Appearing in the proceedings of the 10th International Conference on Computer Applications in Industry and Engineering
San Antonio 1997

Creating and Using Automated Profile Matching Technology

Ian Davidson and Joshua Huang
CSIRO - Mathematical and Information Sciences
723 Swanston Street, Carlton, Victoria, Australia 3053
email: {ian.davidson | zhexue.huang}@cmis.csiro.au

Abstract: Many financial organisations (i.e. banks and insurance companies) have databases on individuals and their major transactions. These databases have different uses such as obtaining a historical record of an individual. Often the (legal) circumstances mean that a unique identifier for each individual does not exist or cannot be used. Instead personal identification details (names, addresses, dates of birth etc.) must be used to "index" and retrieve records. This is an inexact method and often results in ambiguous situations where multiple records are returned for an enquiry. Disambiguating these situations we term the profile resolution problem. Organisations usually use human experts (review clerks) to resolve ambiguous cases. This is time consuming, expensive and does not address the cause of the problem. We have shown that it is possible to automate the decision making expertise of these review clerks with a number of techniques from artificial intelligence and statistics. We illustrate that the profile resolution problem is actually one of three inter-related problems associated with ambiguous cases and discuss how automated techniques can address each.

Keywords: record linking and consolidation, data mining, induction, financial systems

Introduction and Motivation

Many financial service organisations such as banks, insurance firms and credit unions have databases of individuals' personal identification details and their major transactions such as loan/credit applications and insurance claims. We call these *profile databases*. Profile databases can be used to recall the transactional history of an individual, generate summary statistics for all individuals and find interesting patterns to help in marketing, operational issues and even money laundering [1][2]. The database is updated whenever an individual performs a major transaction, usually by filling in a paper form which is then entered by a data entry clerk into the database. If the individual already exists in the database, then the transaction is appended to their existing record, otherwise a new record is created.

There are often privacy or legal limitations to using a unique identifier (key) for each individual even though they exist. In these situations, personal identification details such as names, dates of birth, license numbers, current and previous addresses have to be used. Identification information is usually sufficient to uniquely identify an individual, with the exception being extremely rare situations such as those involving co-habitating twins. Advanced techniques such as cross-validation (of names and addresses), phonetic matching (of names, addresses and suburbs), expanding abbreviations and nicknames are used to generate an advanced retrieval technique. Such techniques, often with incomplete information in the enquiry can retrieve the correct record for an individual in most cases. However, many situations can occur which lead to several records being retrieved for an enquiry. These situations are said to be *ambiguous*, because it is unsure which record(s) represent the individual. Ambiguous cases arise from a large variety of reasons. Some are evidently due to data problems (incorrectly spelt names or absent fields). Others are due to temporal changes such as an individual changing their name or address. Often the customer misinterprets the information required when filling in a form and provides incorrect information.

We postulate that ambiguous cases in profile databases are associated with three problems, all of which are interrelated to some degree. The three problems are illustrated in figure 1 which shows the process of obtaining a transaction history for an individual. The translation problem involves accurately getting identification details from an individual into the profile database. This goes beyond a simple data entry accuracy problem and into issues such as form design and maintaining standards on fields. The integrity problem arises because the database may contain multiple images of the same person or the identification details may not be sufficient to uniquely identify a person. The profile resolution problem involves determining which record(s) of the closest matches to any enquiry is the person in the enquiry. These problems cause numerous unwanted situations with respect to time and cost. The transactional history returned may represent the wrong person, may only be a partial history or may represent many people. Thus the statistics and information derived from the data is limited and potentially incorrect. There may also be legal ramifications if derogatory decisions are made on information that is untrue.

Currently a common solution in complex situations is to employ a collection of case review clerks to resolve ambiguous cases. When an enquiry is raised which leads to an ambiguity, the enquiry and the most likely records to be the individual in the enquiry are sent to a clerk for review. The clerk must then examine the enquiry and piece together the records to provide a complete picture of the individual. This may be as simple as choosing a single record. The ambiguity could have been raised because of the commonality of the surname or because multiple individuals staying at a common address are related and share similar first and middle names. At the other extreme overcoming the

ambiguity may involve combining several records to form a complete picture of the individual. This may be because the individual has been married and changed his/her surname or address.

