ELSEVIER

Decision Support Systems 34 (2002) 369–378

Decision Support Systems

www.elsevier.com/locate/dsw

Data association methods with applications to law enforcement

Donald E. Brown [a,*], Stephen Hagen [b]

[a] Department of Systems and Information Engineering, University of Virginia, P.O. Box 400747, Charlottesville, VA 22904-4747, USA
[b] Complex Systems Research Center, University of New Hampshire, Durham, NH, USA

Accepted 28 February 2002

Abstract

Associating records in a large database that are related but not exact matches has importance in a variety of applications. In law enforcement, this task enables crime analysts to associate incidents possibly resulting from the same individual or group of individuals. In practice, most crime analysts perform this task manually by searching through incident reports looking for similarities. This paper describes automated approaches to data association. We report tests showing that our data association methods significantly reduced the time required by manual methods with accuracy comparable to experienced crime analysts. In comparison to analysis using the structured query language (SQL), our methods were both faster and more accurate.

Keywords: Data association; Crime analysis; Link analysis; Information retrieval

1. Introduction

Information retrieval typically depends on exact matches between the search criteria and the fields searched. In most applications with data stored in relational databases either the user or some intermediate software converts the search criteria into appropriate formal statements in the structured query language (SQL). These SQL statements then return records or documents that exactly match the criteria specified.

Many applications require less than exact matches of the search criteria. Law enforcement is one of these applications. Crime analysts work with incident reports that contain all the relevant information about a criminal event as recorded by a police officer. While the form of the incident reports vary by jurisdiction, the FBI encourages law enforcement agencies to use a standard set of report fields and entries within those fields. This standard is called the National Incident Based Reporting System (NIBRS). Examples of NIBRS fields include location of the incident (bank/savings and loan, convenience store, etc.), type of victim (individual, business, etc.), and hair color of the offender. NIBRS provides a standard way of reporting incident and arrest data for 22 offense categories that consist of 46 crime types.

After the police have entered the incident or arrest data into these reports, crime analysts search these reports for indications of multiple incidents committed by the same person or group of persons (e.g., a gang). Once they have this associated set of incidents, they can look for patterns in the way the criminal or criminal gang operate and use these patterns to aid in apprehension.

* Corresponding author.

0167-9236/02/$ - see front matter
PII: S0167-9236(02)00064-7

In order to find the associated set of incidents the analysts use inexact matches of search criteria. For example, the suspect in one incident can be reported as "heavy" and in another incident as "stocky" or even of "medium build." We would not want to eliminate suspects simply because they did not exactly match the descriptions given in one or more reports.

While most major police departments have computerized records management systems (RMS), these systems do not support the creation of the associated set of incidents given inexact search criteria. This task of grouping data records according to similarity measures is called data association. In the next section, we describe some research on prototype data association systems to aid police analysis. However, none of these systems completely handled the data association tasks. Nor were they fully tested or developed. As a result, crime analysts without access to these prototypes must perform this function manually. Manual searching for associations is both tedious and time-consuming. To better understand the demands of manual searching, we studied three crime analysts in the Richmond, Virginia Police Department as they performed this task for a series of incidents. An accurate comparison of two cases with all information present takes over 1 min for an experienced analyst. Performing pairwise comparisons on just 500 cases (less than half the average number of robberies in Richmond in a year) would require more than 1 million hours. Even the relatively simple task of comparing one new case to 500 existing incidents requires a full day's work for the analyst. Clearly, crime analysts can only perform this task for the most important crimes and hope that perpetrators of other crimes will reveal themselves in other ways.

This paper describes automated approaches to data association motivated by the need to aid crime analysts. However, the methods described have applicability beyond crime analysis to the general problem of associating records in databases using inexact search criteria. In the next section, we provide an overview to existing tools for crime analysis and related literature. Section 3 gives the details of our methods and Section 4 provides results from testing these methods using real data from Richmond. The last section contains our conclusions and recommendations for continued work.

2. Related research

Over the last two decades researchers have developed a number of automated tools to support law enforcement activities. These tools assist in a wide variety of tasks from identifying potential suspects to assisting in hostage negotiations. For our work, we were interested in tools and methods to support crime analysis. We have organized examples of this work into three categories: expert systems, investigative support systems, and non-automated crime analysis methods. The remainder of this section describes work in these three areas.

