Applying Data Mining in Investigating Money Laundering Crimes

Zhongfei (Mark) Zhang
Computer Science Department
SUNY Binghamton
Binghamton, NY 13902-6000
(607) 777 2935
zhongfei@cs.binghamton.edu

John J. Salerno
Air Force Research Laboratory
AFRL/IFEA
Rome, NY 13441-4114
(315) 330 3667
John.Salerno@rl.af.mil

Philip S. Yu
IBM Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532
(914) 784 7141
psyu@us.ibm.com

ABSTRACT

In this paper, we study the problem of applying data mining to facilitate the investigation of money laundering crimes (MLCs). We have identified a new paradigm of problems --- that of automatic community generation based on uni-party data, the data in which there is no direct or explicit link information available. Consequently, we have proposed a new methodology for Link Discovery based on Correlation Analysis (LDCA). We have used MLC group model generation as an exemplary application of this problem paradigm, and have focused on this application to develop a specific method of automatic MLC group model generation based on timeline analysis using the LDCA methodology, called CORAL. A prototype of CORAL method has been implemented, and preliminary testing and evaluations based on a real MLC case data are reported. The contributions of this work are: (1) identification of the uni-party data community generation problem paradigm, (2) proposal of a new methodology LDCA to solve for problems in this paradigm, (3) formulation of the MLC group model generation problem as an example of this paradigm, (4) application of the LDCA methodology in developing a specific solution (CORAL) to the MLC group model generation problem, and (5) development, evaluation, and testing of the CORAL prototype in a real MLC case data.

Categories and Subject Descriptors

I.5.2 [**Pattern Recognition**]: Design Methodology – *classifier design and evaluation.*

Keywords

Money Laundering Crimes (MLCs), MLC Group Models, Uni-Party Data, Bi-Party Data, Community Generation, Link Discovery based on Correlation Analysis (LDCA), CORAL, Clustering, Histogram, Timeline Analysis.

SIGKDD'03, August 24-27, 2003, Washington, DC, USA.

1. INTRODUCTION

Money laundering is considered as a major crime in criminology, and is identified as one of the top group crimes in today's society. With the development of the global economy, increasing applications of the Internet, and advancement of e-business (especially e-banking), it is predicted that money laundering crimes (MLCs) will become more prevalent, more difficult to investigate, and more detrimental to the healthy development of the economy and the stabilization of the financial systems.

The investigation of MLCs involves reading and analyzing thousands of textual documents in order to generate (or hypothesize) crime group models. With such models, investigators are able to provide evidence to support prosecution of defendants, to identify other individuals who might also be involved in the crimes, and to predict and prevent crimes with similar patterns from occurring. At present, this model generation process is completely manual, and thus, extremely expensive, very tedious and labor-intensive, (typically requiring several man-months of effort). Consequently, it is highly desirable to automate this process as much as possible. By doing so there is plentiful of significant manpower savings and reduced prosecution time.

This paper describes part of an on-going government-sponsored research project involving multiple universities, industrial labs, and government agencies to develop a theory and related tools for semi-automatic MLC group model generation and analysis. This research project consists of an automatic component of model generation, and a manual interactive component for user analysis. While user analysis and validation are important and indispensable to model analysis for effective crime investigation, this paper focuses on the automatic component.

2. RELATED WORK AND CHALLENGES

Data mining has been recently extended from the traditional structured data mining to the unstructured one, including time-series, text, and Web. This work specifically addresses automatic generation of a community of data items in a particular application (MLC) and focuses on mining tagged free text data to generate MLC group models. Community generation, though there is extensive research in recent years, in general is still considered as one of the open and challenging problems in data mining research [1].

Of the reported community generation efforts in the literature, all the work focuses on automatic generation of a community based on a binary relationship given between data items. Examples of these efforts include mining on Web community [8] or topic related documents [2], collaborative filtering [7], and social network analysis [6]. All of these community generation efforts assume that there is explicit link information given between data items (e.g., Web links, user-item mappings, or scoring of items assigned by users). We refer to this paradigm as the one of the bi-party data community generation problems.

