Cyberspace Situational Awareness Model based on Spatial Traffic Clustering

Ying Zhuo, Qiang Zhang, Zhenghu Gong
School of Computer Science
National University of Defense Technology
Changsha, Hunan, China
zhuoying@nudt.edu.cn

Abstract—As the certain development direction of network management, fusion based Cyberspace Situational Awareness (CSA) provides the comprehensive macroscopic view of network operating status, enhances the management and control on network, and offers the decision-support for the high-level commander. Today, most research concentrates on security situation. On the basis of traditional network management, we integrated its two function units: traffic analysis and topology discovery as the information source of CSA and established Cyberspace Situational Awareness Model based on Spatial Traffic Clustering – STCSA. We then proposed a top-down subspace clustering algorithm for network data stream oriented situation pattern partition. STCSA can integrate the existing unit network management, mine multi-source multi-feature data by unsupervised machine learning, provide whole-network situation view, and its speed is great. At the same time, STCSA is easy to detect various known and unknown network anomaly and extract anomaly features. The experiment results on real datasets demonstrate the efficiency, effectiveness and scalability. At present, STCSA prototype has been deployed in network management system.

Keywords—Cyberspace Situational Awareness; Spatial Traffic Analysis; data stream; clustering

I. INTRODUCTION

With the rapid expansion of the information network scale, the complexity and uncertainty of system also increase. Each function unit of traditional network management works in an independent state. Due to the absence of effective information extraction and information fusion mechanism, network management system is unable to establish the contact between network resources, and the representation ability of overall information is poor. The massive network management information could not strengthen management, but instead has increased the burden on administrators. Modern network management should be able to provide diverse, individual managements; supplying the detailed information about the managed objects, understanding the operating status of the whole-network, providing the service according to the commanders’ demand, etc. Therefore, Fusion based Cyberspace Situational Awareness will certainly become the development direction of network management in the future [1,2].

Cyberspace Situation is that the current state and trend of the whole-network which is composed of operating status of various network equipments, network actions, user behaviors and other situation factors. It is worth noting that situation is a state, a trend, an overall concept, and no single state can be called situation. The so-called situation factor is an element which can bring changes in network situation, and the set is a subset of monitor index set. Cyberspace Situational Awareness (CSA) refers to the acquirement, comprehension, assessment, visualization of situation factors and forecast the trend of future development in the large-scale network environments. In short, CSA is a mapping from the situation factor set $R$ to the situation space $\theta$, $f: R \rightarrow \theta$.

CSA’s goal is to introduce situation awareness techniques into the network management field, organize each kind of information efficiently in the rapid-change complex environment, synthesize existing indices which represent network partial features, integrate each function of traditional network management unit working independently and provide the comprehensive macroscopic view of network operating status, so as to enhance the network comprehension ability of administrators and provide the decision support for the high-level commander. CSA research includes three main aspects: model, knowledge representation and assessment method. The majority research focuses on the security situation, while a small amount touches upon transmission, information superiority, survivability, system evaluation and so on. The related works as well as the existing problems have been discussed detailed in another paper [3].

2010 International Conference on Intelligent Network and Computing (ICINC 2010)

V1-428
II. STCSA

To address the above problems, combining the spatial traffic analysis and clustering technique, this paper proposed a CSA model – STCSA. First CSA architecture is presented; main idea follows; next STCSA modeling and awareness process are introduced in detail.

A. CSA architecture

Inheriting JDL data fusion model as well as Endsley situation awareness model and combining the concrete network application, CSA architecture is proposed. CSA architecture defines the function of CSA as well as the function division and organization, as shown in Fig. 1: CSA architecture consists of six functional modules; its main function is divided into three levels, bottom-up in turn is object level, situation level and decision level; paralleling with the main function is system management, human computer interaction, database/knowledge base and other system support functions.

CSA architecture highlights the important role of feedback, in which CSA takes up the position of core in data fusion and plays the role of connecting link between the preceding and the following. Comparing with other classical models, the differences lie in: (1) integrates the traditional network management unit, and reflects features of spatial traffic analysis; (2) highlights the importance of knowledge discovery and data mining, by which establishes the scientific and objective situation assessment and forecast methods and separates the modeling process from the situation awareness process; (3) emphasizes the nature of dynamic cycle, and shows a "assessment-forecast-visualization" cyclic process.

B. Main idea

The so-called spatial traffic analysis [4], compared with temporal traffic analysis, not only analyzes the temporal behavior of a single link, but also seeks a new traffic analysis method which can consider both the link traffic features and the network connection manner synthetically, analyze traffic pattern across multiple links or whole network, and support a whole-network view. Spatial traffic analysis has realized the interaction between the network topology and traffic features, and established a highly summarized and complete network situation view.