2.1. Expert systems

Expert systems provide the most common approach to the data association problem for law enforcement. This subsection examines these expert systems approaches and compares them to our approach to the data association problem.

Badiru et al. [3] built the Armed Robbery Eidetic Suspect Typing (AREST) system to help simplify and enhance the investigative process and from our perspective to perform data association. The system searches for similar information contained in multiple reports and attempts to associate the reports.

AREST uses an expert systems approach to accomplish its tasks. Working with the Comanche County Sheriff's Department, the authors assembled a list of 13 suspect traits that could be identified by a witness. Each of these traits was assigned a confidence level that indicates "the level of certainty that a given suspect is probable for a crime for which his observed traits match a suspect file" [3]. The experts from the police department then established a set of rules based on these traits. Using this knowledge base, the expert system classifies people in a suspect file as probable suspects, possible suspects, or not a suspect. The expert system uses the inference engine Personal Consultant Plus. The paper only reports the results of a feasibility study with no testing and no follow-on studies have been reported.

While both AREST and our data association methods described in this paper attack similar problems the approaches taken have important differences. First, our data association methods can be used for both suspects and incidents (i.e., we can link incidents

based modus operandi not just suspect descriptions). Second, our data association methods rank incidents or suspects (i.e., they return a partial order of the alternatives) rather than classifying them into three classes: probable, possible, or not a suspect. Third, our data association methods have been tested on real data and compared to the performance of professional crime analysts. Fourth, details of the association methodology for AREST are hidden in the rules while in our approach they are accessible to the user and can be changed to support specific search requirements. Finally, our data association methods generalize to a number of different types of crimes and criminal incidents.

The FBI has developed a system to aid in crime analysis called VICAP [10]. VICAP compares over 100 modus operandi (MO) attributes of an incident with the other cases in a database. The system returns the top 10 "matches" that were most similar to the case being examined. As with AREST, VICAP uses expert systems technology. This system, unlike our approach, only handles incidents (i.e., modus operandi) and does not associate suspects. The approach returns only the top 10 matches instead of a possibly larger set of equally or closely similar cases. Finally, as with AREST, there are no published studies providing testing results for the approach or follow-on reports.

2.2. Investigative support systems

Other systems have been built to support law enforcement investigations but not specifically data association. Examples of these investigative support systems include APES [6] and the Baltimore County Burglary System [12].

Also of interest are systems that perform something called link analysis. For example, the commercial product Watson [13] looks for links between suspect attributes. For example, a suspect report could indicate that the suspect has an acquaintance with a first name of Ken who drives a red truck. Watson then searches for other records of suspects with the first name of Ken who own red trucks. If it finds one or more records then these are linked to the first record. Notice that this linkage indicates a different type of association than the one found by our techniques. In particular, Watson looks for exact matches of attributes that link different people (or objects) together but it does not associate incidents possibly committed by the same person through inexact or partial matches. Other commercial products in the same general category as Watson include: Memex, iGlass, ALTAanalytics, and SIUSS.

2.3. Non-automated crime analysis methods

Also of particular relevance to our work are results from non-automated crime analysis. Crime analysis has begun to look more carefully at statistical models to analyze criminal behavior. An early example of a model-based approach is in Ref. [9]. More recently Gottlieb et al. [8] have codified this approach and shown how to use statistically identified patterns to fight crime. However, all of Gottlieb's work assumes that the crime analyst has successfully associated incidents or suspects. Our work provides the means to accomplish this critical first step more efficiently and effectively than existing approaches.

3. Similarity-based approaches to data association

Our approaches to associating records in databases depend on a measure of similarity (or dissimilarity). The similarity measure shows how closely two records match on the values of individual attributes. The measure of interest for the analyst is the similarity between any two cases in the database. (For a complete discussion of similarity measures see Ref. [2].) Let $\alpha_k(A,B)$ denote the similarity on attribute k between records A and B. We compute the total similarity measure, TSM(A,B), between records A and B as a weighted sum of the attribute similarity measures, α:

$$\mathrm{TSM}(A,B) = \frac{\sum_k w_k \alpha_k(A,B)}{\sum_k w_k} \quad (1)$$

The denominator restricts the TSM to the interval between zero and one.