In this research, we have identified a new paradigm of problems in which there is no explicit binary relationship given between the data items, while the goal is to generate communities based on a yet-to-be-determined binary relationship between the data items. We define such a paradigm as the uni-party data community generation problem. In the MLC documents collected by the law enforcement agency, most of them contain only uni-party data, e.g., the monetary activities of a single person. An example of uni-party activity includes: "Fred Brown took $950 cash from his bank account on Feb. 2, 1994" or "John Smith purchased a used Honda with $1100 cash on Feb. 4, 1994". Clearly there is no explicit relationship between Fred Brown and John Smith reflected in the documents. Moreover, even if for some documents there might be explicit binary relationship available for the financial transactions (money sender and recipient relationship), the current technology of Information Extraction (IE) is unable to robustly capture the verbs from the text, resulting in the explicit binary relationship becoming unavailable. On the other hand, the generation of the MLC group models is essentially building up the communities of a group of persons based on certain relationships between them inferred from the documents. Hence, this is a typical uni-party data community generation problem. Another example of this problem is to generate communities of countries based on the smuggling activities of massive destruction weaponry between them from the news data. Here the smuggling relationships may not be given from IE or may not even be explicitly reported in the news, but a solution to the problem is to "infer" these relationships through the data to generate the communities of these relationships among a group of countries.

Another application of the problem paradigm is that the data per se is intrinsically uni-party data. Examples include the generation of network intrusion models from intrusion data records in all the nodes of a network; generation of a traffic accident correlation model from traffic record data monitored at all the locations in a traffic network. Note that problems in these scenarios actually are a generalized problem of finding associations based on "inferring" the unknown binary relationships among a group of the data items.

While data mining techniques have been applied to many areas in research and commercial sectors including in the related applications of financial fraud detection [3,5], there is little work reported in the applications in the law enforcement community, and to our knowledge, no research has been done in the application of MLC investigation specifically.

3. Problem Statement

The goal of automatic model generation in MLC investigation is to generate a community of data items, the MLC group model. Here the data items are those individuals involved and committed to a specific MLC being investigated. In law enforcement practice, an MLC group model is often referred to a group of people linked together by certain "attributes". These "attributes" typically are identified by the investigators based on their experiences and expertise, and consequently are subjective. They may also differ in various MLC cases by different investigators.

Since no one has addressed this problem before, we propose the use of a certain correlation as the "attributes" for link discovery in order to build up the community for model generation. The correlation is to be defined in different problems, and in this MLC group model generation problem, we have developed a specific method to define and determine the correlation, which is one of the contributions in this work. Given the correlation, we formally define an MLC group model as a graphic representation with the following information: (1) all the crime members of this group; (2) the different role every member plays in the group (e.g., who is in charge of the group; who are the core members of the group; a large crime group may have a complicated organizational structure); (3) the correlation relationships between different group members; (4) all the financial transaction history of each member in the group; and (5) the personal information of each group member.

The input data to the MLC model generation problem are typically free text documents, and sometimes also contain tables, or other more structured data. The types and format of the data may vary from different sources, such as bank statements, financial transaction records, personal communication letters (including emails), loan/mortgage documents, as well as other related reports. Ideally, if the semantics of these documents were understood completely, the link discovery based on correlation analysis would become easier. However, the current status of natural language understanding is far from being able to robustly obtain the full semantics of the documents; instead, what we are able to robustly obtain are the key entities that are relatively easy to identify and extract through IE, which typically include the four W's: who, what, when, and where.

In this project, we have a data set consisting of 7,668 free text, physical documents regarding a real MLC case provided by the National Institute of Justice (NIJ). The documents are first converted to a digital format using an OCR, and then key entities are tagged using a commercial IE tool. The tagged documents are represented as XML files. The tagged key entities include person names, organization names, financial transaction times and dates, location addresses, as well as transaction money amounts; no link information is tagged, and thus a typical uni-party data community generation problem. Figure 1 shows the goal of MLC group model generation using a hypothetical example. Note that the correlation between people in Figure 1 is not illustrated, and typically a model may be a general graph as opposed to a hierarchical tree.

