Restructuring Databases for Knowledge Discovery by Consolidation and Link Formation*

Henry G. Goldberg and Ted E. Senator

U.S. Department of the Treasury – Financial Crimes Enforcement Network (FinCEN)
2070 Chain Bridge Road, Vienna, VA 22182
goldberg@itd.nrl.navy.mil, senator@snap.org

Abstract

Databases often inaccurately identify entities of interest. Two operations, consolidation and link formation, which complement the usual machine learning techniques that use similarity-based clustering to discover classifications, are proposed as essential components of KDD systems for certain applications. Consolidation relates identifiers present in a database to a set of real world entities (RWE's) which are not uniquely identified in the database. Consolidation may also be viewed as a transformation of representation from the identifiers present in the original database to the RWE's. Link formation constructs structured relationships between consolidated RWE's through identifiers and events explicitly represented in the database. Consolidation and link formation are easily implemented as index creation in relational database management systems. An operational knowledge discovery system identifies potential money laundering in a database of large cash transactions using consolidation and link formation.

Introduction

Real databases often contain incomplete, inconsistent, or multiple identifications of entities of interest. For example, a marketing database of purchases with multiple vendors may have different names or account numbers for the same person, or different people with the same name. In order to discover interesting and useful information about purchasing habits, either specific to individual people or about groups of people, it is necessary first to identify accurately the individuals represented in the database. This process, of disambiguating and combining identification information into a unique key which refers to specific individuals, is called consolidation.

Some real databases record transactions involving multiple individuals. Discovering useful information, such as anomalies which may indicate fraud, in these databases frequently requires constructing networks of individuals by linking together related transactions into a pattern of activity. The process of creating these networks is called link formation. These networks can then be analyzed or evaluated by applying techniques for learning about structured information (Stepp 1986), or by other techniques such as visualization (Davidson 1993). Techniques not only for forming, but also for examining, modifying, analyzing, searching, and displaying these networks are collectively referred to as link analysis, and are widely used in law enforcement and other forms of intelligence analysis (Andrews 1990).

Practical KDD applications that have been previously reported are based on databases that do not require consolidation or link formation. For example, (Anand 1993) assumes that products and categories are given at the outset, (Brachman 1993) assumes that customer identities are known, and (Major 1992) assumes that health care providers are uniquely identified. Omission of these operations from today's integrated KDD systems such as those described by (Michalski 92), (Piatesky-Shapiro 1992), (Brachman 1993) and (Carbone 1993) limits their potential utility.

Motivation: Why Consolidate and Link ?

The need to consolidate and link depends both on the objectives of the discovery task and characteristics of the target databases.

Situations which require consolidation commonly feature a large population of fairly interchangeable real world entities (RWE's) with potentially overlapping identifications, such as people who may share exact – or have similar – names. While it is tempting to launch directly into characterization, classification and categorization efforts, this inevitably leads to errors. In many databases, the primary computational problem is deciding which records represent facts about which particular RWE, and then combining features of these records into a complete picture of activity of that RWE. For example, all transactions involving only a particular person, regardless of the spelling of his name or which "id" number was provided, must first be identified, and

* The authors of this paper are employees of the Financial Crimes Enforcement Network (FinCEN) of the U.S. Department of the Treasury, but this paper in no way represents an official policy statement of the U.S. Treasury Department or the U.S. Government. The views expressed herein are solely those of the authors. This paper implies no general endorsement of any of the particular products mentioned in the text.

then grouped in order to obtain accurate information as to that person's income, purchases, or account activity. We refer to this operation, of identifying transactions with a particular RWE and then combining the identified transactions, as **consolidation**. Although, in some lucky circumstances, it is a trivial task (e.g., when a valid and unique id number is available and properly recorded for each entity), in many real database applications, it is not.

Once instances of data about RWE's are reasonably well consolidated, they may be **linked** together to form more complex patterns, which are often the real objects of interest. For example, individual financial transactions are rarely recognizable as criminal until seen in the context of a pattern of activity, often by several distinct but related persons.