The primary challenge of spatial traffic analysis lies in the lack of prior knowledge. The measure indices of CSA are numerous: the effect of a single index on situation is different, and there are complex relationship and interaction between indices; in addition, due to the scarcity of related research, no achievement of situation pattern partition can be used for reference. If dividing the situation pattern by domain expert, the partition will be subjective inevitably; moreover it is very difficult to determine the contribution of each index to situation accurately and fairly.

As the main method of data mining, clustering is a kind of unsupervised machine learning method with capability of knowledge acquisition and law discovery. Clustering analysis can achieve the situation pattern partition automatically according to massive network data and extract typical situation features without any prior knowledge, so it is scientific and objective.

STCSA makes up the insufficiency of simple index or single link. On the one hand, STCSA can fuse many kinds of factors that affect the network situation, detect a variety of known and unknown anomalies and avoid the limitation of hierarchical structure and weight analysis; on the other hand, STCSA can consider the effect of topology synthetically, reveal the complex relationships between NEs (node or link is called as network element, NE for short), and achieve whole-network situation awareness, so it reflects the situation features of integrity and macroscopy.

C. Modeling

Combining the features of spatial traffic data, CSA model is established as follows: STCSA(Topo,Traffic), in which Topo represents the topology information and Traffic represents the traffic information.

Topological information, includes NEs and the connection relations between each other, expressed as Topo(ID,Time,Node,Link,ψ). ID represents the identity of one topology discovery, it is unique. Time is the time of the topology discovery. Node is the node set, expressed as (IDn,C,W,Dsc): IDn is the unique identity of node; C is the processing capacity of node; W is the importance weight of node which is decided by both network topology structure and node processing capacity; Dsc is an extensible item to describe the related information of node. Link is the link set, similarly with the node set, expressed as (IDn,C,W,Dsc) uniformly; the different lies in C is link capacity. ψ is connection relations, ψ ⊆ Node×Node.
Traffic information, includes all the information and knowledge related to CSA obtaining through the traffic analysis and mining, expressed as Traffic(ID, Time, Trace, IS, SF, SP, AR). ID represents the identity of NE. Time is the time of traffic appearance. Trace is the original packet information from traffic collection. IS, Index System, is a characteristic set which is extracted by traffic analysis and reflects the network situation; \( IS = \{ a_1, a_2, ..., a_n \} \), \( a_i \) is a traffic feature. SF, Situation Factor, is a set of important features, obtained through feature selection, which can cause changes of network situation; \( SF = \{ f_1, f_2, ..., f_d \} \), \( f_i \) is a situation factor, \( f_i \in IS \); SF is a subset of Index System, \( SF \subset IS \). SP, Situation Pattern, according to the situation factor value, divides the situation pattern through the clustering analysis; each situation pattern likes: \( sp = (f_i, v_i), (f_2, v_2), ..., (f_d, v_d) \). \( v_i \) is the value of situation factor \( f_i \). AR, Assessment Rule, determines the situation value of each pattern and designs situation assessment rules on the basis of situation pattern partition.

D. Awareness process

According to STCSA, the basic step of situation awareness is described as follows:

**ALGORITHM STCluster**

<table>
<thead>
<tr>
<th>step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>step 1</td>
<td>Gather topology discovery and traffic collection information from level 1 fusion;</td>
</tr>
<tr>
<td>step 2</td>
<td>Topo modeling, and then analyze the NE importance ( W );</td>
</tr>
<tr>
<td>step 3</td>
<td>Traffic modeling:</td>
</tr>
<tr>
<td>3.1</td>
<td>Establish IS by extracting flow features from original trace data;</td>
</tr>
<tr>
<td>3.2</td>
<td>Discretization and standardization of flow features;</td>
</tr>
<tr>
<td>3.3</td>
<td>Flow feature selection, and establish the set of SF;</td>
</tr>
<tr>
<td>3.4</td>
<td>Mine SP in SF set through clustering analysis;</td>
</tr>
<tr>
<td>3.5</td>
<td>Design AR of SP based on Rough Set;</td>
</tr>
<tr>
<td>3.6</td>
<td>Consider the situation development trend synthetically, and establish the timeliness adjustment policy of situation level;</td>
</tr>
<tr>
<td>3.7</td>
<td>Human-computer interaction, refine AR and adjust model; (optional)</td>
</tr>
<tr>
<td>step 4</td>
<td>Complete the situation assessment and forecast according to STCSA, when new data arrive:</td>
</tr>
<tr>
<td>4.1</td>
<td>Decide NE situation based on Traffic model;</td>
</tr>
<tr>
<td>4.2</td>
<td>Decide NE importance based on Topo model;</td>
</tr>
<tr>
<td>4.3</td>
<td>Calculate the whole-network situation;</td>
</tr>
<tr>
<td>4.4</td>
<td>Situation forecast and visualization;</td>
</tr>
<tr>
<td>step 5</td>
<td>Judge the validation of STCSA:</td>
</tr>
<tr>
<td></td>
<td>If no longer suitable, return to step 1, update STCSA incremental;</td>
</tr>
<tr>
<td>step 6</td>
<td>Repeat step 4.</td>
</tr>
</tbody>
</table>