Clearly for us to obtain high accuracy with this total similarity measure we need effective ways to capture attribute similarity and represent it in $\alpha_k(A,B)$. We also need effective ways to weight the importance of each attribute to overall similarity and represent

these weights in w_k. For w_k to be meaningful across different attributes, we find α scores that are on a common scale. The remainder of this section gives our different approaches to these tasks.

3.1. Basic quantitative and categorical similarities

Attributes for each record in the database can be qualitative (categorical) or quantitative (typically, interval valued). Quantitative attributes present fewer problems in measuring similarity. For example, consider height. There is an obvious relationship or similarity level in a suspect described as 5′10 and a suspect described as 5′11. We can compute a similarity for quantitative attributes as

$$\alpha_k(A,B) = 1 - \frac{v_k(A) - v_k(B)}{R_k}$$

Where A and B are the different records, k is the index of the attribute, $v_k(\cdot)$ is the value of attribute k for the specified record and R_k is the range of attribute k.

A simple and commonly used approach for qualitative attributes uses a binary similarity measure. Specifically, the similarity measure, $\alpha_k(A,B)$, between two records A and B is 1 if the attribute values exactly match on attribute k and 0 if the cases do not match. For instance, if the reported hair color for two suspects does not match, then the similarity score is zero for that attribute.

To calculate the weights on the attributes we exploit the following characteristics of the attributes for the crime analysis problem. Each attribute has three characteristics or properties: the difficulty to change or disguise, the difficulty to perceive, and the likelihood of data entry error. These characteristics of an attribute explain how two crimes committed by the same person can be reported differently. That is, the criminal can change his behavior or appearance, the victim or witness can perceive the same action or physical characteristics of a perpetrator differently, or the data can be recorded differently or inaccurately.

With this understanding, we argue that the attributes that are easy to perceive but hard to change or disguise and less prone to error should receive more weight because they tell the most about the similarity of crimes. We call these basic attributes. Attributes that are difficult to perceive or easy to change or disguise should receive less weight because they tell us the least about the similarity of crimes. These are non-basic attributes. Obviously, this idea can be generalized for other problem domains.

Now ideally we would like our procedure to return the same set of records as an analyst would obtain through manual, expert analysis (without errors due to fatigue, etc.). To obtain this level of conformance, we find the weights that minimize the sum of square deviations between the average similarity scores of expert analysts and the TSM produced by the algorithm. Specifically for K attributes we find the w_k (k=1...K) that minimize

$$\sum_{i=1}^{n} (\mathrm{ESM}_i - \mathrm{TSM}_i)^2 \qquad (2)$$

where n is the number of comparisons, ESM is the average expert similarity for the comparison and TSM is as given in Eq. (1). Notice that when we substitute the right-hand side of Eq. (1) for TSM, the unknowns are the values for the weights, w_k. The minimization in Eq. (2) is constrained as follows: $w_k \geq b$ for basic attributes (b is included in the optimization) and $w_k \in (0,1]$ for non-basic attributes.

3.2. Transformed Categorical Similarities (TCS)

Clearly the binary similarity measure used in the BCS approach precludes anything but exact matches on categorical or qualitative variables. A simple extension to BCS is to allow for partial matches on categorical variables and measure these accordingly. The TCS approach does this by creating similarity tables that show the similarities between different values for each categorical variable. The same idea was suggested for clustering by Anderberg [2] who called them disagreement indices rather than similarity tables.

For example, consider the reported hair color of a suspect. We create a similarity table such as Table 1.

Table 1
Example similarity table for hair color

Hair color	B	Br	Bl	G	U
B	1.0				
Br	0.8	1.0			
Bl	0.4	0.6	1.0		
G	0.5	0.6	0.6	1.0	
U	0.5	0.5	0.5	0.5	0.5

We created this table and the ones used in the system through interviews with crime analysts in Richmond, Virginia.