A manual solution only solves the profile resolution problem, is costly and introduces a time delay. We show that an automated profile matching technique can solve to varying degrees all problems. In the following sections we discuss each of the three problems. We then outline our trials of different technologies from artificial intelligence and statistics to replicate the decision making expertise of the review clerks. We have shown that we can replicate the decision making expertise of the review clerk in approximately 80% of ambiguous cases. We discuss the differences in performance for each technique and the apparent strategy they used. In the next section, the three proposed applications, each of which solves a specific problem, are described. We conclude this discourse by outlining the findings of our current commercial study.

The Three Types of Problems

We postulate that there are three problems associated with the need for profile resolution. In this section we detail each problem and discuss their implications.

Translation Problem

Entering information about an individual is a two stage process. The customer must first fill in a paper document which is then transcribed into an electronic form. Errors can occur at both stages. When filling in a form, an individual may give incorrect or incomplete information or even misinterpret the intention of collecting the field (eg. specifying a relative's licence number when they don't drive). More common errors involve not providing information such as dates of birth or specifying initials instead of full names. When transcribing these details into an electronic form the operator may perpetuate these errors or introduce new errors. It should be noted that the data entry step is often completed quickly. A new record is created if the enquiry does not sufficiently match any existing individuals in the database.

Integrity Problem

If there is no unique identifier for an individual, identification details must be used as a composite "key" to index the profile database. As there is no guarantee on the quality of this information, partial, incorrect and inconsistent identification details are stored in the database. This results in the situation where an individual's history may be split into multiple records or there are records with extremely similar identification fields. The integrity of the database may be compromised in two ways. There does not exist one record per individual and it may not be possible to uniquely identify each record.

Profile Resolution Problem

The final problem, profile resolution, occurs because some queries on the database may result in multiple records being returned. These records may be relatives, spouses or children of the individual in the enquiry, multiple records of the individual or totally different people. From these multiple records, the record or combination of records that represent the individual needs to be identified.

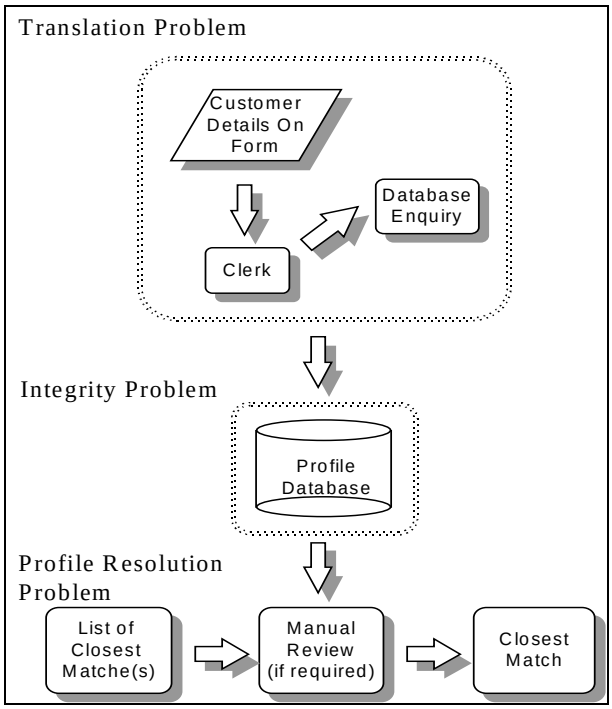

Figure 1: Three problems that automated profile matching technology needs to address.

Implication of Problems

So long as no unique identifier (such as a tax file number or social security number) can be used for each individual these problems will occur to some extent. The causes of these problems will depend on the process used to collect the information. Though the problems are related, fixing one particular problem does not necessarily fix another. If the translation problem is overcome, individuals will still change their names and addresses. If the integrity problem is overcome individuals will still fill in forms incorrectly and clerks will make transcription errors. Solving the profile resolution problem will enable histories of individuals to be retrieved, but the profile database could not be used confidently for other purposes such as generating summary statistics on individuals.

