SYSTEM FOR MATCHING COMPANY DOCUMENTS

J. Finley Carpenter, Troy R. Bishop, Ginger S. Goudie, Bureau of Labor Statistics

INTRODUCTION

A particular application in the growing body of literature on matching techniques is frame refinement in sample surveys. Whenever several records which refer to the same sample unit need to be identified and linked, some technique for "exact" matching must be employed. Exact matching refers to the goal of linking records corresponding to the same unit which may not be clearly and uniquely identified on the record. Examples of frame refinement using exact matching are combining sales receipts by customer name and eliminating duplicate listings in a directory of business establishments. [1]

In the sampling of reporters for the export price indexes of the Bureau of Labor Statistics, Shippers Export Declarations corresponding to the same exporters need to be linked together by assigning a unique identifier--a six digit exporter code. The outcome of the matching process is uncertain because of the differences in spelling and abbreviations in each company name and address found on the Shipper's Export Declarations. For each sample, approximately 12,000 documents need to be linked by exporter codes. Linking of the documents is done to form the first stage sampling cluster of product categories within companies and to calculate the probabilities used in the two stages of sampling.

There are two components of quality in the matching process: false links and missed links. Both errors directly influence the quality of the following two stages of sampling through their effect on probabilities of selection. A false link between two companies amounts to a truncation error since a product category sampled from one of the companies would have zero chance of being collected if the selected address belonged to the wrong company.

The importance of missed combinations arises in part from the need for a second stage sample of product categories within companies to avoid overburdening a company with too many product categories. If a valid link were missed, the probability of selecting a certain product category within that exporter would be conditional on selecting each of the separate links. To calculate the conditional probabilities, a similar matching procedure would have to be performed after subselection of product categories within companies. The upshot is that there is no substitute for an accurate system of matching company records.

The Old Approach

The first approach to solving the problem utilized a computer sorting routine that produced two alphabetical lists: one by exporter name and another by address. The documents which correspond to the same exporter were determined by a manual review of these listings and were identified by document numbers which were keypunched A computer program then assigned a unique exporter code to all documents which correspond to the same exporter. Several interations of this process were necessary in order to produce a sampling frame in which no additional combinations could be found.

The matching procedure was a tedious process which took two people 20 to 25 days and cost $800 in computer time. The 12,000 export documents were combined into approximately 4000 identifiable exporting companies. After the exporters were sampled, a quality measurement study estimated that about 50 of the 4000 identified companies were falsely linked with documents from unrelated companies. Moreover, about 350 of the identified companies could have been linked with other companies. The false and missed links would yield a net reduction of about 300 companies or 8% of the size of the exporter sampling frame.

Besides the tedium and inaccuracies, the old system lacked structural components which could be independently tested and therefore had limited potential for improvement. The efficiency seemed to depend upon unmeasurable quantities such as the alertness of the individual and the degree of effort expended in searching for a particular match. It was also difficult to determine the status of the coding process at any particular time such as the number of records which had been assigned an exporter code.

Design Objectives

Since the criteria for deciding whether two documents pertain to the same exporter are often complex and require expert judgement, it was decided that the primary functions of the new system would be to search for and display the essential information on all documents with names similar to a given record and to assign codes based on the operator's final judgement. Consideration was first given to using a standard text editor available on the computing facility. The text editor allows the user to find and display all records containing a string of characters which identically match a string specified by the operator. However, it was estimated that the computer cost would increase nearly ten fold. Moreover, the text editor does not produce an adequate status report, information on the records could be accidentally altered and modular improvement capability would be limited. For these reasons, it was decided to design our own system for matching company documents.

In particular, the new system was designed to:

- o reduce the tedium associated with thumbing through long lists of names searching for matches. This objective was considered especially important since the system would be used by highly paid professionals.

- increase the quality of the system as measured by number of missed and false links.
- reduce the manpower and elapsed time required.
- trim computer cost.
- have the capability of testing modular improvements.
- maintain summaries of status and effort expended to find matching documents.
- provide protection against accidentally altering any information on the document records.
- code any company identified as a broker or agent according to a numbering system which would distinguish them from regular exporters.

