that records from different sources agree before the data are integrated. The value of a data element can have only one value at a given time, independent of its input source. If there are significant discrepancies between elements, the records will be rejected from data integration and will be printed on hardcopy. Otherwise, the system will attempt to self-correct the problem and continue with the integration. The objective is to eliminate incomplete data and to support the decision-making process. First, the assumptions on which our error detection and correction strategies are based are as follows.

(1) Shipment-descriptive data would exist in the TOPS' database before the shipment cost/claims data get extracted from the Finance Centers and Claims Centers' tape. Otherwise, the transaction will be an insert, not a merge. An insert will occur only in the interim when TOPS is not completely fielded.

(2) It is a fact that shipment data from TOPS are more accurate than the data from the Finance Centers and Claims Centers, because critical data elements are either automatically generated or thoroughly validated in TOPS.

In order for the system to make a reasonable decision on data acceptance, a threshold has to be set. In other words, if a comparison has a distance measure that exceeds the threshold, the patterns are considered "unacceptably" different. A constant threshold can be chosen for acceptance criterion. However, the acceptance level will not be consistent if a fixed threshold is set for all patterns to be matched. For example, let's say a threshold of 2 is set for two patterns in question, one having a length of 4 while the other has a length of 9. The acceptance level for the pattern of shorter length will be 50% whereas that of the other pattern will be about 80%. In order to ensure the data are of good quality, a standard acceptance level must be established for the system. Threshold or cutoff limits are imposed largely by inspection and frequency distribution analysis. For our error detection and correction process, 80% is suggested as a heuristic acceptance level. In other words, records from the two different sources are considered to represent the same shipment record if 80% of the values in those records match. A threshold dynamically set at 20% of the length of the pattern is hence used to meet this acceptance level.

3.4.1 Approach

Since the shipment's Government bill of lading (GBL) numbers and member's social security numbers are the key identifiers of shipment records, comparing values for these two elements will give a good indication of potential problems. However, because of possible human errors on data entry, additional elements will have to be checked to identify irrecoverable discrepancies. It is believed that integration discrepancy errors can be found by examining three more essential data elements. Elements that are essential to traffic management include the destination, the shipping carrier, and the weight of the shipment. Hence, our basic method of determining if discrepancies are recoverable is to compare values for five weighted data elements. Weights are assigned based on the significance of the elements to the shipment record and on the possible error rate of the elements. For example, MEMBER_SOCIAL_SECURITY_NUMBER is assumed to be more significant and to have the possibility of fewer errors than DESTINATION_GBLOC. Listed in decreasing order of relative importance, the five data elements to be examined are

(1) GBL_NUMBER,
(2) MEMBER_SOCIAL_SECURITY_NUMBER,
(3) CARRIER_CODE,
(4) ORIGIN_NET_WEIGHT, and
(5) DESTINATION_GBLOC.

If a GBL number is within the range of numbers known to be assigned to TOPS, the shipment record represented by the GBL number should already exist in TOPS. There is a possibility that the GBL numbers from the Finance/Claims center tape do not match any from TOPS. Even if the GBL number matches one of the GBL numbers in TOPS records, it is possible that the GBL numbers actually represent two different shipment records. Employing the methods of distance measures and weighted matching will help avoid the problem of incomplete data resulting from missing records.

3.4.2 Strategies for Specific Instances

Data integration strategies are proposed for each of the following cases.

(1) When the GBL number from the Finance/Claims Center tape **matches** the GBL number in TOPS and

(2) When the GBL number from the Finance/Claims Center tape does **not match** any of those in TOPS.

The rules and actions suggested for the two cases are different. The rules for Case 2 are a little more stringent than those for Case 1. The reason is that for Case 1, the extra checking is performed to identify that the records are a true match whereas for Case 2, the additional element checks are used to find a match. The rules have to be stringent enough to allow the system to suggest a reasonable match.

3.4.2.1 Case 1: Matching GBL numbers

The integration process designed for Case 1 is as follows.

1. Elements with higher weights are examined first. Based on the assigned weights, the elements are checked in the following order:

   (1) MEMBER_SOCIAL_SECURITY_NUMBER (SSN),
   (2) CARRIER_CODE (CARRIER),
   (3) ORIGIN_NET_WEIGHT (WEIGHT), and
   (4) DESTINATION_GBLOC (GBLOC).

