ISSN: 2278 – 7798
International Journal of Science, Engineering and Technology Research (IJSETR)
Volume 2, Issue 9, September 2013

Fascimile Alias Detection in Malicious Environment Using Mutual Information and Similarity Model

K.J.Jaishri

***Abstract*— Alias detection and Entity consolidation has been a significant problem encountered in various areas. Aliases arise from entities who are trying to hide their original identities, from a person with multiple names or from words that are intentionally decepted. Teasing out these aliases requires effective identification and authentication. Previous system mainly focused on the novel measures with link properties such as the cardinality and uniqueness. These link properties possessed certain disadvantage by providing a matter of degree and difficulty perceived from one analyst to another in the process of alias detection. To resolve this major drawback fuzzy sets has been employed in which this provides a clear cut distinction of the particular entity belongs to a set or not. This also enhances the AOM label sets to provide a greater similarity between the attributes used by the persons who are subject to liability. Though, the previous system used the link based similarity measures it does not provide an effective alias detection to detect aliases present in an malicious environment of different network areas such as emails, financial transactions etc with different servers Gmail and face book. This can be overcome by using the information theoretic framework that comprises of mutual information model and the similarity model to automatically detect the aliases of persons present in the malicious environments.**

***Index Terms*—Alias detection, Order of Magnitude, Fuzzy sets, AOM model, Information theoretic framework, Mutual information model, Similarity model.**

I. INTRODUCTION

Entity consolidation is the problem of uncovering duplicate or near-duplicate entities in a dataset. Problem domains can be as simple as datasets containing accidentally replicated data, or as complex as populations containing criminals or terrorists wielding multiple identities. Teasing out duplicate or near duplicate entities is a serious and challenging problem. Malicious individuals, however, can easily fool such verifications by assigning completely different labels to their identities. But, their *behaviors* are likely to be similar since these are much harder to fake or separate across identities. Behaviors can be observed from various sources such as the identity, attribute usage, communication links, transaction material, social links etc. Previous approaches used the O (M) which is aimed at formalizing reasoning with approximate relations among quantities-relations like “much smaller than” or “slightly larger than”. The order of magnitude operates on an AOM model that provides a finite set of ordered labels or qualitative descriptors achieved via a partition of the real number line R. Each element of the partition represents a basic qualitative class to which a label is associated. The number of labels selected to express each variable of a real problem is subject to both the characteristics and the precision level required supporting comprehension and communication. Significant limitations appears in the AOM model due to ineffective interpretation (i.e.) the nature of magnitude can be differently perceived from one analyst to other analyst if the similarity between the aliases or duplicate attributes possessed by the persons in the “moderate” or in the “high” category. To resolve this problem fuzzy sets have been incorporated.

The theory of fuzzy sets is used to represent the AOM label sets to capture better knowledge and judgment. Fuzzy sets are sets whose elements have degree of membership. Fuzzy sets were introduced in 1965 as an extension of classical notion of set. In classical set theory the membership of the elements in a set is assessed in binary terms according to a bivalent condition- an element either belongs or does not belong to the set. Fuzzy set theory permits the gradual assessment of the membership of the elements in a set. In fuzzy set theory, classical bivalent sets are usually called crisp sets.

Qualitative link analysis introduces a novel order-of-magnitude based measures in which multiple link properties are combined to improve the quality of estimated link-based similarities achieved between the aliases of the persons. The link between the entities may be identified by using both the cardinality and uniqueness measures. The set of shared neighbors between the entities is called as the cardinality. Despite their simplicity, cardinality based methods are greatly sensitive to noise and often generate a large proportion of false positives. This shortcoming emerges because the methods exclusively concern with the cardinality property of link patterns without taking into account the underlying characteristics of a link itself.

As the first attempt to extend this approach by addressing such characteristics, the uniqueness measure of link patterns

Manuscript received Aug, 2013.
***K.J.Jaishri**, M.TECH Information Technology, Veltech Multitech Dr.RR Dr.SR Engineering College, Chennai, India, 9962675885.*

has been suggested as the additional criterion to CT to refine the estimation of similarity values.

Alias problems are commonly encountered in the intelligence community when tracking individuals from a broad population and they arise by providing deceptive identities. As a result to identify the aliases present in the malicious environments information theoretic approach is used. This concentrates on the most informative observations based on the relative importance and then compares the entities exhibiting similar behavioral observations. This comprises of a mutual information model where it generates the ranked view of the observations based on the communicated messages by two persons those who are interacting in an email transaction. The similarity model is used to detect the other false roles played by the persons by their id's, communicated messages, attributes etc.