III. SITUATION PATTERN PARTITION

Clustering divides data into different clusters based on the data character, and its objective is that between the objects belonging to the same cluster the similarity is high, while the difference of the objects in different clusters is distinct. There are many researches on clustering, and the clustering technique is divided into five categories: partitioning method, hierarchical method, density-based method, grid-based method and model-based method.

The features and special requirements of CSA place higher challenges on clustering algorithm. A detailed analysis as follows: (1) As a typical data stream, network data has many features, such as large volume, potential infinity, one-pass, ordinal access, uncertainty arrival rate and so on. (2) Network data fuses multi-source multi-feature data, so the index system is huge and the dimensionality is numerous. High-dimensional data brings some problems, such as the curse of dimensionality, data sparseness, heterogeneous attributes data, difficulty of distance calculation and so on. (3) The dynamic characteristic of network data stream causes concept drift, which requests clustering algorithm to detect concept change timely, update clustering model dynamically and reflect the clustering evolution process. (4) In many cases, as one step of network data analysis, clustering is not the ultimate goal but only a means, therefore the algorithm is requested to complete quickly under real-time online environment and meet the requirement of response time.

According to the above analysis, this paper proposes a clustering algorithm for situation pattern partition -- STCluster. STCluster is described as follows:

**ALGORITHM STCluster**

| step 1 | Partition the clustering space into grids. |
| step 2 | One-pass data stream, statistic grid density information. |
| step 3 | Merge the connected dense grids to build a cluster in fullspace. |
| step 4 | Output the clusters satisfied the cluster threshold density \( r_c \). |
| step 5 | Combine any two clusters with density \( \in [r_d, r_c) \) to build maximum dimensionality candidate projective cluster. |
| step 6 | Statistical the density information of candidate projective clusters through searching other clusters unsatisfied \( r_c \). |
| step 7 | Choose the one of best quality from candidate projective clusters. |
| step 8 | If satisfied \( r_c \), output the projective cluster. |
| step 9 | Repeat step 6,7,8, until no more projection cluster satisfied \( r_c \) or terminal condition. |
| step 10 | If reach update interval, update clustering result incremental. |
STCluster includes three steps: full-dimensional space clustering (FULLSPACE), top-down subspace clustering (SUBSPACE) and incremental update clustering (INCREMENT). STCluster builds high-dimensional grid space through data discretization and encoding (step 1); counts the number of data points in grid as the grid density (step 2); in FULLSPACE, taking the density as the criterion of clustering analysis, merges the connected grids to build a cluster (step 3); in SUBSPACE, searches dense projection clusters in the clusters unsatisfied cluster density threshold τ employing a top-down policy: first, reduces the dimensionality of clustering space to connect the clusters unsatisfied τi, thereby, builds all subspace which contains project cluster possibly (step 5), second, searches the optimal projective clusters [5] one by one in the candidate subspace (step 6-8) until find all of projection clusters satisfied τc; with the arrival of new data, maintains clustering results dynamically by means of incremental update technique (step 10). INCREMENT algorithm has been introduced in another paper [6].

Grid-based method reduces the clustering space of large-scale data stream effectively and with good scalability. Density-based method can discover clusters of arbitrary shape and it is insensitive to the data input order as well as the noise. The combination of density criterion and grid partition avoids the calculation of distance, so it could deal with the heterogeneous attributes data. Subspace approach employs the original dimensions rather than new dimensions to build subspace, so it is simple and easy to understand, and effectively addresses the sparseness of high-dimensional data. Takes full advantage of the distribution character of network data, the top-down policy satisfies the requirement of one-pass and realizes the search of projection clusters in different subspaces with different dimensions. Incremental update reflects the evolution process of data stream and simultaneously reduces the algorithm complexity effectively to satisfy the response time requirement of online clustering.