2. Depending on the degree of discrepancy, there are basically three corrective actions:

   (1) exact match, merge;
   (2) conditional merge, with values replaced by TOPS data; and
   (3) record rejected, no merge; print on hardcopy.

In order to make sure that a value match of the primary element is not the result of an error, supplementary elements are checked according to the rules and actions listed below in Table 1.

Table 1. Rules and Actions for Case 1

RULE	ACTION
If 3 elements do not match	Reject, no merge; print on hardcopy.
If 2 elements do not match	
If both elements are within threshold and if none of the 2 elements are high weighted;	Conditional merge.
otherwise	No merge.
If 1 element does not match	
If the element is within threshold or if the element has the lowest weight;	Conditional merge.
otherwise	No merge.
If 0 element does not match	Accept, merge.

Detailed process flows for Case 1 are shown in Figure 2 on the following page.

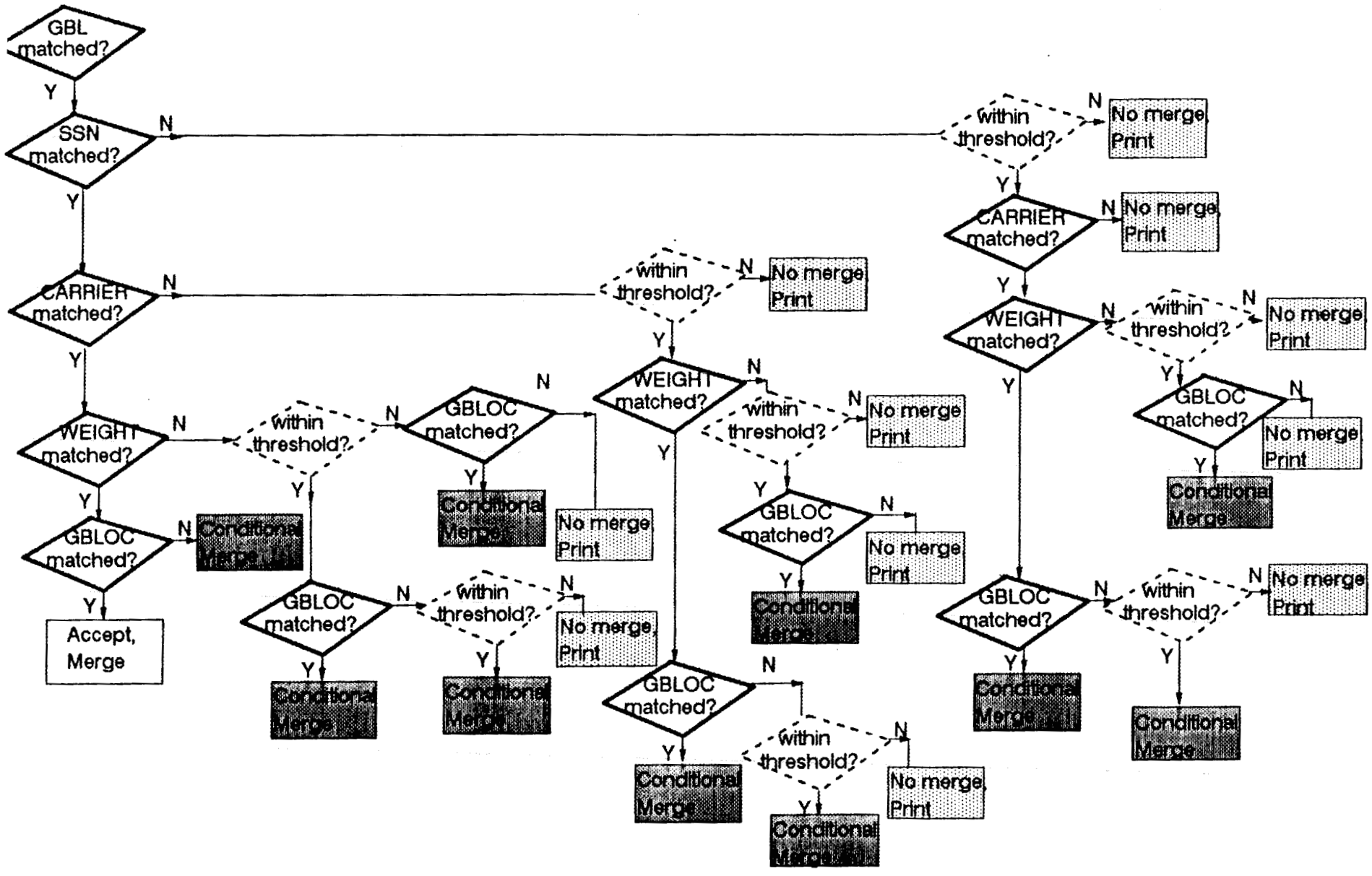


Figure 2 : Matching GBL Numbers

3.4.2.1 Case 2: Unmatched GBL numbers

The integration process designed for Case 2 is shown step by step as follows.