Table 1 uses the following code: B=black, Br=brown, Bl=blonde, G=gray, U=unknown or not reported or other. Weights in this approach are calculated as they were for the BCS. Again attributes can be either basic or non-basic.

3.3. Dynamically Adjusted Weights (DAW)

In both BCS and TCS, the weights assigned to the attributes and used in Eq. (1) are constants. However, in some situations allowing the weights to change can more accurately capture the importance of that attribute in the matching operation. For example, in investigating robberies we might believe that the weapon used is the most important attribute. However, if all robbery suspects use a gun, then this attribute does a poor job in discriminating among suspects despite its prior importance. In contrast if one criminal uses an unusual weapon, say a Japanese sword, then even if the weapon used were not a priori important it becomes an important discriminator based on the unusualness of this occurrence. Hence, the less probable a match then the higher the weight that that attribute should have when determining total similarity (cf. Ref. [2]).

This idea has been explored in the multi-criteria decision making literature (see for example, Churchman and Ackoff [5] and Keeney and Nair [11]). The approach we use here bears closer similarity to the entropy approach of Zeleny [14]. In particular, Zeleny used entropy as a measure of divergence among the values of an attribute. If an attribute has low divergence then its importance to decision making is greatly reduced. As an example, Zeleny notes that the presence of fluoride in toothpaste is important to most consumers but since it has low divergence (all toothpaste brands have fluoride) then its importance to decision making is negligible. To measure divergence among the attribute values, Zeleny proposed using entropy and adjusting the attribute weights according to a normalized entropy score.

For data association we have an analogous situation but one that requires a somewhat different criterion for setting the weights. In particular, we are not as interested in the entropy among values as in the information conveyed by a particular value for a particular attribute. For example, a match on green hair color should carry higher weight given the relatively low numbers of individuals with this color. In other words the value of green hair gives us more information. Also as in our earlier example, matches with an unusual (low probability) weapon, such as a Japanese sword, should carry higher weights than matches among more commonly observed objects. In both examples, these matches on low probability events convey more information to the analyst about the likelihood of association.

We make the same argument about interval variables with values at the extremes of their distributions. Consider matching unusually tall or short suspects. The low probability of multiple occurrences of these extreme values gives us greater confidence or more information that the records associate.

We can measure the information transmitted by the occurrence of a specific value in an attribute using standard concepts from information theory. Let $I(A \approx B; v_k(A), v_k(B))$ represent the information that records A and B associate given the values of attribute k for both records. Now consider the usual axioms of information theory as applied to our data association problem.

(1) $I(A \approx B; v_k(A), v_k(B))$ should be a function only of the prior probability of association before the values of the attribute are observed and only of the posterior probability after their observation.

(2) If the values of two attributes are statistically independent evidence of the association of the records then the combined information in their observation should be the sum of the information provided by their separate, sequential observation. Formally,

$$I(A \approx B; v_k(A), v_k(B), v_j(A), v_j(B)) = I(A \approx B; v_k(A), v_k(B)) + I(A \approx B, k; v_j(A), v_j(B))$$

The left-hand quantity is the information about the association of the records when we get the values of both attributes simultaneously. The first term on the right is the information we would get from first observing the values on one attribute (k). The second term on the right is the information we would get from now updating the information we had from attribute k with the arrival of the values of attribute j.

(3) Finally, we require the evidence of multiple associations to be additive. For example, suppose we have four records, A, B, C, and D. Then let the information that A and B associate given the height of the suspect, evidence contained in attribute k, be $I(A \approx B; v_k(A), v_k(B))$. All other evidence is inconclusive about this association (i.e., either equal or unknown). Now, let the information that C and D associate given the weapon used, evidence in attribute j, be $I(C \approx D; v_j(C), v_j(D))$. Then the additivity assumption requires that the information that both A associates with B and C associates with D given the simultaneous presence of the suspect height and weapon used evidence be the sum of the information for each pairwise asso-ciation given the separate pieces of evidence. Formally,

$$I(A \approx B, C \approx D; v_k(A), v_k(B), v_j(C), v_j(D)) = I(A \approx B; v_k(A), v_k(B)) + I(C \approx D; v_j(C), v_j(D))$$

Taken together these axioms imply (see Ref. [7] for details) that information for association given in the values of attribute k for records A and B should be measured by

$$-\kappa \log \frac{p(A \approx B; v_k(A), v_k(B))}{p(A \approx B)}$$

where κ is a constant, the denominator is the prior probability of association, and the numerator is the posterior probability of association given the values observed on attribute k in records A and B. The prior probability is the probability that the records associate before we observe any attribute data. Typically, we obtain this probability based on just the frequency with which records associate.

