Chapter 11

Record Linkage and Genealogical Files

Nancy P. NeSmith
The Church of Jesus Christ of Latter-Day Saints

Whenever there are large computerized genealogical files the problem of duplication of records for the same individual or family within the file always exist. Many indexing schemes can be used which allow some matching of entries being added to the files with slight variations, but computer technology in the past has been limited in matching entries in which more than one field such as surname, given names, date or locality have disagreement.

Genealogists know that disagreement comes through different record sources used for identification or through transcription errors. Bringing together records with discrepancies has always been a genealogical nightmare. If the records don't even come together using various sorting schemes, how can the records be analyzed for matching or merging decisions? In other words, how does one know if two different records refer to the same individual or family?

One solution to this dilemma is the use of Record Linkage theory. Record linkage refers to a computer program which uses a detailed algorithm based on probability to determine if two records being compared represent the same individual. This technique, developed in Canada by Howard B. Newcombe (1988), has been used in the statistical, demographic, and medical disciplines to identify and link two or more records representing the same person or entity.

The theory underlying record linkage was developed around the need for an algorithm which would mimic human decision making in comparing a record from one file with a record from a second file. To do this, two records which represent the same person are studied and field comparisons made. Fields are items of information in the record, such as given name, surname, birthdate, birthplace, etc. The outcome of this comparison (agreement, disagreement or partial agreement) that is common in linked records is noted. If there is enough agreement, the probability is high the records being compared from the two different files represent the same individual. If the comparison outcome is more common to unlinked records, the probability is high the records being compared represent two different individuals.

Using the comparison statistics, a record linkage system computes the odds in favor of a match or against a match for any two records selected for comparison. For example, if a surname in two records matches, the computer calculates the odds of the two names matching by chance and how often the surname field agrees in the truly linked records contained in the comparison file. From these two statistics the program determines a score which represents the odds above chance the two surnames matching are for the same person.

Each algorithm may be tailored to the uniqueness of the genealogical data elements in its geographic area. This eliminates applying an "English" standard to all geographic areas of the world. Another advantage is the algorithm may be refined to specific cultural or variable record types.

To develop the algorithm, samples of files which need to be matched or merged are examined by

Reprinted with permission from the *Utah Genealogical Journal* (1992), 20, 3 and 4, 113-119.

specialists to locate duplicates for statistical analysis. Based on their analysis and the purpose of the linkage for the file, the specialists choose blocking, weighting, and threshold parameters which will be used by the computer for each geographic area to determine if the records being compared are a match.

Searching the File -- Blocking

When searching a file to see if there is a record which matches the request, it would be ideal to compare every record in the file with the request. However, this is not practical in most data bases so an indexing scheme is used to retrieve only the entries in the file which are most likely to match the request. This is called blocking or retrieval. The intent is to reduce the number of comparisons the computer must make. The implicit assumption is that only records with a reasonable chance of being linked are retrieved.

Blocking effectiveness can be described in terms of "recall" and "precision." Recall is a measure of how many relevant records in the file are included by the blocking scheme. Precision is a measure of how many of the total records retrieved by the blocking scheme are relevant.

For example, if you're looking for a record of Joseph Jones, and the file contained two records for him, one in which he is identified as Joseph Jones and the other as J. Jones, the system would exhibit good recall if it retrieved both records. However, the system tuned to recall near matches such as J. Jones may retrieve irrelevant entries where the letter J. stood for John or James rather than Joseph. These irrelevant entries are known as "noise."

Precision measures the amount of noise. A problem with a system tuned for precision over recall is relevant entries can be missed because the narrower search parameters used to limit the noise also limit the recall. Whenever you tune for recall you increase the noise; when you tune for precision you decrease recall. The two concepts have been found to be in opposition. The goal is to find an acceptable balance between the two which suits each specific application.

It is possible to enhance the recall of a system without greatly reducing the precision by using some form of authority control to bring together the equivalent names of people which are spelled differently and locality names which are different but refer to the same locality. Blocking schemes are tuned to the specific file or part of the file being searched. Those fields which are accurate, discriminating, and most often present in the records are chosen because they help give a balance between precision and recall.