4. GENERAL METHODOLOGY

Given the problem statement, the solution to the general MLC group model generation problem consists of two stages: text processing (including OCR conversion and IE tagging), and community generation. Text processing is not the focus of this

project. In this section, we propose a general methodology, called Link Discovery based on Correlation Analysis (LDCA), as a solution to the general uni-party data community generation problem. LDCA uses a correlation measure to determine the "similarity" of patterns between two data items to infer the strength of their linkage; fuzzy logic may be used in the correlation measure to accommodate the typical impreciseness of the "similarity" of patterns.

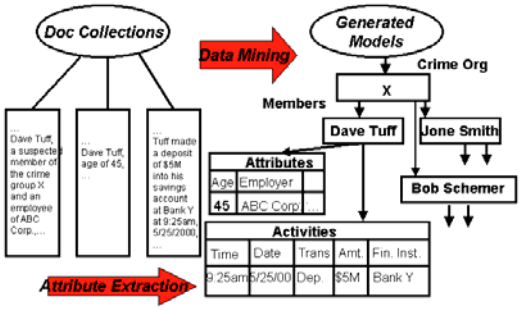

Fig. 1. An example of MLC group model generation.

Figure 2 shows the components of LDCA as well as the data flow of these components. In principle, LDCA consists of three basic steps. For each problem in the uni-party data community generation paradigm, assume that the data item set is *U*. *Link Hypothesis* hypothesizes a subset *S* of *U,* such that for any pair of the items in *S* there exists a mathematical function (or a procedural algorithm) *C* that applies to this pair of items to generate a correlation value in the range of [0, 1], i.e., this step defines the correlation relationship between any pair of items in *S*: $\forall p,q \in S \subseteq U, C: S \times S \rightarrow [0,1]$. *Link Generation* is then concerned with applying the function *C* to every pair of the items in *S* to actually generate the correlation values. This results in a complete graph *G(S,E)* where *E* is the edge set with the computed correlation values. Finally, *Link Identification* defines another function *P* that maps the complete graph *G* to one of its subgraph $M \subseteq G$ as a generated community. In the next section, we present a specific method of LDCA in the application of MLC group model generation.

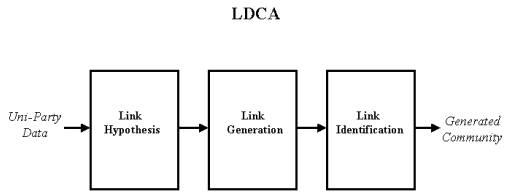

Fig. 2: LDCA components and the data flow.

5. LDCA IN MLC GROUP MODEL GENERATION

The specific method we have developed in solving for the automatic MLC group model generation problem is based on the general LDCA methodology applied in the MLC investigation context along the financial transaction timeline, and thus is called CORAL. Below we follow the general steps of LDCA to present the CORAL method.

5.1 Link Hypothesis

The Link Hypothesis of CORAL states as follows: (1) the data set *U* is the set of all extracted individuals from the collection of the given documents; (2) for each individual, there is a corresponding financial transaction history vector (may be null) along the timeline; (3) the correlation between two individuals is defined through a correlation function between the two corresponding financial transaction history vectors; (4) if two individuals are in the same MLC group, they should exhibit similar financial transaction patterns, and thus, should have a higher correlation value; and (5) any two individuals may have a correlation value (including 0), i.e., $S = U$.

Since we only have access to the isolated, tagged entities in the documents, we must make an assumption to reasonably "guess" the associated relationships between the extracted time/date stamps and the money amount of a specific transaction with the extracted individual. Therefore, when we parse the collection of documents to extract the financial transaction history vector for every individual, we follow the proposed *one way nearest neighbor* principle: (1) for every person name encountered, the first immediate time instance is the first time instance for a series of financial activities; the second immediate time instance is the second time instance for another series of financial activities, etc.; (2) for every time instance encountered, all the subsequent financial activities are considered as the series of financial activities between this time instance and the next time instance; (3) financial activities are identified in terms of money amount; money amount is neutral in terms of deposit or withdrawal; (4) each person's time sequence of financial activities is updated if new financial activities of this person are encountered in other places of the same document or in other documents; and (5) the financial activities of each time instance of a person is updated similarly.