Trials of Automated Techniques

We initially focused on the most accessible problem, profile resolution. We had available several thousand ambiguous cases which had been resolved by a clerical reviewer. Each case contained the enquiry (identification details such as name, date of birth, drivers licence etc.), the identification details of up to the four closest matches (candidates) to the enquiry, the reviewer's actions (a finite set of commands), which identification fields (i.e. surname, date of birth, licence number) were used in the decision and in unusual cases, reasons for the decision. The identification details for individuals is clearly alpha-numeric. To enable our automated techniques to process the alpha-numeric data the identification details are presented to the technique as a series of pairwise (between the enquiry and each candidate) scores for each identification field. The scores represent the similarity between fields and are calculated in a field specific manner making use of prior knowledge. In some situations the scores are "fuzzified". The reviewer's action are commands to the database such as select a record or combine records. It is estimated that the reviewer makes the correct decision 85% of the time. Those decisions which are incorrect are most likely due to the reviewer being conservative, a desriable trait for an automated technique. In our future studies each case will be reviewed twice for additional accuracy. Any automated technique must be able to replicate the decision making expertise of the clerical review. This involves determining the action to take for a given case. We can compare the accuracies of each technique by calculating the proportion of cases where the decision made by the automated technique is the same as the clerical reviewer's. We have tried four different techniques:

1) A machine learning approach using an information theoretic approach to induce decision trees and rules;
2) A conjugate gradient neural network;
3) A case based reasoning approach; and
4) A approach from the field of statistical record linking

The first three techniques are termed *supervised* as they make active use of the decisions made by the review clerk during the learning phase. A decision model is induced from a set of examples called the *training set* and then tried upon the *test set*. The two sets do not overlap. Our trials involved producing multiple training and test sets by randomly sampling from the available cases. The proportion of cases in the training set varied between forty to sixty percent with the remainder of cases in the test set. Our results indicate the average performance of each technique over ten training/test set combinations.

The machine learning approach uses an maximum information gain approach to find the best decision tree for the training set. A penalty function is used to penalise overly complex trees from overfitting the training set. After the decision tree was induced it was parsed to generate an approximately equivalent rule set. The technique is loosely based on the ID3 algorithm [3]. The neural network trials were conducted with the Darwin tool whilst the case based reasoning approach used an inhouse system. The final technique, statistical record linkage [4] [5], does not make use of the decision information during the training phase and is hence termed *unsupervised*. It was tried on all the available cases. Trials were conducted using the AUTOMATCH software product [6].

The techniques varied in the format of the information they processed, the decision model representation plus operational and implementation issues. The results we achieved are illustrated in Table 1. The cases we had available were divided into three categories which corresponded to the number of records (candidate cases) that clerical reviewers had available to them. The break up of cases is shown in Table 2. For each category of case we developed a decision model and evaluated its *accuracy* by determining the proportion of decisions made on cases in the test set (or all cases in the unsupervised technique) that were the same decision made by the human expert.

	Technique			
Category of Cases	Machine Learning	Neural Network	Cased Based Reasoning	Statistical Record Linkage
All Cases	84%	76%	50%	77%
2 Candidates	88%	76%	50%	81%
3 Candidates	70%	Not done	Not done	66%
4 Candidates	67%	Not done	Not done	60%

Table 1: Accuracy of Predictions

Category	Breakup of Examples (%)
Cases with 0 candidates	0
Cases with 1 candidate	2
Cases with 2 candidates	72
Cases with 3 candidates	19
Cases with 4 candidates	7

Table 2: Breakdown of 2098 example cases provided.