General Overview of Frame Refinement System

Figure 1 shows the basic processes of the new system which was designed to meet the above objectives. The first step is to create three utility files and the reformatted version of the file which contains the export documents. The exporter coding program, designed and written by Gordon Sollars, allows the operator to specify a character string and to search for and display records with exporter names similar to the specified string. The operator then decides which records pertain to the same exporter and assigns a common exporter code to them by means of an update command. The program keeps track of the next available code and the next uncoded record, calculates the number of exporters coded and the number of uncoded records, and checks for valid syntax and sequence of commands. In the event of accidental program interruptions or system failure, two recovery programs are executed. The coding at this stage follows a loose set of criteria aimed primarily at grouping all records with similar names and addresses by assigning a common exporter code.

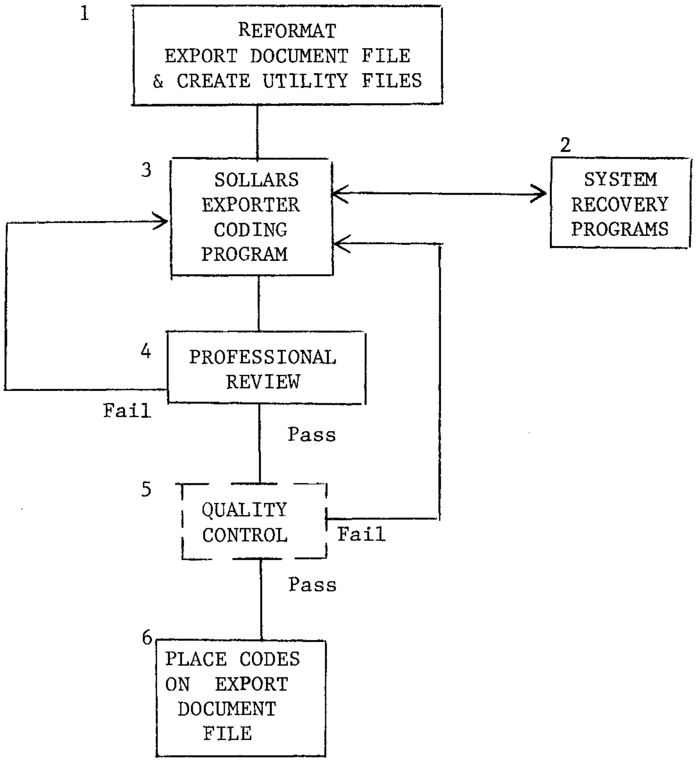
1
REFORMAT EXPORT DOCUMENT FILE & CREATE UTILITY FILES
2
SYSTEM RECOVERY PROGRAMS
3
SOLLARS EXPORTER CODING PROGRAM
4
PROFESSIONAL REVIEW
Fail
Pass
5
QUALITY CONTROL
Fail
Pass
6
PLACE CODES ON EXPORT DOCUMENT FILE

Figure 1 - Basic Processes of Frame Refinement System

When all records have been assigned exporter codes, reformatted lists (sorted by name, address and exporter code) are reviewed by professional economists who are well acquainted with industrial exporting companies. The economists check the listing for missed or false combinations of documents with an emphasis on catching the false combinations. The corrections are then made using Sollars' Exporter Coding program.

After the results of exporter coding pass the economists' review, a sample of documents will be rigorously inspected to ensure at a given confidence level that no more than a specified number of false and missed links were made. This quality control procedure is still in the planning stage. [2]

If the sample fails this quality control procedure, the exporter coding file goes back for review and corrections. When the sample passes, the assigned exporter codes are transferred to the proper records in the original export document file and the file is then ready to be sampled to select exporters and products to be used in the international prices program.

Operating Procedure

The primary function of the interactive exporter coding program is to assist the operator in assigning exporter identifying codes to each record in the Exporter Name File -- the reformated version of the file containing the export documents. It is helpful if the operator works from a listing of the file which is sorted alphabetically by name. Proceeding from the top of the list, the operator searches for the first uncoded record by entering the command IDSEARCH. The program then responds

```
PLEASE ENTER SEARCH KEY OR STOP
```

A null search key entered as "//" then finds and displays the first uncoded record.

The operator then enters a command (IDSEARCH or KWSEARCH) to search for names which are similar to some part or variation of the name (search key). Usually, the first two or three letters of a word in the name are used as the search key and will produce most of the candidates for a match. Initials and probable misspellings are also effective search keys. Either of two commands may be used to search for matching candidates -- each based on different similarity criteria. IDSEARCH finds all records with the first N letters of the name which match the N letters specified by the operator. The operator enters the function name IDSEARCH and responds to the keyword prompt with N letters. For example,

```
/ABC/
```

would locate all names beginning with "ABC".

