View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by AIS Electronic Library (AISeL)

Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2006 Proceedings

Americas Conference on Information Systems (AMCIS)

December 2006

Identity Matching Based on Probabilistic Relational Models

Jiexun Li
University of Arizona

Gang Wang
University of Arizona

Hsinchun Chen
University of Arizona

Follow this and additional works at: http://aisel.aisnet.org/amcis2006

Recommended Citation
Li, Jiexun; Wang, Gang; and Chen, Hsinchun, "Identity Matching Based on Probabilistic Relational Models" (2006). *AMCIS 2006 Proceedings*. 189.
http://aisel.aisnet.org/amcis2006/189

This material is brought to you by the Americas Conference on Information Systems (AMCIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in AMCIS 2006 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Identity Matching Based on Probabilistic Relational Models

Jiexun Li
MIS Department, University of Arizona
jiexun@eller.arizona.edu

Gang Wang
MIS Department, University of Arizona
gang@eller.arizona.edu

Hsinchun Chen
MIS Department, University of Arizona
hchen@eller.arizona.edu

ABSTRACT

Identity management is critical to various organizational practices ranging from citizen services to crime investigation. The task of searching for a specific identity is difficult because multiple identity representations may exist due to issues related to unintentional errors and intentional deception. In this study we propose a probabilistic relational model (PRM) based approach to match identities in databases. By exploring a database relational structure, we derive three categories of features, namely personal identity features, social activity features, and social relationship features. Based on these derived features, a probabilistic prediction model can be constructed to make a matching decision on a pair of identities. An experimental study using a real criminal dataset demonstrates the effectiveness of the proposed PRM-based approach. By incorporating social activity features, the average precision of identity matching increased from 53.73 % to 54.64%; furthermore, the incorporation of social relation features increased the average precision to 68.27%.

Keywords

Identity matching; probabilistic relational models; feature construction.

INTRODUCTION

Many governmental agencies manage identity information for various purposes ranging from providing citizens services to enforcing homeland security. Identity matching is a common practice for them and is used to verify whether a person is who he/she claims to be. As digital government purportedly leads to increased integration and interoperability among agencies, identity matching becomes a key to tying together customers from different agency systems and achieving integration. In the wake of 9/11 terrorist attacks, this issue became one of the critical issues related to national security. The ability to validate identity is expected to help post-event investigation as well as to prevent future tragedies.

Identity matching, however, is a surprisingly complex problem (Camp, 2003, Kent and Millett, 2002). First, the lack of a reliable unique identifier across different agencies makes the task non trivial. For example, the Interval Revenue Services (IRS) uses Social Security Numbers (SSN) or Individual Taxpayer Identification Numbers (ITIN) as a unique identifier, while Motor Vehicle Division (MVD) relies on driver's license numbers to uniquely identify its customers. Moreover, identity information is not always reliable and is subject to unintentional errors and intentional fraud (Wang et al., 2005). Existing identity matching techniques are not adequate in solving the problems mentioned above.

In this paper we propose a probabilistic relational model (PRM) based approach to match identities in databases. Particularly, in addition to common personal identity features such as name and date of birth, features that represent an individual's social information can be constructed and used for identity matching. It is expected to solve the problem of having mismatches for the techniques based on feature value proximity.

LITERATURE REVIEW

In this section, we review the problems of identity matching and some existing techniques to address this issue.

Identity Problems

An identity is a set of characteristic features that distinguish a person from others (Donath, 1998). Identity information, however, is unreliable due to various reasons. First, unintentional errors often occur in data management processes such as data entry, storage and transformation. A study showed that the data error rate in typical enterprises could be as high as 30% (Redman, 1998). Second, identity information sometimes is subject to intentional deception, especially the identities of criminals or terrorists who are known to use false identities to mislead police investigations (Wang et al., 2004). Identity deception also exists in online auction. A customer may use false identities to register multiple user accounts in order to drive up the bidding prices (Snyder, 2000). These problems may result in multiple identity representations for an individual person in one system or across multiple systems. To efficiently manage identities, we need a mechanism to associate identities that belong to an individual. It will also be useful in searching for information about a particular person, which is a critical task for law enforcement and intelligence investigations.