Discussion on Techniques

We feel that of all the techniques, only case based reasoning (CBR) is not suitable. This is not to say that a CBR approach is not suitable only that our variation is not. This was primarily due to the strategy the technique adopted in attempting to determine the correct action to take. Depending on the number of candidates available the

number of actions varied between four and twenty four. The case based reasoning approach tried to *select* the action from all those that were available. We believe this is the reason for its poor performance. We cannot comment on the strategy that the neural network approach used, but by examining the decision tree we induced and the statistical record linkage approach it was apparent they used an alternative strategy. This strategy inherently involved *removing* those actions which were obviously incorrect until the correct action to take was obvious. The strategy was applicable because the retrieval mechanism (given sufficient details of an individual) was felt to always return the images of that individual that existed in the database. Therefore one of the available actions was *always* applicable. Determining if any of the actions were not applicable did not need to be decided. With sufficient effort and encoding of prior knowledge of the decision task, the accuracies of any of these three techniques could be improved.

It was interesting to note that there were intrinsically difficult cases for which several techniques made incorrect predictions. On further examination of these cases, we found they fell into one of two categories. The first was due to the decision being made on information beyond that given to the automated profile matching technique. Each profile contains additional information beyond identification details such as company directorships. This rare information could be used to determine if records should be consolidated. The second is because the decision was made on information which was apparently derived by the clerical reviewers from the raw information. We believe the cases in the first category, due to their rarity and difficulty in solving are not worth the effort in automating. Rather, identifying the conditions of these intrinsically difficult cases will be explored. Cases which match these conditions will then be passed onto a clerical reviewer. Cases in the second category were more common and we attempted to replicate the derivation of information from raw information. We explored the notions of implied sex (derived from first and middle names) and surname complexity which then became additional fields that the automated techniques had available for each case. Using these derived fields was shown to improve the accuracy of our results.

Applications

We have demonstrated that three of the four techniques can be used to automate the decision-making process of clerical reviewers. In this section we focus on how these techniques can be used to solve each of the three problems identified in figure 1, and where appropriate, the most suitable technique to use.

The three possible applications for using the techniques, each of which addresses one of the problems, are:

1) An add on to the on-line profile resolution system: profile resolution problem
2) An autonomous database utility: integrity problem
3) Isolating problem fields: translation problem

On-line Profile Resolution System

Most organisations have an enquiry front end to their database system which can handle the majority of cases. Only ambiguous cases are sent for clerical review. In this application the automated profile matching technology partially replaces the human experts. Ambiguous cases are sent to the automated resolution filter. The filter makes a decision on the correct action to take, which is then performed. The newly formed record is then returned to the user making the enquiry. The entire process would take place in real time. Each decision on a case has a confidence level associated with it which is loosely a function of the past experience on similar cases. The decisions with a relatively low confidence level or cases which are known to be complex would be referred to a clerical reviewer for a decision. It is feasible (though we have not explored this idea) to introduce a feedback loop. The human resolution of these difficult cases could be compared to the automated resolution's decision and changes made to the automated decision making technique if required. This application requires the embedding of the decision model into the on-line software. Whilst this is achievable by any of the techniques, implementing a series of rules would be the most easiest.

Autonomous Database Utility

The use of the automated technology as an real-time add on to a profile resolution system is essentially re-active: ambiguous profiles are resolved only when a user issues a request for one of them. This means that there still may exist multiple records for an individual in the database. The existence of these duplicates means that most information derived from the database is most likely to be incorrect. Average debt per individual for example, would be understated. A pro-active use of the automated profile matching technology would be to directly apply it to potentially ambiguous records in the database. This introduces the question of how to determine potentially ambiguous records. By looking at the actions available to a reviewer, ambiguity arises only when there are a number of very similar files.

Records by themselves are not ambiguous. It is required to find a group of records which together have the potential to be considered ambiguous. Finding groups or clusters of similar records is a well-researched problem [7] [8]. Cases

can be clustered using a number of different technologies ranging in complexity from database queries to advanced clustering software.

After the completion of the clustering process each cluster represents cases which together could be considered to be ambiguous. These ambiguities need to be resolved. Each record in a cluster or an exemplar representing a cluster could be used to represent an enquiry. This artificial enquiry would then be processed and the automated resolution technique used if applicable.