Task Characteristics: The Goal of Discovery

(Piatesky-Shapiro 1994) identifies clustering, data summarization, learning classification rules, finding data dependency networks, analyzing changes, and detecting anomalies as technical approaches encompassed by KDD. These approaches correspond to distinct discovery goals. Accurate discovery with any of these approaches sometimes requires that consolidation or link formation precede their application, in order to ensure that the data of interest – as opposed to the available data – are used as the basis for discovery. After the entities are uniquely identified, or the networks created, additional summary or aggregate attributes must often be computed prior to the fruitful application of discovery techniques. For example, after identifying all transactions belonging to a particular consumer from multiple vendors' databases, attributes such as total expenditures per month or expenditures in distinct categories of goods or services may be of interest. KDD tools may, in fact, be applied to aid in the discovery of concepts which suggest computation of specific derived attributes that are relevant for certain classifications.

Identification of classes of customers, as in (Brachman 1993), for marketing purposes requires consolidation; without it, the total number of customers will appear too large, activity per customer will appear too small, duplicate mailings will result in increased marketing costs, and marketing strategies based on customers with similar behavior may be ineffective.

A frequent task area for KDD applications is the identification of anomalies in databases, which may indicate fraud. Often, one first classifies the databases into categories and then identifies anomalies within each class. Other times, one looks for changes in patterns of activity over time with respect to specific accounts. Because the task is to find specific instances of anomalous behavior with respect to an individual person or account, transactions referring to an individual person or account must be identified as such. In some databases this identification is present (such as fraud detection in credit card or cellular telephone usage, where a change in the pattern of usage could indicate a stolen card or telephone); however, in some databases it is not. For example, identification of potentially fraudulent health care providers based on a comparison of individual provider aggregate claims activity to norms requires that the individual providers be clearly identified. In fact, determination of the appropriate norms themselves from the same database also requires consolidation, suggesting a bootstrap approach to system development. Identification of patients who may be attempting fraud, either by themselves or in concert with a provider(s), requires that patients be clearly identified across providers, and that networks of connections between patients and providers be constructed, because the fraud is not necessarily characteristic of an individual transaction, but rather of a pattern of activities by related individuals and/or providers distributed in time. Summarizing data about health care usage, for purposes of designing cost-effective health insurance policies, would require that individual patients be identified before classes of patients could be discovered.

DB Characteristics: The Raw Material of Discovery

This section illustrates realistic situations in which consolidation is required in order to discover relevant knowledge in a database. It frequently occurs in transaction oriented databases in which RWE's engage in distinct transactions over time. In real situations, various combinations of the following features may occur.

Data Entry Errors: In a database of people who are indexed by name and account number – a fairly common situation that could occur in a customer database – a repeat customer might fail to provide his account number, or might give an alternative spelling of his name (say, without a middle initial, or a different transliteration of a non-English name), or it might be mis-typed, leading to his identity's being recorded differently.

Unforeseen Requirements: (Brachman 1993) points out that many KDD systems use data for purposes other than that for which it was originally collected. Airline reservation systems are a common example of large database systems. When frequent flier programs were introduced, they required new tracking systems, separate from the reservations databases, because reservations databases are organized by flights, not by passenger, and because passengers may use different names on different flights. Airlines, unable or unwilling to review their flight manifests to update and maintain automatically

frequent flier accounts, can rely upon passengers to provide this information because they benefit from it.

Data Collection Cost: Sometimes it is not cost effective to collect complete or accurate identifications for the purposes of authorizing a transaction. For example, a marketing firm that compiles mailing lists for catalogs would not insist on a full name, complete with middle initial, or many requests might be refused. It is cheaper to accept all requests than to require verification.

Multiple Data Reporters: Often data from multiple reporters is combined into a single database, with no common identification required. The difficulties of accurate credit bureau reporting would appear to be an example, as would a marketing firm that obtains mailing list information from several other direct marketers.