Identity Matching Techniques

Biometrics information is often touted as a reliable personal identifier. However, biometrics may not identify individuals uniquely (Camp, 2003) and are subject to falsification (Matsumoto et al., 2002) as well. Not to mention that it would cost enormously to implement biometrics in all government agencies. Other techniques verify identities by comparing personal identity information against existing records in agency systems. Some rely on exact value matching, which is also called all-or-none matching. Marshall et al. provided an exact-value matching technique for law enforcement applications (Marshall et al., 2004). Two identities are considered matching only if their first names, last names, and DOB values are identical. Even if an existing identity record is very similar to the information of a subject, if it is not actually the same, an exact-match query is unlikely to bring up that record. However, as identity information is unreliable, it is possible that identities referring to the same person have disagreeing values.

Approximate matching techniques have been developed for identity matching (Brown and Hagen, 2003, Wang et al., 2004). They rely on feature value proximity to detect matching identities. Brown and Hagen proposed a data association technique for associating suspects or incidents (Brown and Hagen, 2003). It compares corresponding feature values of two records and calculates a weighted total similarity measure (TSM). Similarly, Wang et al. proposed a record comparison algorithm to detect deceptive identities (Wang et al., 2004). Given two identities, the algorithm first computes a similarity rating for the value-pair of each individual identity feature. Assuming features are equally important in making a matching decision, all the similarity ratings are combined into an overall similarity rating using a Euclidean distance function. The two identities being compared are considered matching when the overall similarity rating is greater than a threshold value. These techniques have several disadvantages. First, they may identify mismatches when a subject has common feature values such as name. Especially when the threshold value is set low, the false positive rate would be high. Second, these techniques may fail to identify matching identities when a subject is reporting an identity very different from the ones kept on file. For example, a suspect may use the identity of someone else to mislead police investigation. In those cases, feature proximity based techniques may not be sufficient.

Compared to the common personal identity features, the activity of an individual and the relationships with his/her social contacts may provide additional information in identifying the subject's identity. Moreover, unlike personal identity, social identity of an individual can hardly be altered by the individual. Therefore, social contextual information may help improve the effectiveness of identity matching. However, to the best of our knowledge, there has been little study that attempted to create features representing people's social behavior for identity matching.

RESEARCH DESIGN

We focus on two main research questions in this study. First, how can features that represent people's social behavior be derived for identity matching? Second, how can these social features improve the performance of identity matching? A database is often a rich repository of identity information. The information of individual's personal properties and their social behaviors is often stored in multiple related tables. Standard data mining methods work only with "flat" data representations (i.e., a single table) thereby losing much of the relational structure represented in the database. In this study we adopt a relational learning approach to predict the "matching" relationship between identities using a database relational structure.

Probabilistic Relational Models (PRMs)

Different from standard data mining, relational learning can extract patterns from multiple related tables in a database structure (Dzeroski and Lavrac, 2001). A formal approach for relational learning is called probabilistic relational models (PRMs) (Friedman et al., 1999, Getoor et al., 2002). Applications of relational learning and PRMs span the realms of social

network modeling, citation matching, recommender systems, etc. (Huang et al., 2004, Pasula et al., 2003). This section introduces the essences of PRMs and how to learn them from a database.

A probabilistic relational model consists of a set of classes $X_1,\ldots,X_n$ and a set of relations $R_1,\ldots,R_m$ between the entities. Each class $X \in \boldsymbol{X}$ is associated with a set of *descriptive attributes* $\boldsymbol{A}(X)$ and a set of *reference slots* (*foreign keys*) $\boldsymbol{R}(X)$. We denote the attribute A of X as $X.A$ and the reference slot R as $X.\rho$. X is called the domain of R, while the corresponding class Y that $X.R$ refers to is called the range of ρ. Each reference slot ρ denotes a mapping function from Domain[ρ] = X to Range[ρ] = Y, while ρ^{-1}, called an inverse reference slot, mapping from Range[ρ] = Y to Domain[ρ] = X. A slot chain is defined as $\tau=\rho_1.(\ldots).\rho_k$, where Range[$\rho_i$] = Domain[$\rho_{i+1}$]. Since a relational database contains multiple related entities tables, various dependencies between the attributes of related entities can be explored through slot chains.