1. A distance measure of 1 is used to get a list of closest possible matches; i.e., get a list of GBL numbers that are 1 position different from the one to be matched.

2. For each possible match, four additional data elements are checked to confirm a reasonable match. Elements with higher weights are examined first. Based on the assigned weights, the elements are checked in the following order:

   (1) MEMBER_SOCIAL_SECURITY_NUMBER (SSN),
   (2) CARRIER_CODE (CARRIER),
   (3) ORIGIN_NET_WEIGHT (WEIGHT), and
   (4) DESTINATION_GBLOC (GBLOC).

   A score of 0 to 10 is assigned to each of the possible match. This scoring system is used to guide the system to a best possible match.

3. The record with the highest score is kept.

4. Depending on the degree of discrepancy, there are basically four corrective actions. Each of the scores assigned earlier is associated with an action:

Score	Action
>10	Exact match, merge;
6-10	Conditional merge, with values replaced by TOPS data;
1-5	No merge; insert as new record; and
0	Record rejected, no merge; print on hardcopy.

   The appropriate action associated with the highest-scored record is performed.

Score and action assignments are governed by the rules listed in Table 2 below, which are based on the weights of the supplementary elements. In order for the system to find a possible match of the records when the value comparison of the primary element fails, the rules for checking the supplementary elements are set differently than for Case 1.

Table 2. Rules and Actions for Case 2

RULE	ACTION
If 4 elements match	Accept, merge.
If 3 elements match	
If the element not matched is within threshold;	Assign a score of 7-10; accept as TOPS GBL.
If the element not matched is not within threshold but is low weighted;	Assign a score of 2 or 4; no merge, insert new.
otherwise	Assign a score of 0; reject, no merge.
If 2 elements match	
If both of the elements not matched are low weighted and within threshold;	Assign a score of 6; accept as TOPS GBL.
If one of the elements not matched has the lowest weight and the other element is within threshold;	Assign a score of 1, 3, or 5; no merge, insert new.
otherwise	Assign a score of 0; reject, no merge.
If less than 2 elements match	Assign a score of 0; reject, no merge.

Detailed process flows for Case 2 are shown in Figure 3 on the following page.