The KWSEARCH command is used in the same manner. However, the program locates all records satisfying a more general similarity criteria. A score is calculated for each name based on the number of letters found that are in the same sequence as contained in the keyword. For example, if /GMC/ were used as the keyword, then "Grime Co" would have a score of three, the highest possible, since the name contains the letters "GMC" in the proper sequence. (Only records with perfect scores are located although the program could be modified to locate records with scores above any arbitrary cutoff.)

The records found by IDSEARCH and KWSEARCH are written to a current workspace and the range of new line numbers in the workspace is displayed after each search. The operator may then exhibit these lines using the DISPLAY command. For example,

```
DISPLAY 8/20
```

would exhibit lines 8 through 20 of the current workspace.

If one of the records found by the search commands has a previously assigned code, the operator may wish to write all other records with the same code into the current workspace by using the RETRIEVE command. The operator specifies the code to be retrieved and the program writes the records containing this code into the current workspace and displays the line numbers used.

The operator may then review the lines to decide which records pertain to the same exporter or exporting agent and assign codes using the UPDATE command. There are three options within this command:

1) ADD is used to assign codes to records in the workspace which have not been previously coded. For example,

   ```
   ADD 1/5 EX
   ```

   would assign the next available exporter code to the records in lines 1 through 5 of the workspace. If line 6 was recognized as an agent, the next available agent code is assigned using

   ```
   ADD 6 AG
   ```

   (An agent code is any 6 digit number over 10,000). Any specific exporter or agent code may be assigned instead of the next available code, e.g.,

   ```
   ADD 10/15 EX 586
   ```

   would assign exporter code 586 to lines 10 through 15 of the current workspace.

2) REPLACE is used to change the exporter or agent codes on records which have been previously coded. REPLACE operates in a manner similar to ADD to assign the next available exporter or agent code or any specified code.

3) DELETE is used to delete previously assigned codes by simply specifying the appropriate workspace line numbers.

When the operator is finished assigning codes to records in the workspace, the workspace may be erased using the CLEAR command and the above procedure may then be repeated starting with the search for the next uncoded record.

Status of the process may be instantly checked using the COUNT command which gives the number of exporters coded and the number of uncoded records. The operator also can specify either a long or short form for system prompts by entering the LONG or SHORT command.

Evaluation and Future Plans

The frame refinement system is complete -- except for the quality control function -- and has been used to assign exporter codes to two samples of documents, each covering a different broad class of commodities. The interactive exporter coding program operates smoothly and the operators report a noticeable lessening of tedium compared to the old system.

Although the evaluation of quality improvement is not complete, the quality of exporter coding seems to be an improvement over the old system as measured by the number of false and missed combinations: 3 false combinations and 16 missed links were found in a sample of 1250 of the identified exporters (Results from the second sample are not yet available). Extrapolating to a frame size of 4000, an estimated 13 false combinations and 131 missed links would have been made. (Compared with 50 false and 350 missed in the old system). These estimates are accurate to within 1% for false links and 50% for missed links at the 95% confidence level but are not strictly comparable since different sets of documents were involved. Plans are being made for more valid comparison by an experimental design which would also allow measurement of another dimension of quality: repeatablility. Analysis of the results may reveal classes of names for which the old method is superior to the new or the analysis may also suggest criteria for designing a pattern recognition search module for locating documents with the highest probability of match and lowest probability of mismatch. The criteria may then be adjusted to determine an optimum search routine which may be sufficient for a fully automatic exporter coding procedure as a first step in a modified system. These possibilities represent a fulfillment of the design objective of a system capability for modular improvements.

The objectives of not increasing manpower and computer cost have been met: both have been reduced 20%. However, these reductions do not in themselves justify the development cost of the system since the payback time would be 5-10 years.

REFERENCES

[1] Memorandum from Daniel B. Radner, Chairperson, Subcommittee on Matching Techniques, to Maria E. Gonzalez, Chairperson, Federal Committee on Statistical Methodology, May 5, 1978 (includes partially annotated bibliography).

[2] CARPENTER, J. FINLEY, "Error Analysis in the International Prices Program", presented at 32nd Annual Technical Conference, American Society for Quality Control, May, 1978.