A Framework for Cyber Surveillance of Unlawful Activities for Critical Infrastructure Using Computational Grids

Václav Snášel[1], Ajith Abraham[2], Khalid Saeed[3] and Hameed Al-Qaheri[4]

[1]*VSB-Technical University of Ostrava, Czech Republic*
[2]*Machine Intelligence Research Labs –MIR Labs, WA, USA*
[3]AGH University of Science and Technology, Cracow, Poland
[4]Kuwait University, Kuwait
vaclav.snasel@vsb.cz, ajith.abraham@ieee.org, saeed@agh.edu.pl alqaheri@cba.edu.kw

Abstract

This paper highlights a framework for cyber surveillance of unlawful activities for critical infrastructure protection. The framework uses a computational grid based environment, which is capable of distributed data mining and real time surveillance.

1. Introduction

News reports reveal the usage of cyber infrastructure for framing up unlawful activities against the Government, critical infrastructure etc. Some of these activities are possibly unlawful, e.g. hacking, spamming or a preparation for a crime action against some infrastructure. Although web pages include static and dynamic pages, blogs, discussions and so on, they are not alone the source of information. The important information is included in news, online chats, messaging portals and email conferences as well. Moreover, if we can consider biometrical identification of suspicious criminals (or with some background), we can scan images, videos, keystrokes and mouse movements and so on.

An important issue is that we cannot identify a possible unlawful activity by simply searching for an object. For example, if a term **flood** is identified in a web page, what does it mean? Do we retrieve an attribute of a prepared spam or flood attack? Do we scan an article to a flood on a river? Obviously, such simple scanning is not good enough! It seems that more appropriate way is to build social networks. We can build this network according to the information retrieved from mail conferences, tracing online chats, real-time video of an airport hall and so on. A person's action does not seem as a possible dangerous event at the first look. With the help of social networks, it might be possible to detect a possible unlawful activity according to the relationships between different persons. It is important to build the network according to various kinds of heterogeneous data: blogs, web pages, hyper link structures, emails, images, video, keystrokes, fingerprints, facial images, mouse movements and so on. Of course, volume of such data is very huge; therefore the grid computing is a good tool for managing and mining such data.

Web surveillance has become a hot research topic, which combines two of the prominent research areas comprising of data mining and the World Wide Web (WWW). Web mining has become very critical for developing key intelligence to monitor social networks, cyber terrorism related activities, flooding of abusive contents, Web management, security, business and support services, personalization, network traffic flow analysis and so on. Generally, the Web analysis relies on three general sets of information given:

- past usage patterns
- degree of shared content and
- inter-memory associative link structures

Distributed Web mining is at its infancy and most of existing distributed data mining systems are experimental designs and therefore focused only on a particular data-mining problem. Since this data is geographically distributed (either by the fact that a single database is physically distributed, or that pertinent data is stored in multiple distributed repositories) making the data mining process substantially more tedious or sometimes even impossible. Formulation of meta-knowledge from distributed web data would be an interesting task.

Some of our previous research studies on Web mining have shown the importance of hybrid intelligent systems to obtain optimal knowledge from the data. An important disadvantage of such hybrid systems is the computational complexity of the underlying algorithms. To address the growing need for computational power the interest starts to shift from stand-alone clusters to computational grids. In

addition, the distributed nature of the grids matches the distributed nature of the network traffic data. Due to the sensitivity of the data used, it is important for the computational environment and data to be highly secured. To minimize computational costs, it is important to seek solutions that do not rely directly on cryptography. A secured grid also required for protection from intruders.

2. Proposed Framework

2.1. Reducing and Indexing Voluminous Data

In the literature, we can find a huge number of techniques and strategies that can be utilized to process high-volumes of data in support of cyber terrorism. Data reduction is a critical problem for cyber terrorism; there are large collections of documents that must be analyzed and processed, raising issues related to performance, lossless reduction, polysemy (i.e., the meaning of individual words being influenced by their surrounding words), and synonymy (i.e., the possibility of the same term being described indifferent ways). Our main objective in this research is to investigate data reduction strategies, ranging from data clustering to learning to latent semantic indexing, tensor reduction for web related heterogenous data.

