Falsified Candidate Alias Detection Using Information Theoretic Framework

Mr.M.Rajesh Khanna¹, J.Nagaraj², Mr.P.A.Parthibhan³, Mr.C.Saravanan⁴

¹,²,³,⁴ Assistant Professor, Department of Information Technology, Vel Tech Multitech Dr.Rangarajan Dr.Sakunthala Engineering College, Avadi, Chennai.

ABSTRACT
Alias detection is a common problem encountered in various areas such as intelligence community, social network analysis, biology, marketing, databases etc. Teasing out aliases or multiple falsified identities requires an effective identification. A novel order-of-magnitude based similarity measure that integrates multiple link properties is used to refine and estimate semantic rich similarity descriptions, is incorporated with the theory of fuzzy sets to refine the semantic rich descriptors. This link based similarity measures does not provide an effective alias detection to detect aliases present in a malicious environment. This can be overcome by using the information theoretic framework that comprises of mutual information model and the similarity model to detect the aliases of persons present in the malicious environments and help to find the original roles as well as the other kind of falsified played by the persons.

Keywords: Alias Detection, Order of Magnitude, Fuzzy Sets, AOM Model, Information Theoretic Framework, Similarity Model, Mutual Information Model

1.INTRODUCTION
The O (M) is aimed at formalizing reasoning with approximate relations among quantities-relations like “much smaller than” or “slightly larger than”. O (M) is based on seven primitive relations among quantities, and compound relations formed as implicit disjunctions of consecutive primitives. In the interpretation of the relations, strict interpretation allows exact conservative inferences, while heuristic interpretation allows inferences more aggressive and human-like, by permitting some slack at each inference step.

Inference strategies within O (M) are based on propagation of order-of-magnitude relations through properties of the relations solved or unsolved algebraic constraints and rules. 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.

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.

Link analysis which seeks to discover knowledge based on the relationships in data about people, places, things and events. Link analysis can be conducted on a social network representation of relations among references of real-world entities. This section introduces a novel order-of-magnitude based link analysis in which multiple link properties are combined to improve the quality of estimated link-based similarity measures. 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. Effectively, the higher the cardinality is, the greater the similarity of the entities becomes.

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.

Alias detection has been the significant subject extensively studied for several domain applications, especially intelligence data analysis. Its performance is evaluated against text-based, link-based and other related methods, over a representative of intelligence data that contains a large number of deceptive names.

Alias problems are commonly encountered in the intelligence community when tracking individuals from a broad population and they arise by providing deceptive identities. 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. Tracking out of these aliases requires an effective and authentic deception and the assessment can be further generalized over publication and email data collections.

2. 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).

3. QUALITATIVE REASONING

Qualitative Reasoning about physical systems has been a very active subfield of research in artificial intelligence. Its main aim is both to address the need to deal with physical systems where some magnitudes are not easy to quantify precisely (i.e., numerical data is not available), and to be able to reason at a qualitative or symbolic level (for example, reasoning directly in terms of orders of magnitude). Significant progress toward the development of formal methods for qualitative reasoning about the behavior of physical systems has been made. The simplest formalism used in qualitative reasoning is based on the sign algebra (-, 0, +). Such models are sufficient to represent the sign of quantities and how the increase or the decrease of quantities can affect other quantities. Information about magnitudes or even relative orders of magnitude is not represented. As a consequence, the sign-based approach has too limited an expressive power in some practical cases to be widely applicable.

4. 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. Despite the intuition that the number of labels is not fixed, the most conventional partitions are symmetric. That is, the partition of the underlying domain typically has n positive and n negative labels, which is formally represented by OM (n), and referred to as the AOM model of granularity n. The AOM model thus provides a useful background for the alias detection of the previous system in providing a more detailed description about the quantities.

5. ORDERS 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.

6. 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.

7. **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 \(L^{\text{label}}\) \((\text{very low, low, moderate, high, very high})\) is assigned for the \(V_{a b}\) and where the matching attributes of \(V_{a b}\) is defined in form of Discourse set \(U_{a b}\). Likewise the label values and the discourse set are assigned for \(V_{a b}\) by comparing the neighbors \(m\) and \(n\) of \(V_a\) and \(V_b\).

8. **Information Theoretic Framework**

The information theoretic framework is used for detecting aliases in malicious environments. 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.

9. **Experimental Setup**

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. In order to overcome, the fuzzy set has been employed to the AOM model so that it can provide a order of magnitude for 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 \(L^{\text{label}}\) \((\text{very low, low, moderate, high, very high})\) is assigned for the \(V_{a b}\) and where the matching attributes of \(V_{a b}\) is defined in form of Discourse set \(U_{a b}\). Likewise the label values and the discourse set are assigned for \(V_{a b}\) 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 homogenization process is applied to filter the variable sets of highest priority from the collection of highest valued neighbor attributes from the variable set and then the alias 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 the unsupervised information theoretic approach for automatically detecting aliases in objects and environments by observing the behaviors of the entities is employed.
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 $L_{UQ} = \{\text{very low, low, moderate, high, very high}\}$ 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 $V_{ab}^m$ and $V_{ab}^n$ 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.

Fig. 1: System Architecture

An information theoretic framework that 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. 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, consider the following scenario of a population of South Indian residents and two particular residents Jerry and Alex. If you were told that last year both Jerry and Alex called the Chennai area about 21 times a month, then would this increase your confidence that Jerry and Alex are the same person. Now, suppose we also told you that John and Alex both called Madurai about 21 times a month. Intuitively, this observation yields much more evidence that Jerry and Alex are similar or aliases since not many South Indian residents call Madurai with such frequency. The goal of the framework is to have a better measurement than frequency for the importance of each call and to re-rank them in order of information content. The model leverages this observation by adding importance for a city to which Jerry calls frequently and by deducting importance if many people in the general population call the same city. Therefore the mutual information model concentrates on the most informative observation and then generates a ranked composite view by adding relative
importance for them. Then the similarity model is used to detect and rank the candidate aliases for each entity in the population. This system thus provides the aliases to be found from the entire malicious environment. It also provides the original role played by that particular person as well as the other kind of the roles played by them.

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.

A similarity of 0 indicates orthogonal vectors whereas a similarity of 1 indicates identical vectors. For two very similar elements, their vectors will be very close and the cosine of their angle will also approach 1.

10. CONCLUSION

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 in malicious environments. The fuzzy set formation and qualitative link based methods have some limitations of not providing the effective alias detection in malicious environments. The fuzzy based analysis may provide only the single user aliases among the neighboring contacts information. They do not provide the suitable way to find the aliases of persons from malicious environments. They also do not provide the original role as well as the other kind of fake roles exhibited by the persons. The above said limitations may be overcome by the information theoretic approach by detecting the aliases from the malicious environments. The information theoretic approach uses mutual information model and similarity model for detecting aliases in malicious environments. The mutual information model concentrates on the messages that have been communicated by the persons and then generates the ranked view of the persons of such observations based on the relative importance. The similarity model helps to find out the other fake persons who are acting as the original person in that environment.

References


Author

M.Rajesh Khanna received the M.E degree in computer science and engineering from B.S.A Crescent Engineering College in 2009. During 2009-2010, he stayed in orbit technologies as software engineer to develop health care automation tool. He is now working in Vel Tech Multi Tech Engineering College as Assistant Professor and IBM TGMC Project Coordinator.