

Detection of Anomalous Data using Data Visualization Techniques

Sumalatha Ramachandran, Sindhuja Vijayaraghavan, Radhika Ramadoss and Sathya Marimuthu
Madras Institute of Technology/Information Technology Department, Anna University, Chennai, India.
Email: {sumalatha.ramachandran, vr.sindhu, radhikaramadoss, sathyamarimuthu}@gmail.com

Abstract— One of the most pressing problems in enforcing security in a network is the identification of suspicious nodes and messages in a network. A node's suspicion factor cannot be measured based on the information that is being sent through the network alone. Anomalous nodes are nodes in a network that send strange, malformed data only within a small instance of time. These data are crucial in cybercrime investigations and other kinds of fraudulence. Though existing methods can determine any misbehavior of nodes, little importance has been given to the field of detection of anomalous nodes – nodes that are forced to send suspicious data by the end user at few instances. As the end user is closely related to the anomalous data, even they are treated as 'nodes' in the study. This paper intends to present an approach based on semantic data visualization in order to identify these nodes. The plan is to provide supportive, user-friendly (preferably natural language) explanations that support and verify the anomalous nature of the predicted nodes. Hence, semantic data is being fed as the input to the proposed problem. The intended solution will also use a mechanism for automatic generation of equivalent user-friendly support-data for the detection, thereby leaving it to the end user to evaluate the veracity of the anomalous nature of the node. As initial stages consideration of the data sources that are given to the input data is made. As per the consideration, the data sources used are ASCII text sent over the LAN.

Index Terms— Anomalous data, data visualization, ontology, Tree pruning, domain ontology, intelligent systems.

I. INTRODUCTION

The growing issue faced today is that of deliberate anomalous data transmission over network. Strange, malformed information must be detected and differentiated from the regular pattern based normal information perceived from normal information perceived from normal users or nodes. There is an association of the end user with the behavioral pattern emphasized on the web and implement the concept of knowledge domain visualization through information and data visualization techniques. Visual analytics is the knowledge based discovery that is applied to the domain considered web camouflage and on this basis, are to conclude the association of crucial data with anomalous data.

The input that is considered, deals with generalized abstract data types that the user furnishes on any of the analyzed web based search engines, and provides the sufficient input core information that is to be analyzed and procured either a safe access or a detected anomaly in the data being transmitted. The process of discovery is based on pattern mining and individual's knowledge

based discovery trend which would predict any kind of suspicious environments. The conclusion is intended to be based on the veracity problem hysteresis that normal users associate with other normal users and that a webpage displaying a majority normal behavior will tend to continue to furnish genuine information and that malformed data is always an exception and will also inherently associate with each other evidently or in successive patterns. However, this alone cannot be sufficient to convince the end user of the back end complexities involved in the information he wishes to retrieve. Thereby, in order to produce, a reliable and more accurate detection mechanism, one such that has been significantly unclear in the previously research related papers, a study of the compatible evaluation procedures and through intelligent reasoning and knowledge domain visualization discovery, is present to the end user as an acceptable explanation. The output will be anomalous nodes detected and presented based on any one of the visualization techniques to the user in a security threatened network.

II. EXISTING WORK

Antisocial elements tend to use the internet as their weapon to deliberately send dangerous data[6], which could be used to the investigator's advantage if handled carefully[8]. Careful analysis show that though repetition of data patterns is carefully being avoided for crucial data[1], isolation of crucial data is still possible. This is done by applying knowledge engineering tools and techniques available in hand or customized tools and techniques[4][3]. On those grounds, the intension is to scrutinize the challenge of obtaining the information required to articulate the rules required to define the analogous nature of a node which is yet to be solved[7].

A widespread challenge faced by the Internet is to provide user-friendly, human-interpretable responses to user queries. The knowledge engineering domain offers a wide scope for analysis and a procedural solution to both parts of the above defined problem. Neurological studies and experiments in Science (neuroscience research articles) involve data visualization. Question-answer pairs for student assistance (online student forums). These domains concentrate much on the names, places, objectives, etc., that signal psychological states of persons of interest. Grüniger and Fox's methodology, which includes neither the processes, nor activities and techniques for performing such activities and neither

specified coherently. Furthermore, although studies have been conducted with this methodology, and there are also applications that use ontologies, the domain is confined to psychological states of user.

Biography Builder: Biographical information of people and Medical informatics (DNA pattern matching) on the other hand allows interoperability between systems. Domain ontologies built using SENSUS approach share the same high level concepts (or skeleton). So, systems that use such ontologies will share a common structure of the world, and it would be easier for them to communicate because they share the same underlying structure. The proposed work is intended to be a generalized version of anomalous data detection over a given network and not on a specific domain.