The posterior probability is the probability of association after we observe the values of the attributes. We typically find this probability through the application of Bayes rule:

$$p(A \approx B; v_k(A), v_k(B)) = \frac{p(v_k(A), v_k(B); A \approx B) p(A \approx B)}{\sum_k p(v_k(A), v_k(B); A \approx B) p(A \approx B)}$$

The probability, $p(v_k(A), v_k(B); A \approx B)$, is known as the likelihood. Unfortunately, we do not know this function for the crime analysis data association problem. So we will need to obtain it using techniques other than Bayes rule, as we discuss below.

Now since the records might or might not associate, we want to measure the expected value of the information given the values observed for attribute k. We take the expectation under the distribution for the posterior which gives a measure known as the relative entropy (see, Ref. [4]) or

$$\begin{aligned} i_k(A,B) = {} & p(A \approx B; v_k(A), v_k(B)) \\ & \times \log \frac{p(A \approx B; v_k(A), v_k(B))}{p(A \approx B)} \\ & + p(A \not\approx B; v_k(A), v_k(B)) \\ & \times \log \frac{p(A \not\approx B; v_k(A), v_k(B))}{p(A \not\approx B)} \end{aligned} \tag{3}$$

where $A \not\approx B$ indicates that the records do not associate. Notice that this measure treats attributes that give information of lack of association equally with attributes that contribute to the probability of association. We now have a measure for the information found in the value of an attribute. In other words, a measure that dynamically adapts to the specific data association problem.

However, to use this measure we need to estimate the prior and posterior probabilities. For prior probabilities, we use the average number of records that associate for a specified crime divided by the total number of records. This is the information we have prior to the observation of the attribute values. In practice, the total number of records is time limited (e.g. last year or last 2 months), but without loss of generality we suppress this time dependence in our discussions here.

To find the posterior probabilities let M_k be the total number of records in the database that have a field for attribute k. Let $m_k(A)$ be the number of records that have the same value as attribute k in record A. Then $m_k(A)/M_k$ is an estimate of the upper bound for the posterior probability of association among records A and B with exact matches on attribute k. If the attribute entries for the records were made without error, then this estimate would be an upper bound since associated cases would have to match on this attribute, although non-associated records could also match on a single attribute. Similarly, $[m_k(A) + m_k(B)]/M_k$ is the estimated upper bound for the pos-

terior probability of association for inexact matches between records A and B on attribute k. This means, for example, that the value for attribute k in a record could be either the one in records A or B and it would still be considered a match. Using these estimates in Eq. (3), we compute an estimate of the least amount of information provided by the attribute (i.e. a conservative measure of the worth of the attribute for association).

The new Total Similarity Measure with the information theoretic based weights is

$$\mathrm{TSM}(A,B) = \frac{\sum_k \hat{i}_k(A,B) w_k \alpha_k(A,B)}{\sum_k \hat{i}_k(A,B) w_k} \quad (4)$$

where $\hat{i}_k(A,B)$ indicates our estimate for the information measure in Eq. (3). For the extreme case of attributes that do not cause a change from prior to posterior the value of $\hat{i}_k(A,B)$ is zero and the attribute contributes nothing to our association process. To find the w_k weights, we again solve the mathematical program with the objective function as in Eq. (2) but with the TSM computed by Eq. (4).

This approach provides us with a way to dynamically vary attribute weights depending upon the probability of an observation. However, two implementation problems exists. First, really unusual attributes might not be coded and entered into the database. For example, again consider the case of the robber who uses a Japanese sword. Most police only allow for a small set of entries and would code this weapon as "other." Thus, the information contained in the specification of the weapon as a Japanese sword is lost. However, this more detailed information is typically contained in the narrative information in the police incident report. Given the potential for improved matching we have implemented our approach with the goal of linking it to this more detailed information.