As an extension of Bayesian network learning, PRM learning is to extract the probabilistic dependencies of various descriptive attributes and reference slots over a database. A probabilistic relational model Π is composed of an acyclic directed graph, S, and the parameters associated with it, Θ_S. In particular, S describes the dependency structure of attributes and slots by assigning a set of parents Pa($X.A$) to each $X.A$. Θ_S represents the parameters characterizing the *conditional probabilistic distributions* (CPDs) (Friedman et al., 1999). Each $X.A$ is associated with a conditional probability distribution that specifies $P(X.A \mid \text{Pa}(X.A))$. Given a complete initialization I of objects in each class X as well as their values for each attribute and references slot, a PRM can be learned by finding the model $\Pi^*(S^*, \Theta_S^*)$ that best fits I. A search-and-scoring approach is a standard process to find the best PRM. A commonly scoring metric is $\log P(S \mid I) = \log P(I \mid S) + \log P(S) + C$, where $P(I \mid S)$ is the marginal likelihood $P(I \mid S) = \int P(I \mid S, \Theta_S)\, P(\Theta_S \mid S)\, d\Theta_S$. To constrain the computational complexity, standard greedy search algorithms can be used to search for the optimal structural S^*. Given the optimal structure, parameters of CPDs, Θ_S^*, can be estimated to complete the model specification.

PRM-based Identity Matching

Identity matching is to determine whether two identities refer to the same person. Conceptually, this problem can be regarded as an application of relational learning. The linkage (match or non-match) between each pair of identities is the modeling focus. Figure 1 illustrates an example of the entity-relationship diagram (ERD) for an identity matching database. In this ERD, Person and Incident are entity classes while Participation is the relationship between Person and Incident.

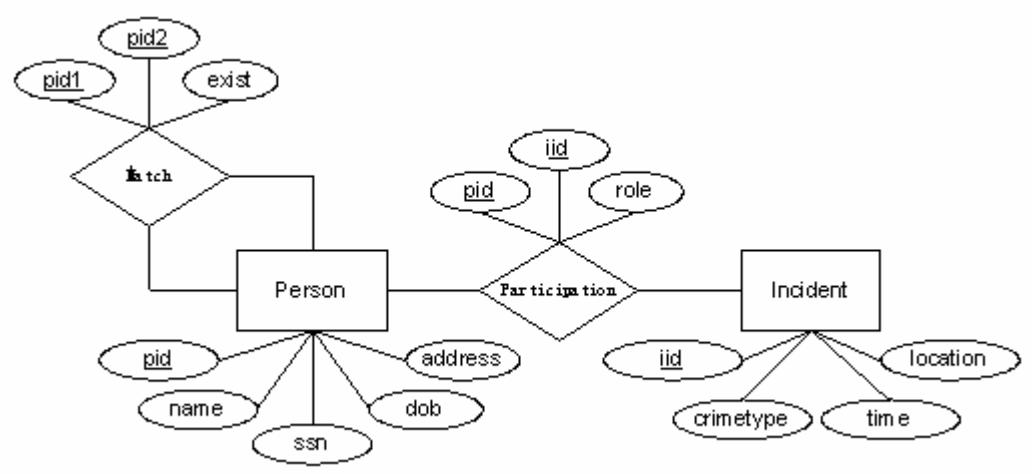

Figure 1. A Entity-Relation Diagram of a Criminal Database

PRMs with existence uncertainty are able to model the existence of certain records. Specifically for identity matching, we introduce a class of Match (pid1, pid2, exist) to model the undetermined match relationship between two individuals. In the Match class, pid1 and pid2 are foreign keys to the Person class; exist is an existence attribute whose value is from {*true*, *false*}. In particular, Match.exist equals *true* if the pair of identities refer to the same person and equals *false* otherwise.

Using the PRM notation introduced previously, the dependency structure S of PRM defines the parents Pa($X.A$) for each attribute $X.A$. For identity matching problem, since the attribute Match.exist is the only concern, we only need to focus on

the partial dependency structure for Match.exist. Potential features involved in this dependency structure can be derived though reference slot chains.

Feature Construction for Identity Matching

Starting from a target pair of individuals in Match class, various features for identity matching can be derived from slot chains. A slot chain of length=1 leads from from the Match class to the Person class. Descriptive attributes of the Person, such as name, date of birth and so on, are the simplest and more straightforward features derived from slot chains. For instance, we use [Match.pid1].dob to denote a target individual pid1's date of birth. These features compose the personal identity features that have been commonly in identity matching techniques based on feature value proximity.

As the length of the slot chain increases and inverse reference slots are introduced, more complex features can be constructed. Unlike personal identity features, these new derived features could reveal an individual's social behavior and contextual information. These social identity features can be further divided into social activity features and social relation features.