III. THE PROPOSED SOLUTION

The data that is being transferred through network is anomalous in nature. So its anomaly cannot be determined by periodic monitoring of the same node. Instead, an unsupervised framework that can identify truly novel results is being used here. The approach described here has the following advantages:

- Providing results that are only semi-tightly coupled with past results to prevent the biasing of the results with only existing ones
- To adapt to a new domain if future demands do.
- Validation facts and data apart from the end percentage at the end- user's disposal for scrutiny.

Keeping these factors in mind, an architecture as shown in Figure 2, has been designed to overcome the problem. The first phase is the framing of the user query based on the user statement. The user statement sets the constraints for the records to be dealt with. Once the SQL query is expanded, patterns of interest – the exceptional anomalous data is detected through tree pruning. To improve the time efficiency without trading off space, the records are hashed and the keys instead of the actual records are pruned. However, apart from the

regular patterns followed, there is a need to provide extensional back up information otherwise stated as support information in order to conclude on anomalous activity. To substantiate evidence of reasoning and conclusive remarks based on these support information. Such data support is retrieved from the database which involves domain ontological phase analysis. The records retrieve here are based on retrieving similar patterns of the anomalous data just recorded, in contrast to the first phase that requires retrieving deliberately diverse data. The final phase of the implementation involves and deals with the data display of the result set of the XML query in a user friendly format at the front end, in which domain ontology plays a crucial role.

IV. IMPLEMENTATION ISSUES

A. Tree pruning approach implementation

Over time, the data that has been analyzed and stored in the database may become insignificant or stale. This will affect the accuracy of the system very badly. So, an automated approach that actually retains the data, yet considers them only if necessary, becomes essential. But implementing the same would actually be an overhead in ontology as the associated time complexity is of non polynomial order. Hence, another plausible approach to the problem is being proposed here. Accordingly, the information chunks in the domain ontology are used as plain data sets itself. Now, the problem in terms of the data sets reduces to 'overfitting data'. And the obvious solution for the same would be to 'prune the tree'. The algorithm can be customized to fit the application domain of the user.

B. Conceptual modeling of the ontology:

Using the Semantic Web's capabilities to monitor anomaly activity and model networks requires an ontology crafted to express complicated and sometimes contradictory information in an accessible and intuitive way. For example, Terrorists and terrorist organizations engage in a wide range of activities that reflect the variety

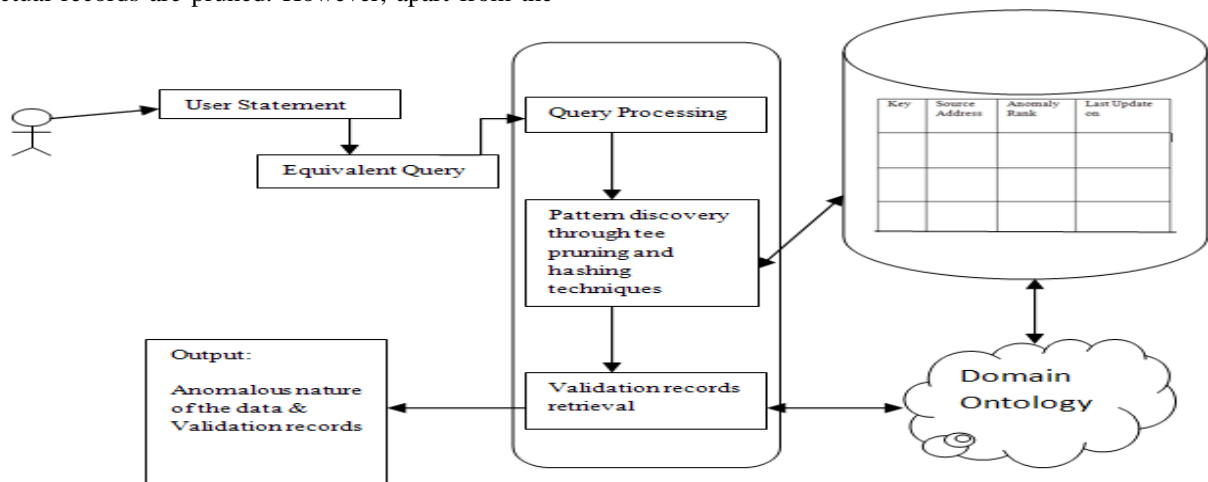


Figure 1. Architecture of the Anomaly Detection scheme

of human endeavor. An ontology that reflects this must maintain a careful balance between being sufficiently comprehensive while not being so specific that it is tailored to particular situations and cannot accommodate others. The complexity and diversity of situations is such that at some points paragraphs of text are necessary – but this is a last resort because doing so defeats the point of and does not use the complete capabilities of the Semantic Web. At the same time, the ontology must be capable of growing and changing to reflect new aspects of a changing phenomenon carried out by adaptable actors. Creating an ontology that is a useful tool for publicly accessible research purposes is a unique challenge.