A second implementation problem is that the technique depends on parameter values that are unique to crime types. For example, the prior and posterior distributions are found by crime type. Fortunately, these calculations are fast the number of crime types is less than 100, so a complete implementation of our techniques could easily handle all available crime types.

4. Evaluation of similarity-based data association

To evaluate the similarity-based data association methods in this paper we collected data on robberies in the city of Richmond, VA. These data consist of records created by the officers in the Criminal Intelligence Unit of the Richmond Police Department. Between January 1996 and January 1998, there were approximately 2900 robberies reported in the City of Richmond. Each of these robbery records has 33 attributes.

There are two high level goals for a data association tool for crime analysis: efficiency and effectiveness. For most association tasks, a computer tool will not improve on the effectiveness of human analysts because the association of records for many problem instances is an easy, albeit tedious, task. Some of the records, however, present challenges to the analysts and will produce different association results depending on the analyst and his/her experience. For these reasons, we can measure the effectiveness of the data association tool by comparing its results with the average of the analysts. Our goal is agreement with this average of expert opinion.

Efficiency is a major reason for using a data association tool. Looking through hundreds, if not thousands of records, is time consuming and tedious. For these reasons, record matching is ideal work for a computer. We measure efficiency by the amount of time required to perform associations with and without the tool. So we want a tool that is fast but in close agreement with expert crime analysts.

The remainder of this section provides a detailed description of our evaluation of the data association methods described in Section 3.

4.1. Evaluation data set

We selected 24 cases from the Richmond Database and then generated an additional 15 realistic cases. The 24 real cases were chosen for their variety on a large number of attributes. The 15 simulated cases were created to be highly similar to each other. This mix allowed us to test on actual data, while also exploring issues in similarity on individual attributes.

Our experts were three crime analysts in the Criminal Intelligence Unit of the Richmond Police Department. These analysts had varying ranges of

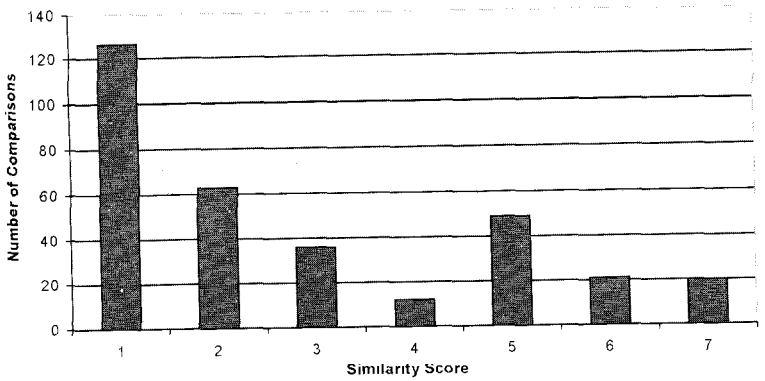


Fig. 1. Distribution of expert ratings.

experience from less than 1 year to just over 5 years. However, even the newest analyst had previous law enforcement experience.

We presented the records to the analysts in a pairwise format that allowed for easy side-by-side comparison of the 33 attributes. Each analyst performed 109 comparisons. Twenty of these comparisons were common to all analysts and used as means of evaluating similarity between experts. The other 89 comparisons were randomly chosen from the pairs of real and simulated cases. Each analyst rated the association between the records on a scale from one to seven, with seven indicating strong similarity and one indicating complete dissimilarity. We did some initial experiments with the range of this scale and found that a seven-point scale produced the most consistent results. Fig. 1 shows the distribution of ratings by the analysts.

We evaluated the agreement between experts using a weighted kappa statistic [1]. For exact matches, the pair was assigned one agreement point. For a case in which one expert categorized the similarity as a "1", very dissimilar and the other expert categorized the similarity as "7", very similar, the pair was assigned zero agreement points. For in-between assessments, the pair was assigned a linearly decreasing value from one.