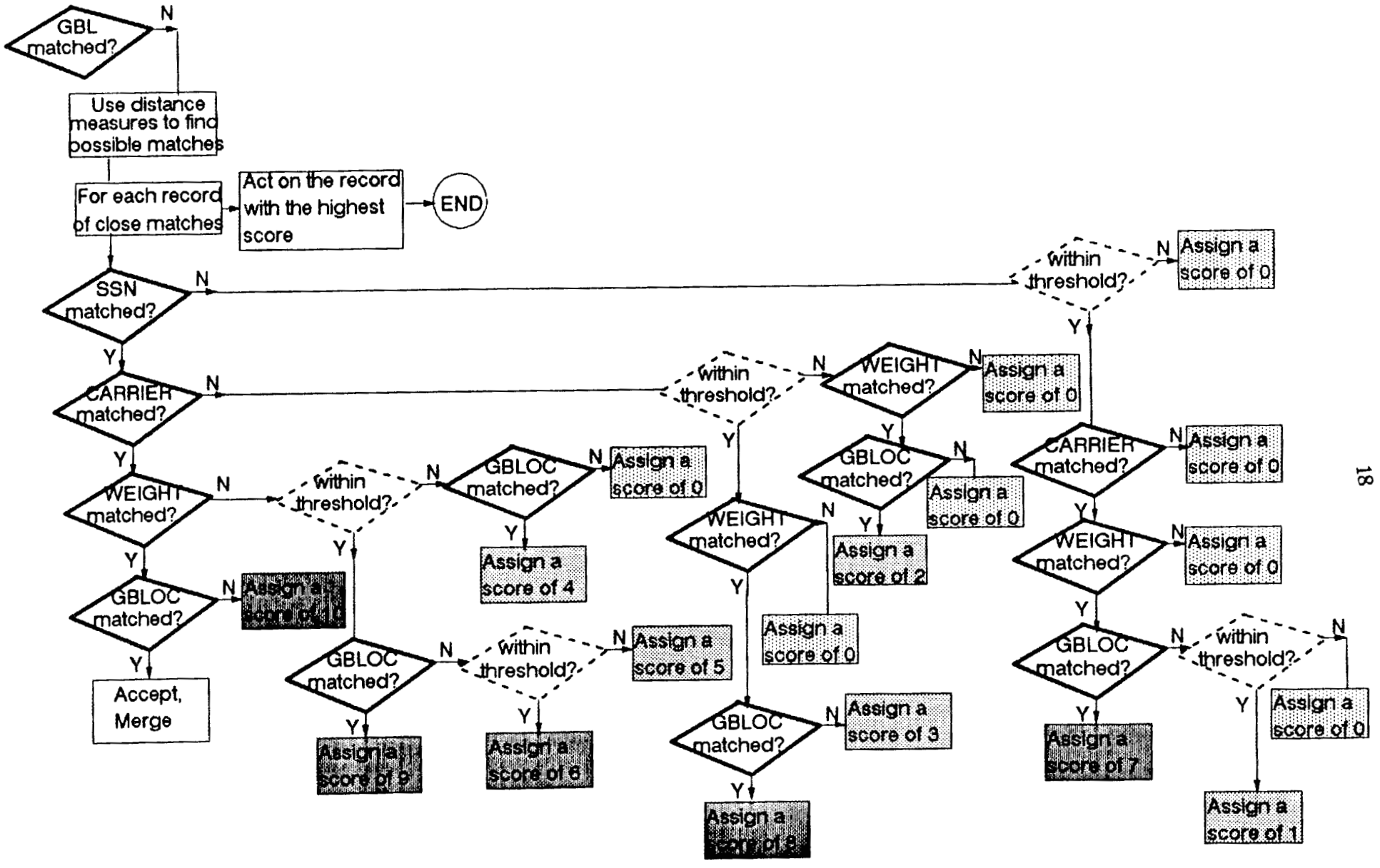


Figure 3 : Not Matched GBL Numbers

To illustrate how different weights of the elements affect scoring, the specific score assignments for Case 2 are summarized below in Table 3.

Table 3. Score Assignment for Case 2

# OF ELEMENTS MATCHED	ELEMENT(S) NOT MATCHED	ELEMENT(S) WITHIN THRESHOLD	SCORE ASSIGNED
3	GBLOC		10
3	WEIGHT	WEIGHT	9
3	CARRIER	CARRIER	8
3	SSN	SSN	7
2	WEIGHT, GBLOC	WEIGHT, GBLOC	6
2	WEIGHT, GBLOC	WEIGHT	5
3	WEIGHT		4
2	CARRIER, GBLOC	CARRIER	3
3	CARRIER		2
2	SSN, GBLOC	SSN, GBLOC	1

3.4.3 Implementation Method

WHIST-MOD is a distributed system involving heterogenous hardware and software. The basic hardware environment for WHIST-MOD includes:

- a Teradata DBC/1012 database computer (DBC) that serves as a historical data repository,
- a UNISYS 5000/80 minicomputer that serves as a data acquisition and data administration computer,

- a Solbourne workstation that serves as an application server within MTPP and as the host for the Teradata database computer, and
- some microcomputers;

all connected to each other via a local area network (LAN).

The WHIST-MOD database is hence distributed over several machines. Historical WHIST-MOD data like shipment data reside on the DBC. Data needed to assist in the user interface of the WHIST-MOD applications (for example, list of valid values) reside on the application server. The data administration machine, however, is required to store those data that have not been stored in any other machines in order to facilitate communications with the TOPS Switcher communications software.

Since only limited information is stored on the WHIST-MOD data administration machine, there is not enough information at the time of data integration to identify discrepancies. Access to the Teradata DBC/1012 database is hence required to implement the discrepancy check between TOPS records and Finance Centers/Claims Centers records. There are two implementation methods.

(1) Perform the discrepancy check on the DBC at the time of insert/update.

(2) Get the pertinent data from the DBC to the WHIST-MOD data administration machine, perform the discrepancy check on the data administration machine, and create the appropriate insert/update record ready for the DBC load.