By extending the slot chain to the class of Participation and Incident, we can derive new features that represent the target individual's social activities, i.e., the incidents in which he or she is involved in. For example, [Match.pid1].[Participation.pid]$^{-1}$.iid, represents the set of incidents in which the target individual pid1 is involved; [Match.pid1].[Participation.pid]$^{-1}$.role represents the roles of the target person in his/her involved incidents; [Match.pid1].[Participation.pid]$^{-1}$.[Participation.iid].crimetype represents the crime types of the pid1's involved incidents.

Furthermore, we can extend the slot chains back to Participation and Person class to construct features that describe the social relationship of an individual. For example, [Match.pid1].[Participation.pid]$^{-1}$.[Participation.iid].[Participation.iid]$^{-1}$.[Participation.pid].pid represents the set of individuals who are involved in at least one incident with the target individual, representing the "*neighbors*" of the target individual. [Match.pid1].[Participation.pid]$^{-1}$.[Participation.iid].[Participation.iid]$^{-1}$.[Participation.pid].[Participation.pid]$^{-1}$.role represents the roles of pid1's neighbors in their involved incidents. [Match.pid1].[Participation.pid]$^{-1}$.[Participation.iid].[Participation.iid]$^{-1}$.[Participation.pid].[Participation.pid]$^{-1}$. [Participation.iid].crimetype represents the crime types of the incidents that the pid1's neighbors are involved.

These social activity and relation features represent an individual from social perspectives. We believe that the incorporation of these features could improve the performance of identity matching. In PRM learning, there are an infinite number of potential features that can be constructed by extending the length of slot chains. However, more features will also increase the computation complexity while searching for the optimal PRM. In addition, as the length of reference slot chain increases, the derived features tend to become less interpretable. In this study we constrain the length of the slot chain and only focus on these three major types of feature: personal identity features, social activity features, and social relation features.

Similarity Measures

For PRM-based identity matching, the existence of a link to predict, Match.exist, is between a pair of individuals, pid1 and pid2. Each of them is represented by the features derived separately from slot chains. Prediction in identity matching is based on the similarity between each pair of feature values. Personal identity features, such as name and date of birth, are often stored in string format. For these features, we can simply use string comparison methods such as edit distance (Levenshtein, 1966) to measure the similarity between two strings.

Social activity and relation features are often derived though long reference slot chains. In cases where chains contain one-to-many mappings, the derived feature will be of multiple values. For example, [Participation.pid]$^{-1}$ indicates a mapping function from an individual to all of his/her involved incidents. The notion of aggregation from database theory is the proper tool to address this issue by converting a multi-valued set into a single-valued feature. There are many useful notions of aggregation, such as cardinality (the number of distinct values in the set), mode (the most frequently occurring value), mean, median, maximum, minimum, etc. For example, *mode*{[Match.pid1].[Participation.pid]$^{-1}$.[Participation.iid].[Participation. iid]$^{-1}$.[Participation.pid].pid} represents the most frequent neighbor of the target individual. In addition, we can also jointly use aggregation operator cardinality and multi-set operations such as intersection and union to derive the Jaccard's coefficient, a commonly used similarity measure. For example, the similarity between the neighbors of pid1 and pid2 can be computed by *cardinality*{*intersection*{[Match.pid1].τ,[Match.pid2].τ}} / *cardinality*{*union*{[Match.id1].τ, [Match.pid2].τ}}, where τ=[Participation.pid]$^{-1}$.[Participation.iid].[Participation.iid]$^{-1}$.[Participation.pid].pid. In this formula, the numerator denotes the number of the two individuals' common neighbors and the denominator denotes the total number of their neighbors.

PRM Learning Process

With the three types of features and their corresponding similarity measures, PRM learning is to find the optimal partial dependency structure involving Match.exist. Standard hill-climbing greedy search algorithms can be employed to search for the optimal structural *S*. With the optimal dependency structure, maximum likelihood parameter estimation can be performed to complete the model specification. In this study we used a Bayesian network classification model for binary prediction (Langley and Sage, 1994) to estimate the probability *P*(Match.exist=1 | relevant features of Match.exist) for undetermined identity pairs.