Weight Calculations

Using the blocking parameters the computer retrieves a set of records which can now be compared in detail with the query to determine their similarity to it. As fields in the query and candidate record are compared, a statistical score or weight is computed which reflects agreement, partial agreement, or disagreement of the two fields being compared. A positive weight is calculated for agreement, a smaller positive weight is calculated for partial agreement, and a negative weight is calculated for non-agreement on that field. If either record has missing information in the field being compared, a weight of zero is assigned. The weights are added to each other to obtain a total weight which reflects the similarity of the pair of records being compared.

The weights are tailored to the locality or record source. For example, surnames for England agree more often than surnames in Denmark and other countries which have patronymic surnames. This affects the weights. England would have a smaller positive weight for agreement on surname than Denmark but would have a higher negative weight for disagreement on surname than Denmark because the surnames seldom disagrees for England. Also taken into consideration for the calculation of the weight is the relative size of the name pool. For example, there are fewer surnames in England than there are in United States records. The fewer the names, the less significant is the agreement. These types of calculations and comparisons are done on the fields for gender, names, localities, and dates.

It is not necessary to weight all of the fields in a record. Generally fields which were used as blocking parameters are not weighted. The fields are not weighted because the records retrieved have already matched on these fields and weighting them only increases the overall score of each record by the same amount. Other fields may not be weighted because they are not statistically discriminating and don't contribute significantly to the equation.

Threshold Determination

Once the file records have been retrieved using the blocking scheme, compared field by field to the query, each field weighted, and each total record's weight determined; then a decision can be made about whether or not a duplicate was found. The total weight which is used to decide whether a record should be considered a match or non-match with the retrievals from the file is called the threshold.

Generally, scores above a certain threshold indicate a match and those below it indicate a non-match. For example, if the weight of 40 is considered the threshold, then all retrievals scoring less than 40 are considered non-matches. All retrievals with scores of 40 and above are considered matches. The threshold decision is based on the total weights of truly matched records for that specific locality (truly matched means the two records are known to refer to the same person).

There is often a small range of scores which includes intermingled matches and non-matches. This is called the gray area. For example in a study of criminals and law abiding citizens, the range of scores could be from -150 to 150. Criminals have scores ranging from -150 to +50. Citizens had scores ranging from +30 to +150. Everyone with a score below 30 is a criminal, everyone with a score above 50 is a citizen. Those with a score between 30 and 50 could be either a citizen or a criminal. This is the gray area for this study, meaning if 40 is picked as the threshold score there is the possibility of a non-match scoring high enough to appear to be a match (false positive) or there is a possibility of a match scoring so low it appears to be a non-match (false negative). Which number to pick is called the threshold decision.

In making a threshold decision it is important to decide the purpose of the links. For optimal linkage it is important to follow the rule which states that before a threshold is picked, decide the purpose of the links. If the goal is to arrest all the criminals in a town and not let any of them go free, then a threshold of 50 would be picked. But as a result, some law abiding citizens would be arrested because their score would be similar to criminals. If the goal is to arrest as many criminals in a town but to not falsely arrest any citizens, then a threshold of 30 would be picked. As a result, some criminals would go free, but no citizens would be arrested.

The Family History Department and Record Linkage

The theory underlying this technology is the best approach known to the scientific community. For this reason, the Family History Department of The Church of Jesus Christ of Latter-day Saints has chosen to implement its usage in their genealogical systems and databases. It is currently being used to retrieve entries within the department's genealogical files, and will be used in match and merge decisions. The results have been very satisfying and its efficiency has been improved by taking full advantage of name and locality authority systems. The use of record linkage for the massive files and record linking needs of Family History makes the most efficient use of the Department's computer resources in eliminating or matching duplicates in their files. This technology has not been employed in the Personal Ancestral File7 as that program was developed before the implementation of Record Linkage in 1988.

References

Newcombe, Howard (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration and Business*, Oxford: Oxford University Press.

Wrigley, E. A. (ed). (1973). *Identifying People in the Past*. London: Edward Arnold.

* Nancy P. NeSmith, 5440 South Lighthouse Road, Salt Lake City, Utah 84123. Miss NeSmith received a BS in Genealogy and undertook graduate studies in Family History at Brigham Young University. She is currently a Systems User Specialist in the LDS Family History Department.

Personal Ancestral File is a registered trademark of The Church of Jesus Christ of Latter-day Saints.