We apply a non-negative matrix and tensor factorization approach for the extraction and detection of concepts or topics from Web. Web mining has become very critical for developing key intelligence to monitor social networks, cyber terrorism related activities, flooding of abusive contents, Web management, security, business and support services, personalization and network traffic flow analysis. Traditional data mining algorithms are enhanced to find specific relationships or patterns in the particular data source. The dimensionality of the data is typically high and data used in the mining is usually non-negative. Non-negative matrix and tensor factorization (NMF) is a recently developed technique for finding parts-based, linear representations of non-negative data. Nonnegative matrix and tensor factorization produce nonnegative basis vectors, which make possible the concept of a parts-based representation. Techniques like *Singular Value Decomposition* (SVD), *Semi-Discrete Decomposition* (SDD), and *High Order SVD* (HOSVD) also generate basis vectors – various additive and subtractive combinations of which can be used to reconstruct the original space. But the basis vectors for these methods contain negative values and cannot be directly related to the original space to derive meaningful interpretations.

2.2. Distributed Data Mining

We require managing and devising a scheme for updating the huge Web databases efficiently. In a distributed environment the data gets updated in a periodic fashion and by managing the synchronization between the master and a replica (periodic updated version) it is possible to manage the data effectively.

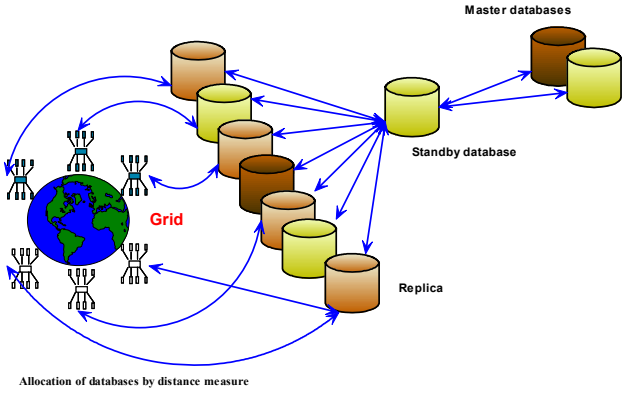

Figure 1. Database management architecture

Traditional approaches to knowledge discovery for Web data repositories concentrated on the process of collecting and synchronizing the data into a single location using some kind of client/server architecture and data warehousing, and then analyzing the data using fast parallel computers along with efficient knowledge discovery algorithms. Some of the architectures are depicted in Figures 2 (a), (b) and (c) regarding discovering knowledge from distributed databases. $Algorithm_1$ ...$Algorithm_n$ represents different data mining algorithms and $Descriptor_1$ ....$Descriptior_n$ refers to the individual outputs of the data-mining tasks.

We propose to analyze the available full data rather than taking samples at regular intervals or random samples. A particular data-mining algorithm will analyze each data and the individual descriptors are finally combined to formulate meta-knowledge as shown in Figure 2(a). In Figure 2(b), the knowledge from the individual descriptors are passed to the neighboring data-mining algorithm and finally the meta-knowledge is obtained from the n^{th} descriptor. Another alternative is to cooperate between the different data-mining tasks and combine the individual descriptor knowledge or as illustrated in Figure 2 (c).