EXPERIMENTAL STUDY

In order to examine the effectiveness of our proposed PRM-based approach, we conducted an experimental study on a real criminal dataset. Particularly, we compared the predictive power of the three types of derived features for identity matching

Test-bed

The test-bed used in our experimental study is a dataset named "Meth World" from the Tucson Police Department (TPD). The Gang Unit Sergeant at the TPD provided a list of 103 major criminals in the dataset and 924 criminals surrounding the major offenders. Theses criminals were involved in 11,704 crime incidents ranging from theft and aggravated assault to drug offences from 1983 to 2002. There were 18,923 identities that appeared in the reports of the 11,704 crime incidents.

Metric

In this study we use precision as the evaluation metric to assess the performance of identity matching. This measure has been widely used in information retrieval domain. In particular, precision is defined as follows:

$$\text{precision} = \frac{\text{Correctly detected matches}}{\text{Total detected matches}} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

where TP and FP represents the number of false positives and the number of false positives in the detected matches, respectively.

Hypotheses

We denote personal identity features, social activity features, and social relation features as F_p, F_a, and F_r, respectively. Particularly, we use the conventional identity matching approach based on the proximity in personal identity features F_p as a benchmark in the comparative study. To examine the predictive power of the derived social features for identity matching, we test the following two hypotheses in our experimental study.

H1: $F_p + F_a > F_p$

In H1, we hypothesize that the incorporation of social activity features can improve the performance of identity matching by reducing false positives.

H2: $F_p + F_a + F_r > F_p + F_a$

In H2, we hypothesize that incorporation of social relation features can further improve the performance of identity matching by further reducing false positives.

Experimental Design

Due to the large amount of records in the criminal dataset, it is impractical to compare all the individuals in a pair-wise fashion. To reduce the search space and improve the matching efficiency, we used the adaptive detection algorithm (Wang et al., forthcoming) to match the identities based on the personal identity features only.

The adaptive detection sorts the list of identities based on a key attribute such as name. Based on the assumption that matching identities have similar values on the key attribute, they should be located close to each other after the sorting. Thus, each identity is only compared to its neighboring ones in the sorted list within a window size. For each comparison, the similarity score of the value pair for each personal identity feature f_i is calculated and further combined into an overall similarity score, $Sim(i, j)$, by calculating a normalized Euclidean distance:

$$Sim(i, j) = \sqrt{\frac{Sim(f_1(i), f_1(j))^2 + Sim(f_2(i), f_2(j))^2 + ... + Sim(f_n(i), f_n(j))^2}{n}}$$

If the overall similarity score of two identities is greater than a pre-specified threshold, the algorithm matches the two identity records and considers them to be the same person.

The threshold value determines the accuracy of matching decisions. A large threshold value tends to yield a low false positive rate with a high false negative rate. A small threshold value decreases the false negative rate at the cost of a high false positive rate. In this study a threshold of 0.9 is arbitrarily chosen for the adaptive detection to make matching decisions. The algorithm matches 778 pairs of identities out of 18,923 individuals.

To evaluate the correctness of these matched identity pairs, we followed a rule of thumb: a pair of identity was considered as a match only if their DOB values disagreed at no more than one digit and their names disagreed at no more than two characters. Each identity pair is labeled by either 0 (i.e., non-match) or 1 (i.e., match). According to this evaluation criterion, 418 out of the 778 pairs are true matches and the other 360 are not. In other words, identity matching based on personal identity features alone yielded a precision of 418/778 = 53.73%.

Furthermore, we incorporated the social activity features and social relation features in an ordinal fashion and learn a probabilistic relational model to rectify the matching decisions. Table 1 presents the three types of features used in our experiments.