Isolating Problem Fields

When an individual begins filling in a form, he or she possesses all the information required to uniquely identify themselves apart from exceptional cases. However, when converting this information into electronic form numerous errors occur which lead to ambiguities being introduced. These may occur in a multitude of forms. The customer may introduce errors when filling in forms. Data entry clerks when transcribing their details may perpetuate these or introduce new errors. Some of these errors are inevitable and are only errors from a continual identification viewpoint (someone changing their surname due to marriage is hardly an error). These type of errors, we assume, will occur uniformly across all localities. However, other errors may be due to reasons which are due to the process of how the data is collected and transcribed. A profile matching technique such as rule induction, where the decision model can be interpreted can provide insight into identifying the last type of errors.

The automated profile matching technology contains a decision model on how to resolve ambiguities in particular cases. The decision model compares names, addresses and other information for potential matches to the enquiry which raise the ambiguity. By determining how parts (fields such as date of birth) of the decision model are used, we can determine why the ambiguity occurred. By then looking at details regarding the location of the origins of the enquiries, we can identify a scope of application. If this scope is distinctly non-uniform it is possible to identify a region in which the error is most likely to occur. For example, if a rule which focuses on missing middle names is constantly activated by queries coming from a particular suburb in Sydney, then it would seem likely that the method of capturing this information is flawed in this region. The flaw may be due to problematic paper forms or instructions given to the customers.

Other Applications

We have only discussed one use of the automated profile resolution technology. The technology inherently allows us to determine if two profiles are linked. The technology could also be used to merge together physically different profile databases (one say for insurance claims and another for credit history) to obtain a more complete picture of an individual. Furthermore the removal of all potentially ambiguous records allows the generation of various summary information and interesting anomalies with great confidence in the information derived as has been advocated before [2].

The automated profile resolution technology based on rule induction allows us to determine the nature of the relationship between the profiles. The rules induced could be labeled as effectively resolving a particular situation. Examples of these situations are when an enquiry returned individuals from the same family or the same person at different stages in their lives. The rule(s) to resolve these situations can also be used to *identify* them. Enumerating each possible situation, identifying rule(s) with each situation and using these rules to identify the situations could be beneficial to determine more complex patterns of a transactional history of units larger than an individual. This is useful for identifying fraudulent behaviour [2].

Conclusion

Organisations sometimes must use personal identification details to index and retrieve records on individuals. This can lead to ambiguous situations where an enquiry returns multiple records. We have identified three problems associated with these ambiguous situations. To address one of these problems, profile resolution, organisations use review clerks. We have shown that automated profile matching techniques can be built by analysing their actions in the form of a set of worked examples. We found that using derived information from the raw data and adopting a strategy of removing obviously inapplicable actions worked well. Three of the techniques we tried handled approximately 80% of cases correctly. We illustrate how automated techniques could be used to address all three problems.

Acknowledgments

We would like to thank the other group members, James Gleeson, Geoff Laslett and Dong-Mei Zhang for their considerable effort on this project.

References

[1] T.E. Senator et al, “The FinCEN Artificial Intelligence System: Identifying Potential Money Laundering from

Reports of Large Cash Transactions," Proc. of the 7th Annual Conference. IAAI, 1995.

[2] H.G. Goldberg, T.E. Senator, "Restructuring databases for knowledge discovery by consolidation and link formation," KDD 95, pp 136 - 141, 1995

[3] P.H. Winston, Artificial Intelligence, Third Edition Addison-Wesley, 1992.

[4] H.B. Newcombe, "Record linkage: the design of efficient systems for linking records into individual and family histories," American Journal of Human Genetics, 19, pp. 335–359, 1967.

[5] H.B. Newcombe, Handbook of Record Linkage. Oxford: Oxford University Press, 1988.

[6] M.A. Jaro, "Probabilistic linkage of large public health data files," Statistics in medicine, 14, pp. 491–498, 1995.

[7] G. Dunn, B.S. Everitt, An Introduction to Mathematical Taxonomy, Cambridge University Press. 1982

[8] T. Clifford, W. Stephenson, An introduction to numerical classification, Academic Press. 1975