Based on this parsing principle, we define and generate an event-driven, three-dimensional, nested data structure for the whole data set *U*: whenever a new individual's name is encountered, a new PERSON entry is created; whenever a new time instance is encountered, a new TIME event is created under a PERSON entry; whenever a new financial transaction is encountered, a new TRANSACTION event is created linked to both corresponding TIME event and PERSON entry. All the events and entries are represented as vectors. Figure 3 illustrates the data structure. After parsing the whole collection of the documents, we map the data structure into a timeline map illustrated in Figure 4, where each timeline represents the financial transaction history vector of each individual. The time axis of the timelines is "discretized" into time instances. Each node in the timelines is called a *monetary vector* that records the part of the financial transaction history of the corresponding person between the current time instance and the next time instance.

While the above "one way nearest neighbor" parsing principle may not be necessarily true in all the circumstances, we propose this principle based on the following two reasons: (1) this is the best we can do with the absence of the actual link information in the data; (2) the experimental evaluations show that the generated models based on this principle are reasonably accurate.

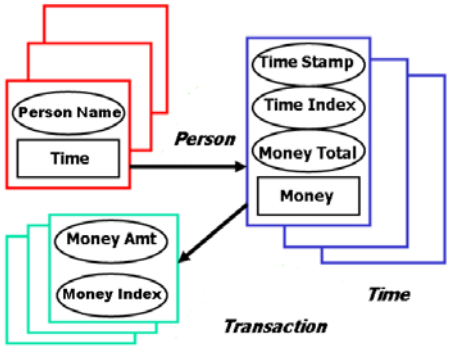

Fig. 3: Event-driven, 3-dimensional, nested data structure.

5.2 Clustering Algorithm

Given the generated timeline map, based on the Link Hypothesis, in order to accurately determine the financial transaction correlation between two individuals, ideally we wish to be able to determine which monetary vectors are "useful" (i.e., they are truly related to the MLC being investigated), and which are just noise (e.g., a "normal" purchasing activity, or a false association between one's monetary activity and someone else due to the one way nearest neighbor parsing principle). However, since we do not have the true semantics of the documents, this information is not available, and hence we would have to "guess" based on an assumption. Fortunately, during the data collection process (i.e., the law enforcement investigators manually attempt to collect all the documents that might be related to the case) the investigators typically have the intention to collect all the documents that are related to those indicted in the case, or those either suspiciously or routinely related to the case; thus, it is expected that for those individuals who might be involved in the crimes, the majority of their monetary vectors should be well clustered into several "zones" in the timeline axis where the actual MLCs are committed. We call this assumption as the *focus* assumption. Based on the focus assumption, we only need to pay attention to the "clusters" of the monetary vectors in the timeline map, and can ignore those monetary vectors that are scattered over other places. This allows us to maximally "filter" out the noise when determining the correlation between two individuals.

Assume that there are n individuals extracted in total. This clustering problem is then a standard clustering problem in an $n+2$ dimensional Euclidean space (n PERSON dimensions, 1 TIME dimension, and 1 TRANSACTION dimension). This problem may be solved through applying the standard K-means algorithm. However, taking advantage of the fact that all the n individuals share the same timeline, we can further simplify this general $n+2$ dimensional clustering problem as follows.

When we discretize the whole timeline into different time instances, each monetary vector is viewed as a node in this one-dimensional timeline space. We first simplify the problem by collapsing all the monetary vectors into scalar variables w.r.t. either accumulated money amount or accumulated transaction frequency for each monetary vector. We then project all the monetary vectors of all the individuals into the timeline axis to form a histogram. Consequently, the clustering problem is reduced to a segmentation problem in the histogram [4]. Figure 4 illustrates this concept. Since the projection and the histogram segmentation may be performed in linear time in the timeline space [4], this clustering algorithm significantly improves the complexity and avoids the iterative search the K-means algorithm typically requires. The resulted number of "hills" (i.e., segments) in the histogram becomes the K clusters.