Feature Annotations	Similarity Measures	Descriptions
Personal Identity Features (F_p)		
[Match.pid1].fname	Edit distance	First name
[Match.pid1].lname	Edit distance	Last name
[Match.pid1].dob	Edit distance	Data of birth
[Match.pid1].ssn	Edit distance	Social security number
[Match.pid1].address	Edit distance	Address
Social Activity Features (F_a)		
[Match.pid1].[Participation.pid]$^{-1}$.iid	Jaccard's coefficient	The identity's involved incidents
[Match.pid1].[Participation.pid]$^{-1}$.role	Jaccard's coefficient	The identity's role in previous incidents
[Match.pid1].[Participation.pid]$^{-1}$.[Participation.iid].crimetype	Jaccard's coefficient	The crime type of the identity's involved incidents
[Match.pid1].[Participation.pid]$^{-1}$.[Participation.iid].time	Jaccard's coefficient	The time period of the identity's involved incidents
Social Relation Features (F_r)		
[Match.pid1].[Participation.pid]$^{-1}$.[Participation.iid]. [Participation.iid]$^{-1}$.[Participation.pid].pid	Jaccard's coefficient	The identity's neighbors
mode{[Match.pid1].[Participation.pid]$^{-1}$.[Participation.iid]. [Participation.iid]$^{-1}$.[Participation.pid].pid}	Jaccard's coefficient	The most frequent neighbor
[Match.pid1].[Participation.pid]$^{-1}$.[Participation.iid]. [Participation.iid]$^{-1}$.[Participation.pid].[Participation.pid]$^{-1}$.role	Jaccard's coefficient	The role of the identity's neighbors
[Match.pid1].[Participation.pid]$^{-1}$.[Participation.iid]. [Participation.iid]$^{-1}$.[Participation.pid].[Participation.pid]$^{-1}$. [Participation.iid].crimetype	Jaccard's coefficient	The crime type of the identity's neighbors' involved incidents
[Match.pid1].[Participation.pid]$^{-1}$.[Participation.iid]. [Participation.iid]$^{-1}$.[Participation.pid].[Participation.pid]$^{-1}$. [Participation.iid].time	Jaccard's coefficient	The time period of the identity's neighbors' involved incidents

Table 1. Three Types of Features Constructed for Identity Matching

Experimental Results

To examine the predictive power of the derived social activity and social relation features, we adopted a Bayesian classifier for binary prediction to predict whether an identity pair matches or not. A standard 10-fold cross validation was conducted on the dataset of 778 identity pairs to estimate the performance of the classification model. First, four social activity features were incorporated into the classification model to predict identity matching. Second, five social relation features were added into the model. For each feature set, we repeated this cross-validation process for 30 times by randomly reconstructing the 10 folds to get enough data points for statistical comparison.

The estimated TP rates of identity matching using different features are presented in Table 2. Using only the five person identity features, the adaptive detection algorithm found 778 matching identity pairs from the dataset with precision = 53.73%. By incorporating the four social activity features into the classification model, the average precision increased to 54.64%. Furthermore, the incorporation of social relation features increased the average precision to 68.27%. Furthermore, paired *t*-tests between these identity matching models demonstrated that both hypotheses, H1 and H2, were supported. Specifically, the incorporation of social activity features (F_a) significantly improved the identity matching performance (*p*-value < 0.001); the incorporation of social relation features (F_r) also significantly improved the identity matching performance (*p*-value < 0.001).

Features	True positives	False positives	Precision	*p*-value
F_p	418	360	53.73%	/
$F_p + F_a$	256.87	161.13	54.64%	< 0.001
$F_p + F_a + F_r$	249.20	115.83	**68.27%**	< 0.001

Table 2. A Comparison of Identity Matching Performance

Discussions

To better understand how the derived social features help match and differentiate individuals, we looked into the identity matching process and found several interesting identity pairs. For example, two individuals, A and B, share the same first name and last name as well as similar date of birth and address (the detailed personal information is disguised here due to the issue of confidentiality). Hence, based on personal identity features alone, they were predicted as the same person. However, a lot of discrepancy between these two individuals can be found by reviewing their social activity and relation features. Particularly, for their social activities, A was often involved in assault and offence crimes as a suspect, while B was involved in many theft and drug crimes as an arrestee. In addition, by looking at their social relationships, most of A's neighbors were victims in offences, while many of B's neighbors were suspects or arrestees in drug-related crimes. These disagreements in feature values represent their different social behavior and therefore indicate that A and B are not the same person. Another example is a pair of identities, P and Q. Again, the overall similarity score of their personal identity features were very high. However, their social activity features indicate that they were both arrested in two incidents together. Since one individual can not have two different identities in the same incident, these two are obviously not the same person. These examples demonstrate that both social action and relation features can capture social contextual information about individuals and rectify the mismatches based on personal identity features.

CONCLUSIONS AND FUTURE DIRECTIONS

Identity matching is an important and challenging task in large database management. Personal identity features such as name and date of birth have been commonly used in identity matching but usually yield to false positives. In this study we introduced a PRM-based approach to construct features for identity matching though reference slot chains in the database structure. Not only does this approach covers personal identity features commonly used in previous studies, but also derives some novel features that represent the social contextual information of individuals, i.e., social activity and social relation features. Experimental study on a criminal dataset demonstrated that the incorporation of these social identity features could significantly improve the precision for identity matching.