II. AOM REASONING

Qualitative Reasoners aim at being able to model physical systems and reason at a qualitative or symbolic level. For this purpose, the initial formalism used to represent variable values is based on sign algebra (-, 0, +), which is sufficient to present the sign of quantities and implication of changes among them. However, without information about magnitudes, it has too-limited expressive power to be applicable on most realistic application domains. In order to reduce qualitative ambiguity caused by this weak abstraction of the real numbers, a number of order-of-magnitude models have been developed to permit a more detailed description of quantities, including the absolute order-of-magnitude model (AOM).

A. Absolute Order Of Magnitude Model

The absolute order of magnitude (AOM) model operates on a finite set of ordered labels or qualitative descriptors achieved via a partition of the real number line. Each element of the partition represents a basic qualitative class to which a label is associated. The number of labels selected to express each variable of a real problem is subject to both the characteristics and the precision level required to support comprehension and communication.
Submit your manuscript electronically for review.

B. Order Of Magnitude

There exists another line of research in qualitative reasoning which focuses on reasoning with relative orders of magnitude. The ultimate aim of the proposed approach is to build automated reasoning systems that mimic the process of simplifying and approximately solving equations from the knowledge of relative orders of magnitude of involved parameters.

This type of activity corresponds to a particular form of commonsense reasoning where the ideas of closeness, comparability and negligibility are involved. Relative orders of magnitude using fuzzy relations, is a promising approach for solving some ambiguity problems in qualitative reasoning.

This approach can also mechanize the commonsense reasoning of engineers simplifying complex equations and computing approximate solutions. Moreover, this approach can be applied to provide a fuzzy finest semantics to plausible reasoning with qualitative probabilities. The software realized so far can solve linear equations under closeness and negligibility assumptions.

The fuzzy relations can provide an appropriate semantics for inference rules for reasoning about relative orders of magnitude. Modeling closeness and negligibility relations in fuzzy semantics captures in a rigorous way some attenuation of transitivity for the closeness relation, as well as its reinforcement for the negligibility relation. It also provides a natural interface between numbers and qualitative terms.

III. QUALITATIVE LINK ANALYSIS

Generally, identity is a set of characteristic descriptors unique to a specific person, which can be principally categorized into three types of attributed, biographical and biometric identity, respectively. Organized criminals uses a variety of false identities such as the names, telephone numbers, date of birth etc. Among several attributes personal names one of the attributed identity is greatly subject to deception and much easier to falsify. With present high-quality equipment, it is easy to generate false identity documents. On the other hand, it requires a great deal of time and experience to distinguish between true and false copies.

With the textual attributes given as the aliases by a person identity verification and name detection system solely relies on the inexact search of the textual attributes for e.g. John dev and John Merlin. In the given e.g. by applying the text based measures the aliases with the same textual content can be detected but if the aliases given are John and Alex then the unconventional truth between the deceptive identities is found by the link analysis technique.

Attributed identity is a person's description like name, details of parents, date and place of birth, biographical identity constitutes the personal information of a person. Comparing to biometric identity like fingerprints and DNA features, the first two types are greatly subject to deception as they are much easier to falsify. So in this we are going to disclose on possibility of attributed identity based on Qualitative link analysis.

Link analysis is based on examining relation patterns amongst references of real-world entities. In Qualitative link analysis technique, the similarities among the social members are predicted by common neighbors of the social members.

The similarity between social members is determined by the "cardinality "of their shared neighbors. Intuitively, the greater the cardinality is, the higher the similarity of these members becomes.

In Uniqueness property frequency of the link occurring in social members is found by applying the orders of magnitude values in the attributes set.

A. Cardinality(CT)

In this property, the similarity among social members is decided upon "common neighbors. Consider a social network

represented as a graph G= (V, E), where v_a and v_b are the corresponding "social members" each variable is compared to another variable based on the common neighbors $Nv_a^l \in V$ where $v_a \in V$ and $Nv_b^l \in V$ where $v_b \in V$ to find the similarities among the social members. By this the common neighbors of v_a $v_b \in V$ can be identified as $[N_{va}^l \cap N_{vb}^l]$.The similarities among the common neighbors is obtained by cardinality property.

B. *Uniqueness(UQ)*

Despite their simplicity, cardinality based methods are greatly sensitive to noise and often generate a large proportion of false positives, so to further refine the estimation of similarity values uniqueness property is employed.

In this property a uniqueness value is measured by the frequency of link between objects. In the given Graph G= (V,E) ,uniqueness U_{ab}^1 of the objects a and b and the joint neighbor l is obtained by the frequency of link(F_{ab}^1) between the "common neighbor" l of the variables. The cardinality and the uniqueness property thus used to refine the similarities among the false identities.