Combination of Databases: A major trend in recent years is to combine information from multiple databases; unless the databases which are being combined have identical keys for the entities of interest – or unless there is a one-to-one mapping from one to the other – consolidation is necessary. An example would be the construction of an overall consumer profile based on purchases from multiple vendors and credit information from multiple accounts.

As a second example, a tax agency might want to combine information about people's incomes and automobile ownership to identify potential tax evaders. The income information might be indexed by name and social security number, while automobile information might be indexed by name and driver's license number. Current KDD technology could be used to identify which combinations of values of attributes such as income, age, occupation, automobile make, model, and age are useful predictors of tax evasion; however, a precursor to doing so would be to identify accurately all individuals appearing in both databases. It might also be desirable to treat some distinct owners as equivalent, such as husband and wife.

Transactions Occurring over Time: Multiple transactions by the same individual, with non-identical identification information being supplied with each transaction, is the most common condition leading to the need for consolidation. This feature is shared by the examples discussed above. Identification information can change over time simply at random; a customer could arbitrarily supply one of several credit card account numbers. Some information, not normally considered an identifier but useful for distinguishing between individuals, such as address or telephone number, could legitimately change if a person moves. Finally, an individual who was concerned about privacy could intentionally vary identifications in order to make consolidation more difficult.

Transformations of Representation in the DB

Both consolidation and link formation may be interpreted as transformations of representation from the identifications originally present in a database to the RWE's of interest. Although a general formulation is not yet available, certain realistic assumptions lead to practical implementations of these transformations in commonly available commercial relational database management systems (RDBMS).

Consolidation

The simplest case is a flat-file database, in which a row refers to a single party transaction, an assumption also made by most machine learning algorithms (Frawley 1993). Viewing the database as a set of transactions {T} and the set of RWE's as {R}, consolidation may be implemented by assigning a partition, i.e., a set of subsets of {T} such that every element, T, is an element of exactly one subset, which corresponds to a unique R. All transactions about a particular RWE may be grouped, and then summarized and/or aggregated, to describe its behavior. In database terms, producing an index for every R and storing it in T allows efficient access to data by RWE. In knowledge representation terms, we have transformed a transaction-based representation to a RWE-based representation, as depicted in figure 1.

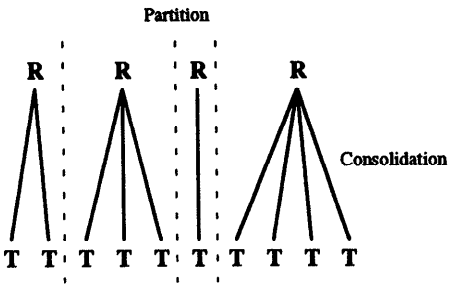

Figure 1 - Simple Consolidation

If the database structure is more complex – as is typical in most real relational databases – it can be flattened or denormalized easily by use of the relational "join" operation, resulting in the simple situation discussed in the previous paragraph. Thus it is not only theoretically possible, but also practical, to view a more complex database as a simple set of transactions.

If transactions permit multiple parties drawn from the same population of RWE's (e.g., people), then the partition applies to the subparts of the transactions identifying each party. If transactions involve distinct populations of RWE's (e.g., people and businesses),

partitions are applied to each population independently, resulting in the situation depicted in figure 2. A transaction can be indexed by any number of partitions. KDD techniques can then be used to discover information about the individual types of RWE's from each distinct population.

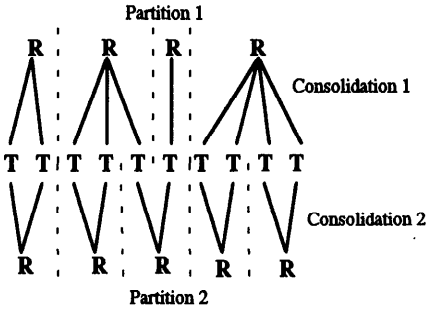

Figure 2 - Multiple Consolidations

The criteria for assigning a particular partition are highly domain dependent. If the correct partition – corresponding to the actual population of RWE's – can not be determined from the information in the database (e.g., where only names of persons are available), it may be useful to compute and maintain alternative consolidations defined by distinct criteria. Each consolidation corresponds to a different choice of partition function and results in its own database index.