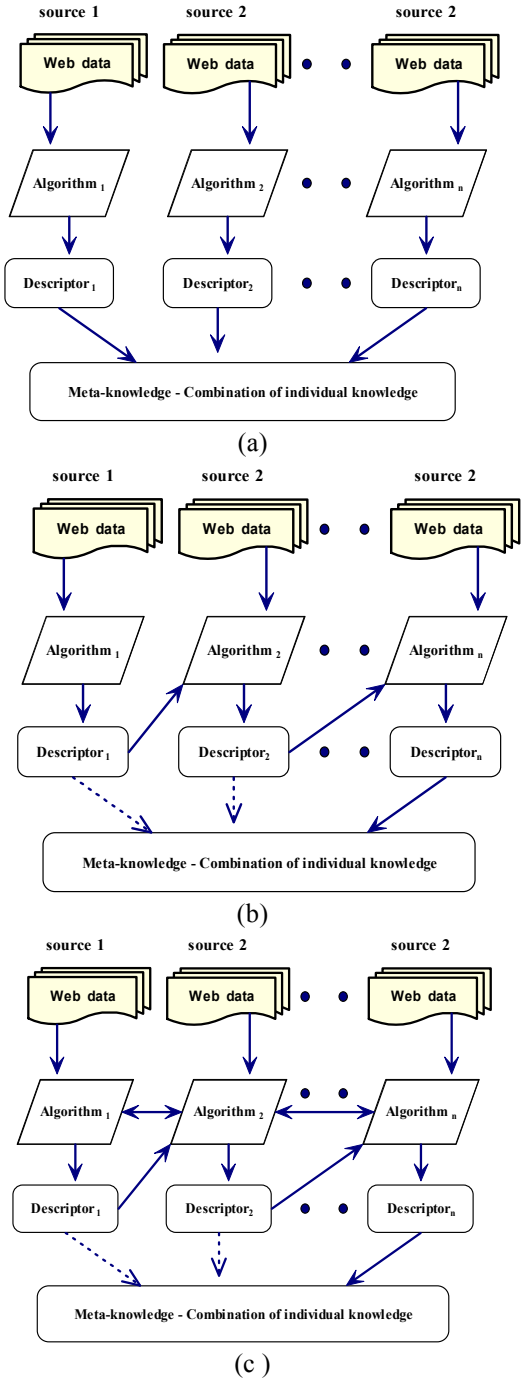

Figure 2. Distributed mining architectures

2.3. Web Traffic Flow Monitoring

It has been found that new aspects of Internet behaviour emerge that are either unknown or poorly understood can be revealed sometimes by accurate traffic flow monitoring. Network monitoring and measurement would also help to spot abnormal traffic flow within a suspected social network. Traffic flow monitoring also helps us to analyze network infrastracture trends and user behavior and to improve the security of the cyber-infrastracture. LOBSTER is a successful pilot European passive Internet traffic monitoring project that is capable of providing early warning alerts for traffic anomaly, security incidents, and providing accurate and meaningful measurements of performance [1]. We plan to deploy passive monitoring sensors at speeds starting from 2.5 Gbps and possibly up to 10 Gbps based on LOBSTER technology for advanced network traffic flow monitoring. Based on the alerts received from the passive monitoring sensors, more efforts will be focused on to investigate the reason for the anomaly traffic flow in the network (social network).

3. Surveillance

Automatic surveillance is relatively new area of application for galloping video and sound technologies in communication. Monitoring the environment for security or general information has already proven its importance and possibilities. It involves gathering and analyzing data from various sources. Such sources may supply data of various types, may be located in distant spots and operated under different conditions. Automatic processing of such diverse and large records requires cooperation of many specialized components, often distributed logically or/and geographically. Computing Grid architectures provide excellent platform for this kind of processing. It allows better resource management, scalability and application of more computationally intensive algorithms. Table 1 summarizes the purposes, applications and data used for surveillance. *Applications related to* ***voice/sound*** *data* can serve as a support for the video data, in particular for:

1. **Person identification** – voice of the person may serve as an additional data for the purpose of identification. Assignment of voice to a specific individual may be problematic when the feed is taken in the street.
2. **Detection of certain events** – examples of interesting sounds are: gunshots, explosions, breaking glass, crashes, screams, foot steps, closing doors. This kind of activity could be independent of video data; the advantages are relatively low cost of microphones when compared to cameras and high efficiency in detecting loud events that are frequently related to unlawful activities.

Applications of ***meta-data*** *obtained from video/sound analysis.* **Entrance and exit timestamp** – analysis of time passed by individual/vehicle in certain location may raise suspicions when it is unusual. Examples: pickpockets enter a hypermarket only for several

minutes, pickpockets pass long time in public transportation instead of using it for short time, vehicles that enter parking only for a short time may be suspected of passing illegal goods or suspicious meetings. The *biometric measures suited for the surveillance* may be, face image, ear image, iris, hand shape and geometry, skin texture, gait (way of walking) and voice.