C. Designing of the ontology:

Reflecting events and sequences of events in time is a central challenge to building successful misinformation ontology. A class of events to describe the odd (anomalous nature) in which properties included location, date, casualties etc. based on the existing databases was created, but, it was insufficiently specific. This solution proved useful in many other places as well. The problem of describing minute to minute events was solved by creating a “Moment” class to describe specific moments in time. “Contacts” class describes meetings between operatives. Subclasses were created to describe different types of contacts such as meetings, telephone and internet communications, and information delivered by messenger. Important properties, besides location and participants, reflect what was transacted in the contact. Special subclasses had to be created to describe financial transfers that were done through financial institutions. Another important type of event that is important to tracking misinformation is wordlist and sentence structure. A class for the same was created.

One property consistently shared by every event is “Involved User” which has the range of “Person.” Events described, be they meetings between people or terror attacks, described specific action. Meetings in which specific information and items were transferred are important to understanding misinformation but equally important are the ongoing relationships that cement the small cliques that become terrorist cells. To reflect this, a different class for relationships was created. This class allows two or more involved members.

There are a number of existing issues when we are working on the system. A few of them would be the issues related to running on a Linux based operating system. Another prevalent issue would be that it is a client-server based architecture in the sense that the user (client) is reported with the anomalous nature of the data from the server.

D. Insertion of the new node into the ontology:

Searching the ontology and insertion of new nodes at the right position will be the most decisive phase of the entire system. The pseudocode for this phase is described in Fig. 2. Though so many searching mechanisms are available, the Breath first searching mechanism is

1. Pick one source node, S, from the set of nodes with anomalous information.
2. Traverse through the list of its neighbors.
3. Insert the source node, S, into the queue.
4. WHILE the list is not empty
 - a. IF the node has the required anomalous information)
 - i. THEN Output the trail that led to the destination and clear the queue.
 - ii. ELSE Add the node to the queue and retrieve the next node.
5. IF the node was found, THEN Halt.
6. ELSE IF Other nodes remain unsearched,
 - a. THEN Search the next node in the queue
 - b. ELSE Insert the new node, whose position is determined from the inferences obtained from the rules designed in the ontology.

Figure 2. Node Insertion Algorithm

preferred over the rest. The reason is pretty straightforward. The anomaly of the information node in the ontology is inversely proportional to its distance from another anomalous information node. In terms of graphs, nodes with anomalous information are:

- Adjacent to each other or
- At the least possible distance from each other.

This enables us to cluster the anomalous information together. Hence Breath First Searching process will reduce the time elapsed at searching the required information greatly.

The source node from which the search must be triggered will be chosen from a set of nodes with anomalous information that is already prevalent in the database. Once inserted, Anomaly Factor A_f is calculated as follows:

$$A_f = \frac{N - A_r}{N} \quad (1)$$

Where

A_f = Anomaly Factor

N = Total number of records in the database

A_r = Rank of the related anomalous content in the database

From A_f , the anomaly percentage is calculated by simply multiplying it with 100.

V. RESULT ANALYSIS

Two criteria that are of major influence while the framework is being tested will be speed and accuracy of the anomalous nature of the user input. To scrutinize these two criteria exhaustively, an input set consisting of a gamut of anomaly levels ranging from 0 to 100% is a good choice of input. For these input sets, the anomaly factor, that was calculated using (1) is the measurement of the required output.

The Anomaly factor for a sample of 200 queries has been simulated in Fig. 3.