Two of the experts agreed quite well, scoring a 0.816 for the weighted kappa. However, the third expert had kappa values of 0.673 and 0.679 with the other two. We examined this expert's responses more closely and also noted that this was the least experienced crime analyst. Given these facts, we decided to remove this expert's results from further use.

We randomly divided the remaining data set in half and used one part to estimate weights for the TSM as described in Section 3 and the other half to test the performance of the methods as discussed below.

4.2. Evaluation of efficiency

Table 2 summarizes the efficiency comparisons by showing the average time to perform the association tasks with the various methods. All of the new methods had comparable results so we only show those for TCS. User effort shows the amount of time needed to set-up and formulate the query. In the case of the manual approach, this includes the amount of time needed to get the records for comparison, although it does not include the time normally needed to arrange the attributes for easy viewing.

The compare time is the amount of time needed to complete the comparisons once we have the query. Both the SQL queries and the new methods had compare times considerably under a minute but we rounded up to give a more conservative sense of the comparisons given the small number of trials.

The standard error is reported for the total times. Since we had only one analyst who could formulate

Table 2
Efficiency results

		User effort (min)	Compare time	Standard error (min)
Existing	SQL query	15	<1 min	
	Manual analysis	30	2 h	6
New	TCS	5	<1 min	<1

these queries, we had only one trial with SQL queries. The standard error for the new methods is based on three trials.

Even without the standard errors the automated approaches clearly dominate the current manual method. The SQL approach is also slower than the newer methods primarily because of the need for analysts to formulate queries. Since only one analyst knew how to create these queries, this method is not even a viable option for any of the other analysts in the unit.

4.3. Evaluation of effectiveness

To measure the effectiveness of the data association tools we want to score the similarity between the results produced by the experts and those produced by the tool. This approach is consistent with our earlier comments that we are not trying to beat a human expert in accuracy, but rather to do the tedious parts of the job with similar accuracy but much faster.

An appropriate measure of similarity in this case uses the Spearman's rank correlation coefficient (SRCC). The expert and the tool could rank the cases from most similar to least similar. We then use the SRCC to compare the two rankings. The SRCC is defined as

$$\mathrm{SRC} = 1 - 6\frac{\sum_i (x_i - y_i)^2}{N(N^2-1)}$$

where x and y are the expert and algorithm ranks, respectively, for the ith data point and N is the total number of data points. We normalized the SRCC by dividing by the maximum SRCC possible on the data set and will use ASRCC for this adjusted version.

The test data consists of 117 comparisons produced by two experts. We calculated the ASRCC for both experts for each data association method and took the average. Table 3 gives the results.

Table 3
ASRCC for alternative data association methods

Method	ASRCC
SQL queries	0.2482
BCS	0.7474
TCS	0.8442
DAW	0.8442

Clearly, the SQL approach does not provide the accuracy with respect to the experts of the similarity-based approaches. Also, the DAW and TCS produced identical results because the data set did not have low probability values in any of the attributes. As mentioned in Section 3, the police currently do not record these unusual attribute values, although the information is available in the narrative section of the written report.

We found that the results in Table 3 were sensitive to a few extreme data points. Even during the several hours the experts worked on this task they became tired and more prone to misreading records. Not surprisingly, this condition was most evident when they read records near the end of the group. The three largest differences between the ranks provided by the experts and those provided by the tool fell into this late evaluation category. When they were re-evaluated later (after a day of rest) we recalculated the ASRCC. The ASRCCs for BCS and TCS (DAW) rose to 0.8457 and 0.9200, respectively.

Finally, we wanted to know if the difference in performance between the BCS and TCS (DAW) procedures was statistically significant. In order to measure the significance of the performance difference with the available set of cases, we used six-fold cross-validation. That is, we divided the entire data set into six groups, and estimated the weights using 5/6th of the data set and tested on the remaining 1/6th. We did this six times. The results showed that the better performance by TCS is statistically significant at the 0.10 level under the assumption that the ASRCCs have a large sample Gaussian distribution.