IV. FUZZY SET BASED AOM MODEL

In Fuzzy based orders of magnitude, each variable set operates on a label value depending on the degree of relevance of "Common neighbor". Orders of magnitude are defined for each variable and neighbor set based on the underlying data of the common neighbor of the variables. Consider variables like V_a, V_b and the common neighbor of V_a and V_b are l, m and n.

The common neighbor l of V_a and V_b are compared to obtain the similar underlying data of *l*. Based on the data the label set value L^{UQ}= {very low, low, moderate, high, very high} is assigned for the V_{ab}^l and where the matching attributes of V_{ab}^l is defined in form of Discourse set (U_{ab}^l). Likewise the label values and the discourse set are assigned for $V_{ab}^m V_{ab}^n$ by comparing the neighbors *m* and *n* of V_a and V_b.

A. *Aggregation Based Similarity Evaluation*

The data set of the variables is analyzed to achieve a measure by aggregating the values of different neighbors. In general, each examined variable set is assigned with a degree of relevance. Finally, these assigned variable set are encapsulated together, by which variables Fuzzy set (F^v) with the relevant Orders of magnitude values are obtained. And these aggregated values of the variables are examined to find the neighbors of highest value to predict the similarity of the variables.

B. *Homogenization of AOM Domains*

In Homogenization of multi-granularity values, the attributes of the variable discourse set (U_{ab}^l) compared on the attributes of the Universal Discourse (U^D).The Unified discourse set is the collection of the neighbor attributes ordered in a priority manner. The Fuzzy set (F^V) of the variables are compared to the universal discourse to filter the variable set of highest priory based on the Discourse set of the variable set. By this the similarity between the variables are predicted.

V. INFORMATION THEORETIC FRAMEWORK

An information theoretic approach is used for automatically detecting aliases in malicious environments by observing the behaviors of the entities. This model discovers the most informative observations between entities and then compares them to identify entities exhibiting similar behaviors.

By this information theoretic framework all the pieces of the puzzle to detect aliases from a given population and a set of observations are first processed through our mutual information model to generate a ranked composite view of the important observations. Then, our similarity model is used to detect and rank candidate aliases for each entity in the population.

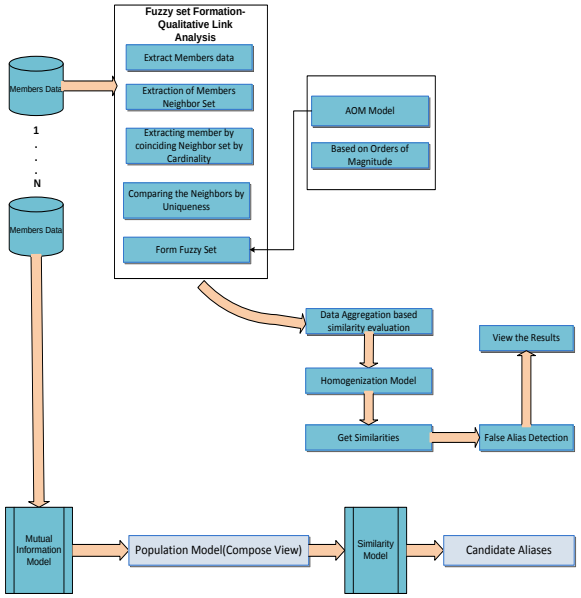

Fig. 1 System Architecture

In this system, the members have been extracted from the database and their neighbors have been extracted. The members and their neighbor sets are processed to estimate the similarities among them by using the link properties called the cardinality and uniqueness and with the help of the AOM model the label sets has been created to provide an order of magnitude for the aliases based upon their similarities.

The AOM model does not provide a detailed expression for the quantities so that it leads to vagueness and uncertainty among the data analysts because they feel a degree of difficulty to measure the values in some cases (e.g.) if the uniqueness value lies in between a moderate or a high magnitude. It also provides a single person acting as a sender communicating with a unique message with a different receiver.

Orders of magnitude are defined for each variable and neighbor set based on the underlying data of the common neighbor of the variables. Consider variables like V_a, V_b and the common neighbor of V_a and V_b are l, m and n. The common neighbor l of V_a and V_b are compared to obtain the similar underlying data of *l*.

Based on the data the label set value L^{UQ}= {very low, low, moderate, high, very high} is assigned for the V_{ab}^{l} and where the matching attributes of V_{ab}^{l} is defined in form of Discourse set (U_{ab}^{l}). Likewise the label values and the discourse set are assigned for $V_{ab}^{m}V_{ab}^{n}$ by comparing the neighbors *m* and *n* of V_a and V_b.