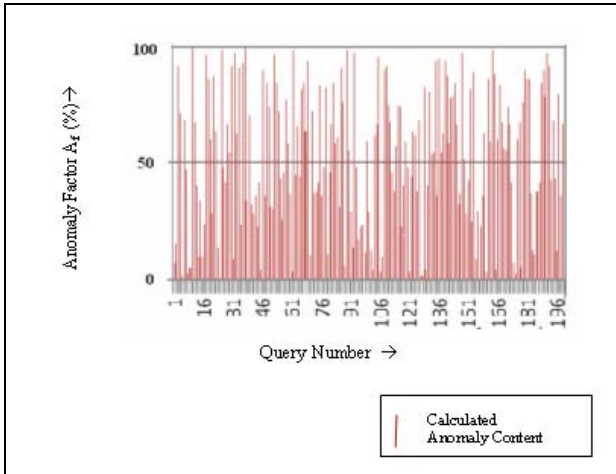


Figure 3. Anomaly content in the user input

When compared with existing systems, the accuracy levels of the proposed system, the standard deviation of the ideal accuracy factor is lesser than existing systems, which implies better efficiency of the system. The results suggest that it is useful to consider information provided by nodes and links that are multiple steps away from the source node in the proposed system. Anomaly factor, as stated above was calculated based on a cluster of data-nodes. It was observed that, for various central nodes, the magnitude of the anomaly factor remained stable. This balance must be maintained, in order to ensure the proper functioning of the system.

Furthermore, the execution speed must also be realistic. The execution time for an exhaustive range of queries has provided the execution time is shown in Fig. 4.

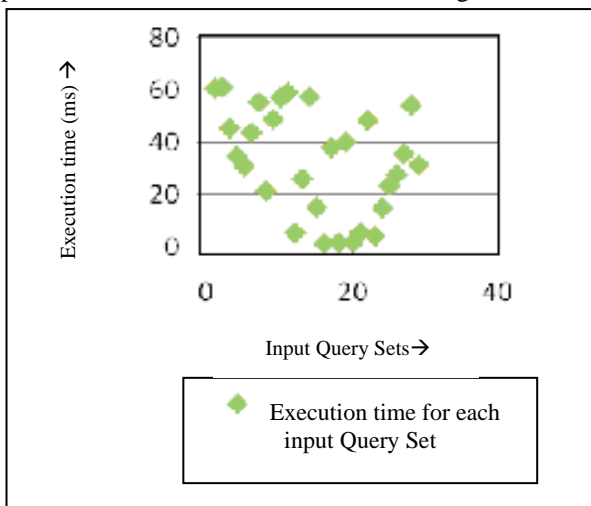


Figure 4. Calculated Execution Time for a set of given inputs

The execution time is defined as the total time elapsed from the time of the user providing the input to the time of providing the output. The variance of the execution times of the query set is observed to be an optimal minimum. This implies that the system performs efficiently over a varied range of queries.

VI. CONCLUSION AND FUTURE WORK

Thus, several data mining functions, each with a rich set of methods, has been applied on a database of anomalous data. It is tightly coupled with the same and further analyzes the classification and sequential patterns among the data. Once this is obtained Intelligent Query answering mechanism is applied to provide the sufficient data to support the anomalous nature of the data that has been predicted.

REFERENCES

- [1] J. Wang, F. Wang, and D. Zeng, "Rule Exception Learning-Based Class Specification and Labeling in Intelligence and Security Analysis," Proc. Workshop Intelligence and Security Informatics (WISI '06), vol. 3917, 2006, pp. 181-182.
- [2] G. A. Fink and C. North, "Root polar layout of internet address data for security administration.", In Proc. IEEE Workshop on Visualization for Computer Security (VizSEC), October 2005, pp 55-64.
- [3] S. Lin, "Generating Natural Language Descriptions for Paths in the Semantic Network," final project report, Dept. of Linguistics, Univ. of Southern California, 2006.
- [4] Marie B. Synnestvedt, "CiteSpace II : Visualization and Knowledge Discovery in bibliographic Databases", AMIA Symposium Proceedings, pp 724-728,2005.
- [5] Markus Jakobsson, Sid Stamm, "Web Camouflage Protecting your Clients from Browser Sniffing Attacks", Published by IEEE Computer Society pp 16-24,2007
- [6] H. Chen and J. Xu, "Intelligence and Security Informatics for National Security: A Knowledge Discovery Perspective," Ann. Rev. of Information Science and Technology, vol. 40, pp. 229-289, 2006.
- [7] Segev Wasserkrug," Inference of Security Hazards from Event Composition Based on Incomplete or Uncertain Information", IEEE transactions on knowledge and data engineering, vol. 20, no. 8, pp 1111-1114, August 2008
- [8] Chaomei Chen, "Top 10 Unsolved Information Visualization problems", IEEE publication July/August 2005, pp 12-16,2005G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," *Phil. Trans. Roy. Soc. London*, vol. A247, pp. 529-551, April 1955.