We are in the process of extending our work in the following directions. (1) Under the probabilistic relational model, an infinite number of features can be derived by extending slot chains and applying aggregations. We are interested in identifying the key features from these derived features so as to improve the performance and efficiency of identity matching.

(2) The current study only investigated how the derived social features help reduce the false positives for identity matching. We will also study their impact on the false negatives in the future. (3) In addition, we plan to extend the PRM-based feature construction approach to other applications such as customer relationship management in marketing domain.

ACKNOWLEDGEMENTS

This project has primarily been funded by the following grant: NSF, Digital Government Program, "COPLINK Center: Social Network Analysis and Identity Deception Detection for Law Enforcement and Homeland Security," #0429364, 2004-2006.

REFERENCES

1. Brown, D. E. and Hagen, S. C. (2003) Data Association Methods with Applications to Law Enforcement, *Decision Support Systems*, **34,** 369-378.
2. Camp, J. (2003) Identity in Digital Government, *Proceedings of 2003 Civic Scenario Workshop: An Event of the Kennedy School of Government*Cambridge, MA 02138.
3. Donath, J. S. (1998) In *Communities in Cyberspace*(Ed, Smith, M. a. K. P.) Routledge, London.
4. Dzeroski, S. and Lavrac, N. (2001) *Relational Data Mining*, Springer-Varlag, Berlin.
5. Friedman, N., Getoor, L., Koller, D. and Pfeffer, A. (1999) Learning Probabilistic Relational Models, *Proceedings of the 16th International Joint Conference on Artificial Intelligence,*, pp. 1300-1307.
6. Getoor, L., Friedman, N., Koller, D. and Taskar, B. (2002) Learning Probabilistic Models of Link Structure, *Journal of Machine Learning Research*, **3,** 679-707.
7. Huang, Z., Zeng, D. and Chen, H. (2004) A Unified Recommendation Framework Based on Probabilistic Relational Models, *Proceedings of the Fourteenth Annual Workshop on Information Technologies and Systems (WITS 2004)*.
8. Kent, S. T. and Millett, L. I. (2002) *IDs--Not that Easy: Questions About Nationwide Identity Systems*, National Academy Press, Washington, D.C.
9. Langley, P. and Sage, S. (1994) Induction of selective Bayesian classifiers., *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence* Morgan Kaufmann, Seattle, WA, pp. 399-406.
10. Levenshtein, V. L. (1966) Binary Codes Capable of Correcting Deletions, Insertions, and Reversals, *Soviet Physics Doklady*, **10,** 707-710.
11. Marshall, B., Kaza, S., Xu, J., Atabakhsh, H., Petersen, T., Violette, C. and Chen, H. (2004) Cross-Jurisdictional criminal activity networks to support border and transportation security, *Proceedings* Washington, D.C.
12. Matsumoto, T., Matsumoto, H., Yamada, K. and Hoshino, S. (2002) Impact of Artificial Gummy Fingers on Fingerprint Systems In *SPIE, Optical Security and Counterfeit Deterrence Techniques IV*, Vol. 4677.
13. Pasula, H., Marthi, B., Milch, B., Russell, S. and Shpitser, I. (2003) Identity Uncertainty and Citation Matching, *Proceedings of Advances in Neural Information Processing Systems 15 (NIPS 2002)*Cambridge, MA: MIT Press.
14. Redman, T. C. (1998) The Impact of Poor Data Quality on the Typical Enterprises, *Communications of the ACM*, **41,** 79-82.
15. Snyder, J. M. (2000) Online Auction Fraud: Are the Auction Houses Doing All They Should or Could to Stop Online Fraud?, *Federal Communications Law Journal*, **52,** 453-472.
16. Wang, G., Chen, H. and Atabakhsh, H. (2004) Automatically Detecting Deceptive Criminal Identities, *Communications of the ACM*, **47,** 71-76.
17. Wang, G. A., Atabakhsh, H., Petersen, T. and Chen, H. (2005) Discovering Identity Problems: A Case Study In *Intelligence and Security Informatics: IEEE International Conference on Intelligence and Security Informatics (ISI 2005)*Atlanta, GA.
18. Wang, G. A., Chen, H., Xu, J. and Atabakhsh, H. (forthcoming) Automatically Detecting Criminal Identity Deceptions: An Adaptive Detection Algorithm, *IEEE Transactions on Systems, Man and Cybernetics (Part A)*.