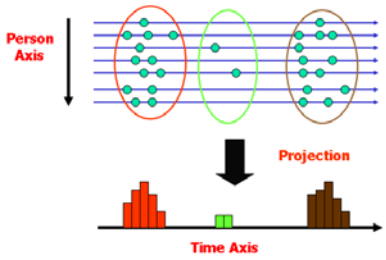

Fig. 4: Histogram segmentation based clustering.

5.3 Correlation for Link Generation

After the clustering, each individual's financial transaction history vector may be represented as a timeline histogram partitioned into K clusters, which may in turn be represented as K histogram functions of time t: $\langle f_i(t) \rangle$, where $f_i(t)$ is the financial transaction histogram of this individual in cluster i. Hence, the correlation between two individuals $\langle x,y \rangle$ is defined as a combined global correlation of all the local correlations between the two individuals, where the local correlation is defined as the correlation between two clusters of the timeline histograms of the two individuals. Figure 5 illustrates the process of determining the global correlation from local correlations between two individuals x and y. The reason why the correlation is defined as this "two level" function is due to the unique nature of the problem --- individuals in the same MLC group may exhibit similar financial transaction patterns in different time "zones" (which constrains the local correlation), but the difference in the timeline of their financial activities should not be too large (which constrains the global correlation). While the local correlation is defined following a standard approach in Pattern Recognition literature to determining a fuzzified "similarity" between two functions [9], the global correlation is defined based on the unique nature of this problem to further constrain the overall "similarity" between the financial transaction patterns along the timeline of two individuals.

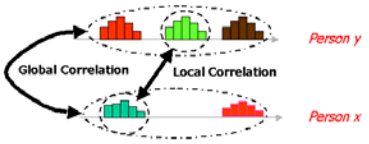

Fig. 5. An illustration of the algorithm to determine the correlation between two individuals *x* and *y* in CORAL.

To define a reasonable correlation function, it is noted that the concept of similar financial transaction patterns is always fuzzy (e.g., if two individuals belong to the same crime group and are involved in the same MLC case, it is unlikely that they would conduct transactions related to the crime simultaneously at the exact time, nor is it likely that they would conduct transactions related to the crime at times that are of a year difference; it would be likely that they conduct the transactions at two

different times close to each other). Consequently, we apply fuzzy logic in both definitions of the local and global correlations to accommodate the actual "inaccuracy" of the occurrences in the extracted financial transaction activities between different individuals at different times.

5.3.1 Local Correlation

Let $fx_i(t)$ and $fy_j(t)$ be the financial transaction histogram functions of individuals *x* and *y* in clusters *i* and *j,* respectively**.** Following the standard practice to define a fuzzified correlation between two functions [9], we use the Gaussian function as the fuzzy resemblance function *within* cluster *i* between time instances *a* and *b:*

$$G_i(a,b) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(a-b)^2}{2\sigma_i^2}}. \quad (1)$$

$$\sigma_i = \frac{2}{W_i(W_i-1)} \sum_{a=1}^{W_i} \sum_{b=a+1}^{W_i} |a-b|. \quad (2)$$

where σ_i is defined accordingly based on the specific context in this problem, and W_i is the width of the cluster *i.*

The purpose of using the Gaussian function is that it gives a natural decay over the time axis to represent the fuzzy resemblance between two functions. Consequently, two transactions of two individuals occurred at closer times results in more resemblance than those occurred at farther away times.

It is clear that after applying the fuzzy logic using the Gaussian function as the resemblance function, the resulting fuzzified histogram is the original one convolved with the fuzzy resemblance function [9].

$$gx_i(t) = \sum_{t'=1}^{W_i} fx_i(t') G_i(t,t'). \quad (3)$$

Thus, the local correlation between $fx_i(t)$ and $fy_i(t)$ is defined as the maximum convolution value

$$g(x_i, y_j) = \max_{t=0}^{W_i} \sum_{t'=-W_j}^{W_j} gx_i(t') gy_j(t-t'). \quad (4)$$