If transactions arrive over time, the partitions may be computed incrementally by careful indexing. It will also help performance to choose partition functions that are decomposable into relational selects on only a few indexed fields. Obviously, this sort of processing cannot be done in an *ad hoc* manner for large databases. However, we view the choice and design of such database transformations as part of the domain knowledge that is often a necessary precursor to KDD.

Link Formation

Consolidation produces a one or more transformations of the database from transactions to RWE's, after which KDD techniques may be fruitfully applied. However, discovery of knowledge that depends upon the structure of groups of RWE's requires computing of linkages between RWE's. Consolidation produces sets of transactions relating to individual entities; linkage produces sets of transactions relating to structured groups of entities. This is one of the most fundamental operations of link analysis. Figure 3 illustrates such linkage groups combining RWE's. Note that a transaction or an RWE may be included in multiple linkage groups, unlike the case with consolidation, so linkage groups may not be represented by a partition function. For example, a person may be a member of several unrelated organizations.

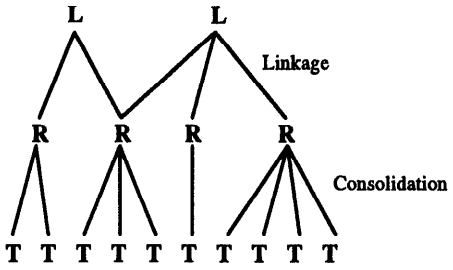

Figure 3 - Consolidations and Linkages

As an example we consider a database of two-party transactions, e.g. telephone calls between people who are identified by name. We begin by producing two consolidations, say, one that requires an exact match on first, middle and last names, and another that uses only first initials and last names. We then construct linkages, grouping the transactions into sets that are connected, and producing "calling circles." We may produce two linkage sets, one which looks for all linkages and another which requires several calls between people before they are considered connected. Obviously *a priori* domain knowledge is essential. In a database of financial transactions, for example, we would not want to link people according to which bank they use, since that would produce too few, broad networks of financial structures to be able to classify particular patterns of activity.

As with consolidation, above, a transaction oriented database can be practically maintained with the indices corresponding to linkage groups, or networks. If the consolidations are being incrementally computed and stored with the transactions, it is relatively easy to envisage an incremental process in which a network index is generated for each transaction. If a transaction shares a RWE with another, then their network indices – and those of all other transactions with that index – are merged.

A Real Example: The FinCEN AI System

The FinCEN Artificial Intelligence System (FAIS), which is described fully in (Senator 1995), is an example of an operational knowledge discovery system containing many of the characteristics described above. FAIS integrates intelligent software and human agents in a cooperative discovery task on a very large data space. The goal of FAIS is to identify leads which may be indicative of money laundering in a database of large cash transactions. Its architecture employs a database as a blackboard, because information relevant to particular problem solving instances arrives in fragments distributed over time. A set of asynchronous processes implement the operations of consolidation, link formation, derived attribute calculation, and evaluation of entities of interest (persons, businesses, accounts, etc.) according to the likelihood of

representing money laundering. During its construction, several approaches to knowledge discovery were attempted with limited success, partially due to the lack of completion of the underlying database restructuring functions that form the subject of this paper, and partially due to the inapplicability of these approaches to structured data.

FAIS confronts most of the data and task characteristics which necessitate consolidation and link formation. The database is large enough to challenge performance (~20M transactions), with a large number of attributes (~100 fields), most of which are nominative. Because of the data collection process, data is errorful and incomplete. Data is received incrementally over time. It is reported by a large number of filing institutions, introducing variability to other sources of error. Alternative forms of identification are considered acceptable. Collection of the data is a cost both to the subjects of the transactions and to the reporting financial institutions. Information from different form types is combined. Names are often non-English, making for alternative transliterations, and standard identification numbers such as social security number, are frequently non-existent for non-US residents.