The dataset of the variables has been aggregated by the values of different neighbor sets and finally a fuzzy set of relative order of magnitude has been obtained and the similarity among the variables is predicted. The labels are provided to further estimate the similarities in the name, date of birth, address, alternate mobile number and mail id's. Then the similarity between the subject content and main message content is extracted The homogenization process is applied to filter the variable sets of highest priority from the collection of highest valued neighbor attributes from the variable set (i.e.) the persons those who have communicated more than once are being displayed in the AOM check process then the aliases of those persons those who have used the same attributes is detected.

In other side members from the database is extracted and based on the behaviors of the observations stored in the mutual data store is then processed to the information theoretic approach. An information theoretic framework models the importance of observations by capturing the intuition compares them to identify entities exhibiting similar behaviors. The model discovers the most informative observations between entities and measures the relative importance of such observations and leverages them to detect aliases.

A. *Mutual Information Model*

The mutual information model commonly used to measure the association strength between two events or entities. It essentially measures the amount of information one event gives to another event. For example, the message that has been transacted or communicated is namely, mission. This word mission has been checked in the database of the user profile and those persons who have communicated this message has been displayed in a ranked view of based on the relative importance by the mutual information model and then the similarity model is used to find out the other persons acting as the honorable persons as well as it helps to find out the duplicate attributes used by them as the original persons.

B. *Similarity Model*

A method of ranking observations according to their relative importance, still need a comparison metric for determining the likelihood that two entities are aliases. The requirement is that the metric handles large feature dimensions and that it not be too sensitive to 0-valued features. That is, the absence of a matching observation is not as strong an indicator of dissimilarity as the presence of one is an indicator of similarity.

VI. CONCLUSION

Alias problems are commonly encountered in the intelligence community when tracking individuals from a broad population and they arise by providing deceptive identities. As a result to identify the aliases present in the malicious environments information theoretic approach is used. This concentrates on the most informative observations based on the relative importance and then compares the entities exhibiting similar behavioral observations. This comprises of a mutual information model where it generates the ranked view of the observations based on the communicated messages by two persons those who are interacting in an email transaction. The similarity model is used to detect the other false roles played by the persons by their id's, communicated messages, attributes etc.

Instead of detecting aliases by looking for morphological, phonetic, or semantic cues in entity labels the attention has been focused on the behavioral cues exhibited by the entities (e.g. communications, financial transactions, social links, etc.), in malicious environments. Previous system uses a fuzzy set formation and qualitative link based methods but they have some have some limitations of not providing the effective alias detection in malicious environments. The above said limitations may be overcome by the information theoretic approach by detecting the aliases from the malicious environments from different networks with two different servers such as the Gmail and face book. The information theoretic model which comprises the mutual information model measures the relative importance to detect the aliases. The mutual information model then provides a ranked view of persons based on the communicated message based on the relative importance. Then the similarity model is used to detect the other kind of duplicate roles acting as the honorable persons as well as the attributes used by them.

REFERENCES

[1] A.H. Ali, D. Dubois, And H. Prade, "Qualitative Reasoning Based On Fuzzy Relative Orders Of Magnitude,"*IEEE Trans. Fuzzy Systems*, Vol. 11, No. 1, Pp. 9-23, Feb. 2003.

[2] P. Pantel, "Alias Detection in Malicious Environments,"Proc. Aaai Fall Symp. Capturing and Using Patterns for Evidence Detection, Pp. 14-20, 2006.

[3] T. Boongoen and Q. Shen, "Order-Of-Magnitude Based Link Analysis for False Identity Detection,"Proc. 23rd Int'l Workshop Qualitative Reasoning, Pp. 7-15, 2009.

[4] T.Boongoen and Q.Shen, "Nearest- Neighbor Guided Evaluation of Data Reliability and Its Applications," *IEEE Trans. Systems, Man and Cybernetics, Part B*, Vol. 40, No, 6, Pp. 1622- 1633, Dec. 2010.

[5] T. Boongoen and Q. Shen, and C. Price, "A Hybrid Link Analysis Approach for False Identity Detection," Ai and Law, Vol. 18, No. 1, Pp. 77-102, Mar. 2010.

BIOGRAPHY OF THE MAIN AUTHOR

K.J.Jaishri pursuing the post graduation in the field of Information Technology from Vel Tech Multi Tech Engineering College and received the B.Tech degree from Prince Shri Venkateshwara Padmavathy Engineering College. Presented the paper titled "XML based pushing technology for mobile users" in the national conference held at Apollo engineering college, chennnai. Presented the M.Tech Project "facsimile alias detection in malicious environment using mutual information and similarity model" in various national and international conferences.