Both methods require a utility program, preferably written in a portable third-generation programming language such as "C", that steps through the five data elements and performs the distance measure and weighted matching as detailed in Figures 2 and 3.

Since WHIST-MOD is not a real-time system, applying method 2 seems to be a more viable choice because this method will truly support the concept of a data acquisition/data administration dedicated machine.[8]

4. SUMMARY

The primary objective of data quality management during data integration is to eliminate incomplete data due to missing records or incorrect integration so as to support the decision-making function of a DSS. We found that our methodology of using distance measures with weights should be able to suggest with fair degree of certainty

(1) whether two members from different objects having distinct keys (due to errors) are indeed functionally duplicates, and

(2) whether two members from different objects having the same keys (due to errors) are actually functionally discrete.

Section 3 showed an actual application in which this methodology could be used for error detection and correction during data integration. The strategies proposed for the WHIST-MOD system were accepted by the program management staff of the WHIST-MOD system. However, because of budgetary constraints, this data integration concept was not implemented in WHIST-MOD as of the date of this paper. The implementation plan was to start with a skeletal framework for data integration. The error detection and correction strategies described in this paper will be incorporated into the system in a later version.

REFERENCES

1. V.W. Ng, *A Case Study in Specifying Data Requirements for a Decision Support System Database*, ORNL/TM-11583, Oak Ridge National Laboratory, September 1990.

2. Gunar E. Liepins, V.R.R. Uppuluri, editors, Data Quality Control (July 1990).

3. W.W. Peterson, E.J. Weldon, Error-Correcting Codes, 2nd ed., MIT Press, Cambridge, Mass., 1972.

4. J.L. Peterson, "Computer Programs for Detecting and Correcting Spelling Errors", Comm. ACM, Volume 23, No. 12, 676-685 (December 1980).

5. C.N. Alberga, "String Similarity and Misspellings", Comm. ACM, Volume 10, No. 5, 302-313 (May 1967).

6. J.B. Leber, "A Fuzzy Approach to Data Repair", Database Programming & Design, Volume 3, No. 1, 26-33 (January 1990).

7. P.H. Sellers, "On the theory and computation of evolutionary distances", SIAM J. Applied Math., Volume 26, 787-793 (1974).

8. P.F. Daugherty, V.W. Ng, T.L. James, *Final Functional Description for the Worldwide Household Goods Information System for Transportation Modernization (WHIST-MOD)*, ORNL/M-906, Oak Ridge National Laboratory, September 1989.

9. Oak Ridge National Laboratory, "Requirements Analysis for the Worldwide Household Goods Information System for Transportation Modernization at Headquarters", February 1989.

INTERNAL DISTRIBUTION

1. P. R. Barnes
2. F. P. Baxter
3. V. Baxter
4. M. Bronzini
5. T. S. Darland
6. R. T. Goeltz
7. R. K. Gryder
8. C. W. Hagan
9. H. L. Hwang
10. T. L. James
11. J. O. Kolb
12. M. Kuliasha
13. R. Lee

14-18. V. Ng
19. R. Noe
20. P. Singley
21. M. Spears
22. T. Truett
23. V. Wheeler
24. T. J. Wilbanks
25. ORNL Patent Office
26. Central Research Office
27. Document Reference Section

28-30. Laboratory Records
31. Laboratory Records (RC)

EXTERNAL DISTRIBUTION

32. B. G. Buchanan, Department of Computer Science, University of Pittsburgh, 206 Mineral Industries Building, Pittsburgh, PA 15260

33. J. J. Cuttica, Vice President of Research and Development, Gas Research Institute, 8600 W. Bryn Mawr Avenue, Chicago, IL 60631

34. S. W. Jebo, Military Traffic Management Command Headquarters, PPM, 5611 Columbia Pike, Falls Church, VA 22041-5050

35. D. E. Morrison, 333 Oxford Road, East Lansing, MI 48823

36. A. Hirsch, President, Dynamac Corporation, The Dynamac Building, 11140 Rockville Pike, Rockville, MD 20852

37. M. Williams, Professor, Department of Economics, Northern Illinois University, DeKalb, IL 60115

38. Office of Assistant Manager for Energy Research and Development, DOE/ORO, P.O. Box 2001, Oak Ridge, TN 37831-8600

39-48. Office of Scientific and Technical Information, U.S. Department of Energy, P.O. Box 62, Oak Ridge, TN 37831