5. Conclusions

We have described our research and development of new data association tools for crime analysis. These tools show benefits for both efficiency and accuracy. They provide clear efficiency gains for crime analysts since they reduce search times by a factor of 1/3 over SQL-based search and provide for considerably greater accuracy than this current approach. They also provide several orders of magnitude improvement over manual record search times.

The techniques also do well in ordering the records in the same fashion as an experienced crime analyst.

The high value of the TCS method for the adjusted Spearman's rank correlation coefficient shows excellent agreement with the experts. Further, this testing revealed that the automated methods described here can actually do somewhat better than a fatigued analyst, since they never get tired of tedious record searches.

In comparing the methods among themselves we found no difference between the DAW and TCS approaches. As noted in Section 3, for the DAW method to perform well the records must contain low probability values. Because these values are unusual, the police currently do not record them. Clearly, a tool like the DAW method provides good reason for them to begin recording these values. We believe that when this occurs this method will help the crime analysts more quickly narrow their searches. We are also currently investigating methods to recover the unusual values from the narrative information.

The comparison between TCS and BWS was less equivocal. The TCS method does significantly better. Clearly, the use of similarity tables for non-exact matches on categorical variables gives an important lift to the performance of the search methods.

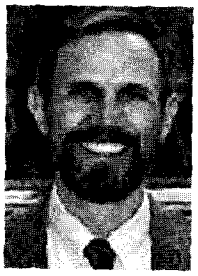
Based on the results of this work, we have implemented the TCS approach in the Regional Crime Analysis Program (ReCAP) used by several police jurisdictions in Virginia. We will continue testing the approach and work toward a more extensive test of the TCS and DAW methods.

Acknowledgements

The work reported in this paper was partially supported by grants from the Virginia Department of Criminal Justice Services and the National Institute of Justice, Crime Mapping Research Center.

References

[1] A. Agresti, Categorical Data Analysis, Wiley, New York, 1990.
[2] M.R. Anderberg, Cluster Analysis For Applications, Academic Press, New York, 1973.
[3] A.B. Badiru, J.M. Karasz, B.T. Holloway, AREST: Armed Robbery Eidetic Suspect Typing Expert System, Journal of Police Science and Administration 16 (1988) 210–216.
[4] D.E. Brown, R.L. Smith, A correspondence principle for relative entropy minimization, Naval Research Logistics 37 (1990) 191–202.
[5] C.W. Churchman, R.L. Ackoff, An approximate measure of value, Operations Research 2 (2) (1954) 172–187.
[6] W.F. Coady, Investigating with APES, Security Management 31 (1987) 67–70.
[7] A. Feinstein, Foundations of Information Theory, McGraw-Hill, New York, 1958.
[8] S. Gottlieb, S. Arenberg, R. Singh, Crime Analysis from First Report to Final Arrest, Alpha Publishing, 1994.
[9] D.H. Harris, Development of a computer-based program for criminal intelligence analysis, Human Factors 20 (February 1978) 47–56.
[10] D.J. Icove, Automated crime profiling, Law Enforcement Bulletin 55 (1986) 27–30.
[11] R.L. Keeney, K. Nair, Nuclear siting using decision analysis, Energy Policy, September 1977.
[12] E.C. Ratledge, J.E. Jacoby, Handbook on Artificial Intelligence and Expert Systems in Law Enforcement, Greenwood Press, New York, 1989.
[13] Software Solutions, Security Management, March 38 (1994) 17.
[14] M. Zeleny, Multiple Criteria Decision Making, McGraw-Hill, New York, 1982.



Dr. Brown is Professor and Chair of the Department of Systems and Information Engineering, University of Virginia. Dr. Brown received his BS degree from the United States Military Academy, West Point, the MS and M.Eng. Degrees in Operations Research from the University of California, Berkeley and the PhD degree in Industrial and Operations Engineering from the University of Michigan, Ann Arbor. Dr. Brown is a past president of the IEEE Systems, Man, and Cybernetics Society and a Fellow of the IEEE. His research interests are in the areas of data fusion, data mining and pattern analysis with applications to problems in security and safety.

Mr. Hagen is currently a research scientist with the Complex Systems Research Center at the University of New Hampshire, where he does research in remote sensing. He received his BS and MS degrees from the University of Virginia in Systems Engineering.