STCluster is an efficient subspace clustering algorithm for high-dimensional data stream. Under the assumption that, there are total n data points distributing among g grids in practice, so the time complexity of FULLSPACE is O(n+g²); if the number of clusters is c, the number of clusters unsatisfied τi is c approximatively (slightly less than c), candidate subspace is s, projection clusters is p, the time complexity of SUBSPACE is O(c²+pcs). The algorithm complexity may be reduced further by limiting the density τc of dense grids participating in clustering as well as the density τc of candidate clusters building subspace. If there are d situation factors and w windows in INCREMENT, the space complexity of STCluster is O(c²+pcs).

IV. EXPERIMENT RESULTS AND ANALYSIS

In this section, we do some experiments to test STCSA. Experimental platform is configured as follows: AMD Athlon Dual Core 4200 + GHz/2GB/Windows XP. The code is realized by ActivePerl (5.8.8). The experimental data, Abilene network flow data, are provided by the U.S. National Laboratory for Applied Network Research (NLANR) Passive Measurement and Analysis (PMA) Working Group, who established multiple measurement points and collected Internet data passively in the HPC network [7]. Abilene [8] is US education and research network, as shown in Fig. 2, its core topology consists of 11 nodes and 14 bidirectional links. NLANR data take samples from Abilene packet information, 8 times one day and 90 seconds each time. And in July to September 2001, January 2003 and January 2004, records the burst of Code Red Worm, Slammer worm and W32 Mydoom respectively.

STCSA chooses the number of packets, bandwidth, estimated delay, packet length distribution, protocol distribution, application distribution, new connections, active connections, connection volume, average connection duration, TCP flag distribution and other indices to assess network situation of Abilene. Fig. 3 shows a one-week situation assessment result of one link from 2001/07/10 to 16. It can be seen that network situation represents obvious periodicity taking 24 hours as the unit in normal state.

At the same time, the two-week situation curve shown in Fig. 4 reflects the joint effects of situation factors on situation trend, such as the number of IP packet, the number of ACK flag, destination port distribution entropy, and so on.

We use STCSA to assess the situation when the burst of Code Red Worm, Slammer worm and W32 Mydoom. As shown in Fig. 5, when the network anomalies occur, the situation curve changes obviously and the situation value is larger relatively.
Besides detecting a variety of known and unknown network anomaly, STCSA can extract anomaly features from different points of view. Corresponding to the situation trend in Fig. 5, Fig. 6, 7, 8 show the change of the number of packet, distribution entropy, TCP flag respectively. It can be seen that the behaviors of different factors in different anomalies are widely divergent. For example, the number of TCP packet increases sharply at the burst of Code Red Worm, while at the burst of Slammer worm the number of UDP packet is exceptionally prominent; similarly, RST flag reflects the burst of Code Red Worm, while ACK flag isn’t exceptional; distribution entropy works well at the burst of both Code Red Worm and Slammer worm; ACK flags, ICMP-reply and the distribution entropy of source port behave unusually at the burst of W32 Mydoom.

In order to reveal the detail of anomaly features comprehensively and present the behaviors of different factors in different anomalies, STCSA draws a features chart of situation factors using radar plot, as shown in Fig. 9. In radar plot, each kind of anomaly is clear at a glance. Not only shows many kinds of situation factors simultaneously, but also facilitates the comparison between anomalies each other as well as normal state. Obviously, STCSA is easy to find network exceptional behavior, and extracts anomaly features in time.

V. CONCLUSION

In this paper, oriented the requirement of network management and combined the advanced technique of situation awareness, we established Cyberspace Situational Awareness model based on Spatial Traffic Clustering, and proposed a top-down subspace clustering algorithm for high-dimensional data stream which can mine the network traffic data for situation pattern information. The experiment results on real datasets NLANR demonstrated the efficiency, effectiveness and scalability of STCSA. STCSA could fuse the diverse unit network management information, and grasp the network situation macroscopically, which reflected the application value and practical significance of prototype system. This paper is an attempt on CSA, and the related research still has huge space for development. In the future work, we will research the key issues of STCSA as well as situation forecast method and pay attention to the latest research simultaneously, so as to further improve the accuracy and efficiency of CSA.

ACKNOWLEDGMENT

This work was supported by a grant from the National Fundamental Research and Development Program of China (973 Program) (2009CB320503) and the National High Technology Research and Development Program of China (863 Program) (2008AA01A325).

REFERENCES