5.3.2 Global Correlation

Assuming the timeline axis is clustered into *K* segments, based on the definition of the local correlation, for each individual *x*, at every cluster *i*, there is a set of *K* local correlations with individual *y,* i.e., $\{g(x_i, y_j), j = 1, \ldots, K\}$. We give the fuzzy weights to each of the elements of the set based on another Gaussian function to accommodate the rationale that strong correlations should occur between financial transactions of the same crime group closer in time than those farther away in time. Thus, we have the following series:

$$\{g(x_i, y_j) S(i,j), j = 1, \ldots, K\} \quad (5)$$

where

$$S(i,j) = e^{-\frac{(c_i - c_j)^2}{2\sigma_i^2}}. \quad (6)$$

and c_i and c_j are the centers of clusters *i* and *j* along the timeline.

The correlation between individual *x* in cluster *i* and the whole financial transaction histogram of individual *y* is then defined based on the *winner-take-all* principle:

$$C(x_i, y) = \max_{j=1}^{K} \{g(x_i, y_j) S(i,j)\}. \quad (7)$$

Finally, the global correlation between *x* and *y* is defined as:

$$C(x,y) = \sum_{i=1}^{K} C(x_i, y) C(y_i, x). \quad (8)$$

5.4 Link Identification

After applying the correlation function to each pair of individuals in the data set *U*, we obtain a complete graph *G(V, E),* where *V* is the set of all the individuals extracted from the given collection of the documents, and *E* is the set of all the correlation values between individuals such that for any correlation *C(x, y),* there is a corresponding edge in *G* with the weight *C* between the two nodes *x* and *y*.

For the problem of MLC group model generation, we define the function *P* in Link Identification as a graph segmentation based on a minimum correlation threshold *T.* The specific value of *T* may be obtained based on a law enforcement investigator's expertise, which also allows the investigator to play with different thresholds to be able to validate different models generated based on his/her expertise.

Given *T*, from the literature there are efficient algorithms available such as the breadth-first search with complexity *O(|E|)* to conduct this segmentation. Note that there may be multiple subgraphs *M* generated, indicating that there may possibly be multiple MLC groups identified in the given document collection. It is also possible that the original graph *G(V, E)* may not necessarily be connected (the complete graph *G* may have edges with correlation values 0, resulting in virtually an incomplete graph).

6. EXPERIMENTAL RESULTS

We have implemented the CORAL method into a prototype system. In this section we first discuss the scenario of a real MLC case used in the experiments with the data given by NIJ. The CORAL prototype system is tested and evaluated based on this data set. Since the data is not considered as public domain data, we have replaced all the real names of the involved individuals and the organizations with fictitious names in this paper for the purpose of the discussion and references.

6.1 The Case Scenario

The documents used in this project were collected concerning the practices of a group of businesses, their clients and associates involved in an alleged money laundering case. The documents were obtained from an investigation of a fraudulent scheme to offer and sell unregistered prime bank securities throughout the United States. The U.S. Securities and Exchange Commission, the Securities Division of the Utopia Corporation Commission, the U.S. Customs Service and the Utopia Attorney General's Office jointly investigated the case.

It was alleged that Bob Schemer and his company, Acme Finance, Ltd., along with a group of other individuals and organizations developed a fraudulent trading scheme. Religious and charitable groups and individuals investing retirement funds were targeted. Approximately $45 million dollars were raised from more than three hundred investors. To encourage investors, Schemer, et al, misrepresented the use and safety of investors'

funds. Investors were told that their funds would be transferred to a foreign bank, secured by a bank guarantee and used as collateral to trade financial instruments with the top fifty European banks. The investors were also told that this trading activity would provide annual returns of 24% to 60%. This was not the case. Schemer, et al, did not send any of the funds to Europe for use in a trading program, and the funds were not secured by any type of guarantee. Instead, Schemer, et al, misappropriated the investment funds for unauthorized and personal uses. He also used the funds to make Ponzi payments, which is an investment scheme in which returns are paid to earlier investors entirely out of money paid into the scheme by new investors.