The goal of discovery in FAIS is to identify anomalies, i.e., potentially suspicious behavior. Consolidation and link analysis are required because money laundering is rarely, if ever, manifested by a single transaction or by a single RWE (in this case, a person, business, or account), but rather by a pattern of transactions occurring over time and involving a set of related RWE's.

Consolidation in FAIS

FAIS consolidates data as it is received. Each person or business on a new transaction is compared to persons or businesses already in the database, and determined to be either identical to one already in the database or to be new. The consolidation criteria were obtained from knowledge engineering with expert analysts. They are implemented as a combination of SQL stored procedures and C programs. They involve a number of sequential tests on a variety of numerical and textual fields, and non-exact matches are allowed. The performance depends on the number of already identified subjects to which comparisons are made. At present the process takes about 1/3 second per new transaction on a 6 processor Sun SparcCenter 2000 server. The particular consolidation heuristics that were adopted were chosen to be conservative, in order to avoid the over-generation of potentially suspicious activity and because it is simpler to combine unconsolidated information than to separate information which should not have been consolidated. FAIS creates aggregate and summary information for each consolidated subject; not all the calculations for these derived attributes are invertible. A capability for an analyst to manually consolidate subjects that the system considered separate is also supplied, providing for complete flexibility and evaluation of any subject. Experimental evaluation of alternative consolidation heuristics against the entire database is simply too expensive on the operational system server.

Transactions included in FAIS allow for the roles of party and owner, and permit multiple individuals in these roles. Consolidation is performed across all individuals and all businesses independent of role, because the same individual or business may appear in different roles on different transactions. The number of possible networks that may be created by link formation is, therefore, extremely large. As of January 1995, 20 million transactions have been entered and linked together, resulting in 3.0 million consolidated subjects and 2.5 million accounts. On average, approximately 200,000 transactions are added per week.

Link Formation in FAIS

Creation of networks is performed manually in FAIS. An analyst starts with a seed, which is a particular subject or account of interest. The seed is selected by examining the result of the knowledge-based evaluation of suspiciousness, which is run periodically, or by examining the results of queries based on analyst defined criteria of interest. The analyst can then direct the system to find all other subjects, accounts, or transactions linked to the seed. He can then iteratively repeat the process, incrementally building a network of subjects, accounts, and transactions which appears to be suspicious.

Issues and Challenges

Consolidation and link formation can, in principle, be computed according to several techniques, e.g. clustering, equality, or domain specific heuristics. Similarity-based methods would use similar names, perhaps with some background knowledge such as a model of typing errors or knowledge of name variations in particular languages, identification numbers, and the like. Even if similarity-based methods are used for the purpose of consolidation, it is unlikely that the type described in (Stepp 1986) would apply; they attempt to learn a set of simple descriptions or a small set of descriptions while consolidation requires discovery of a large set of entities, perhaps on the order of the number of transactions in the database.

Sometimes consolidation is necessary due to lack of data standardization and could be addressed by a preprocessing step. For example, an address standardizer could be used to put street addresses in a canonical form or a name standardizer could standardize order, titles, capitalization, or initials. Unfortunately, any such

standardization comes at a cost; as soon as the original data is lost the possibility of over-consolidation is immediately introduced. The alternative is to build the standardization into the consolidation operation itself, by allowing for equivalence between elements believed to be the same. A hybrid approach, which would retain all the original, reported data but adopt a canonical form for the consolidated entity, is what we chose to adopt in FAIS.

Implementing a particular consolidation algorithm on a particular database – usually in an incremental mode – is feasible on typically available computing resources (as with FAIS), but exploring the space of potential consolidations - usually in a batch mode - could require massive computing power. Constructing the set of all possible linkages on a particular database requires constructing the transitive closure of all possible linkages, a computationally prohibitive operation in any but the most minimally connected databases.