6.2 The Test and Evaluation Results

There were 7,668 documents in total as the whole collection that were provided by NIJ. Due to the gross OCR errors and the IE tagger errors (which was partially caused by the OCR errors), we had to manually clean up all the documents before they could be used as the CORAL input. We manually cleaned 332 documents and used this collection for testing the CORAL prototype. From the set of analyzed documents there were 252 individual names with 2,104 monetary vectors extracted in total. The distribution of the monetary vectors along the timeline was not even, with the majority obtained from those involved in the MLC case, which verifies that the focus assumption was correct.

The prototype system analyzed the collection of the 332 documents in about 20 minutes to complete the model generation on a P-III/800 with 512 MB memory running Windows 2000. Compared with the typical effort required in manual model generation, this demonstrates the significant savings automatic model generation can offer.

At this time we do not have access to the ground truth in terms of the complete list of the individuals convicted in this case as well as the role every convicted individual played in the MLC group. However, from what is reported in the news, we know that for the models we have generated with sufficiently high correlation thresholds, the individuals identified by CORAL are all convicted major crime group members. This shows that CORAL model generation method may have the capability to identify the correct MLC group members as well as to link them together to generate the groups based on the proposed correlation analysis. On the other hand, taking the example of the model generated at the threshold 0.18, we have 7 individuals that were identified as the MLC group members from the original 252 individuals extracted in the 332 documents. This shows that the elimination rate is 245/252 = 97%! Based on this "qualitative" evaluation, we are confident that CORAL method as well as the LDCA general methodology offers great promise for automatic uni-party data community generation.

7. CONCLUSIONS

We have identified a new paradigm of problems in this project, which is community generation from mining uni-party data. Unlike the traditional community generation problems such as Web mining, collaborative filtering, and social network analysis, in which the data sets are given as bi-party data, here we do not have direct and explicit access to the link information between data items. We have proposed a general methodology to solve for the problems in this paradigm, called Link Discovery based on Correlation Analysis (LDCA). As an example of these problems, we formulate and address the money laundering crime (MLC) group model generation problem, and based on the LDCA methodology, we have developed and presented a specific method to generate the MLC group model based on correlation analysis along timeline, called CORAL. We have implemented a CORAL prototype, and tested and evaluated the prototype using a data set of a real MLC case provided by NIJ. The preliminary testing and evaluations have demonstrated the promise of CORAL in automatically generating MLC group models, as well as validating the LDCA methodology.

8. ACKNOWLEDGMENTS

This work is supported in part by Air Force Research Laboratory through contract F30602-02-M-V020.

9. ADDITIONAL AUTHORS

Jingzhou Hua and Ruofei Zhang, Computer Science Department, SUNY Binghamton, Binghamton, NY 13902, and Maureen Regan and Debra Cutler, Dolphin Technology, Inc., Rome, NY 13441.

10. REFERENCES

.

[1] Domingos, P. and Hulten, D., Catching up with the data: research issues in mining data streams, *Proc. Workshop on Research Issues in Data Mining and Knowledge Discovery*, 2001.

[2] Gibson, D., Kleinberg, J., and Raghavan, P., Inferring Web communities from link topoploy, Proc. HyperText98, 1998.

[3] Goldberg, H.G. and Senator, T.E., Break detection systems, *Proc. AAAI Workshop on AI Approaches to Fraud Detection and Risk Management,* AAAI Press, 1997.

[4] Jain, R., Kasturi, R., and Schunck, B.G., *Machine Vision,* Prentice Hall, 1995.

[5] Jensen, D., Prospective assessment of AI technologies for fraud detection: A case study, *Proc. AAAI Workshop on AI Approaches to Fraud Detection and Risk Management,* 1997.

[6] Scott., J. Social Network Analysis: A handbook, SAGE Publications, 1991.

[7] Shardanand, U. and Maes, P., Social information filtering: algorithms for automating "world of mouth", *Proc. ACM CHI,* 1995.

[8] Toyoda, M. and Kitsuregawa, M., Creating a Web community chart for navigating related communities, *Proc. ACM HT,* 2001.

[9] Vertan, C. and Boujemaa, N., Embedding fuzzy logic in content based image retrieval, *Proc. 19th Int'l Meeting of North America Fuzzy Information Processing Society,* 2000.