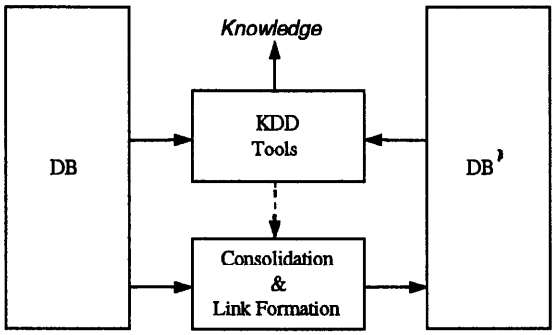

Figure 4: KDD System Augmented by Consolidation and Link Formation

Techniques for automating consolidation and link formation are in their infancy, as is a formal theory. Current systems handle these issues in an ad-hoc manner, if at all. Further research is required in both areas. A future KDD application would likely include a data clean-up and/or standardization module, a consolidation and/or link formation module, and a derived attribute calculator, in addition to those present in today's systems. A possible architecture is depicted in Figure 4. In this system, a module of KDD tools would be available for discovery in both the original DB and one transformed by consolidation and link formation, DB'. Results of KDD analysis could feed back into the consolidation and link formation modules to improve their performance. Finally, as it becomes more common to mine databases created by integrating information from multiple sources, the need for consolidation and link formation will increase.

References

Anand, T., and Kahn, G. 1993. Opportunity Explorer: Navigating Large Databases Using Knowledge Discovery Templates. In *Proc. 1993 Workshop on Knowledge Discovery in Databases (KDD-93).* Menlo Park, CA:AAAI.

Andrews, P. P. and Peterson, M. B. eds. 1990. *Criminal Intelligence Analysis.* Loomis, CA: Palmer Enterprises.

Brachman, R., Selfridge, P., Terveen, L., Altman, B., Borgida, A., Halper, F., Kirk, T., Lazar, A., McGuiness, D., and Resnick, L. 1993. Integrated Support for Data Archaeology. In *KDD-93.* Menlo Park, CA:AAAI.

Carbone, P. L., and Kerschberg, L. 1993. Intelligent Mediation in Active Knowledge Mining: Goals and General Description. In *KDD-93.* Menlo Park, CA:AAAI.

Davidson, C. 1993. What Your Database Hides Away. *New Scientist* 1855:28-31.

Frawley, W.J., Piatesky-Shapiro, G., Matheus, C.J. 1993. Knowledge Discovery in Databases: An Overview. In *Knowledge Discovery in Databases*, G. Piatesky-Shapiro and W.J. Frawley, eds., 1-27. Cambridge, MA:The MIT Press.

Major, J.A. and Riedinger, D.R. 1992. EFD: A Hybrid Knowledge/Statistical-Based System for the Detection of Fraud. *Intl. J. Intell. Sys.* 7(7):687-703.

Michalski, R.S., Kerschberg, L., Kaufman, K.A., and Ribeiro, J.S. 1992. Mining for Knowledge in Databases: The INLEN Architecture, Initial Implementation and First Results. *J. Intell. Info. Sys.* 1(1):85-113.

Piatesky-Shapiro, G. and Matheus, C. 1992. Knowledge Discovery Workbench for Exploring Business Databases. *Intl. J. Intell. Sys.* 7(7):675-686.

Piatesky-Shapiro, G.; Matheus, C.; Smyth, P.; Uthurusamy, R. 1994. KDD-93: Progress and Challenges in Knowledge Discovery in Databases. *AI Magazine* 15(3):77-81.

Senator, T.E., Goldberg, H.G., Wooton, J., Cottini, M.A., Khan, A.F.U., Klinger, C.D., Llamas, W.M., Marrone, M.P., and Wong, R.W.H. 1995. The FinCEN Artificial Intelligence System: Identifying Potential Money Laundering from Reports of Large Cash Transactions. In *Proc. 7th Annual Conf. IAAI.* Menlo Park, CA:AAAI. Forthcoming.

Stepp, R.E., and Michalski, R.S. 1986. Conceptual Clustering: Inventing Goal Oriented Classifications of Structured Objects. In *Machine Learning: An Artificial Intelligence Approach, Vol.3*, R.S. Michalski, J.G. Carbonell, and T.M. Mitchell (eds.). Los Altos, CA:Morgan Kaufmann.