Table 1. Surveillance data and purposes

Purposes	**Possible applications**	**Data used**
Person identification	Access control in restricted areas	biometric features, face recognition, voice
	Detection and apprehend of suspicious/wanted individuals	
Unusual/suspicious person behavior detection	Preventing unlawful activities	Individual behavior – trajectory, speed, gait etc.
	Detecting unhealthy humans (Alzheimer disease, epilepsy)	
Crowd flux statistics	Public transport management	Person detection
Road traffic statistics	Road traffic management	Vehicle detection
Anomaly behavior detection	Theft detection in parking lots, hypermarkets	People/vehicle identification and time of occurrence
Vehicle plate recognition	detection of stolen vehicles or vehicles with illegitimate plates (that may serve for unlawful activities)	Image of plates
Animal unusual looks detection	Detection of terrorist-related (i.e. bombing) threat connected to animals	Animal features (like silhouette) and behavior
	Detection of animals dangerous to people	
Vehicle unusual looks detection	Detection of terrorist-related threat connected to vehicles (i.e. rigged)	Vehicle features (i.e. unusually heavy load)

Techniques that could be a very interesting possibility could be the use of thermographic cameras (capable to register infrared image). That provides information additional to the standard "visible" image and permits the use of biometric techniques like *facial thermogram*. Thermography makes it also easier to detect and track moving targets in real-time. The disadvantage of the technique is however fragility and high cost of high quality thermographic cameras. Some Exemplary scenarios are given below:

1. Street abnormal activity detection – detecting suspicious/unlawful activities like theft, robbery, fights, gatherings, etc.
2. In-street (office, etc.) suspects identification and tracking – global (country, world) system with database of wanted criminals at disposal uses video/sound surveillance in other to extract people biometric features and use them in order to identify and track wanted criminals (within coverage of surveillance network)
3. Hooligans identification at sport event – system analogous to the previous but with smaller scale – aimed at identification of known hooligans and preventing them from entering the sport event. The identification is much easier as the people are usually waiting in the line, so the human location and direction of looking are predictable. Additionally the users' population is smaller.
4. Airport threat detection – access control, detecting of wanted criminals, detecting of suspicious behavior.

Alternative to webcams: secure private network of cameras monitoring high security areas (like airports, train and metro stations, offices, shops) connected to distributed analysis centers. This fulfils the paradigm of computing grid, so let us call the system Surveillance Computing Grid. Nobody wants the data from high security area to be publicly available for everyone (security and privacy trait), as the attacker may use the information for malicious purposes. Therefore the security of such an area is usually "an island" that is maintained locally by human experts looking at monitors. The network that connects such areas may prove to be advantageous if there is a standard automated procedure for monitoring for unlawful activities. That would replace or reduce the need for training and paying to human "experts" for manual monitoring. Surveillance Computing Grid advantages are as follows:

1. Computing Grid advances in technology may be used for organization and secure communication inside network.
2. Potentially lower cost, because "machine replaces human" - no need to train and maintain so many human experts for the purpose of detecting threats
3. One security center may serve for several high-security areas.

4. Once standard automated procedure for monitoring for unlawful activities is created it may be replicated easily.
5. Database of suspects and suspicious activities may be distributed and updated along network
6. All above-mentioned advantaged should benefit to efficiency of area protection.
7. Human surveillance experts are prone to mistakes (miss useful information) from boredom or tiredness; this setback is absent in automated surveillance systems.

Surveillance Computing Grid disadvantages:

1. Need for unobstructed secure communications channel and procedures between crucial parts of the system (Data gathering –> Data analysis -> Physical action)
2. The security conditions of specific area may vary from the standard automatic procedure, then the procedure needs to be modified (at a cost).

3.1. Surveillance Systems

We predict the substantial growth of importance and the need for automatic surveillance systems.

- Detect and analyze unusual/suspicious individual behavior in public places or private buildings – this can be used in many ways, basically for detecting unlawful activity, but also for detecting accidents/emergencies of deserted people (like old people living alone for example).
- Create a database of "wanted" individuals and identify them in the crowd – by combining many biometric techniques we are expecting to achieve high accuracy of identification. Furthermore many terrorist activities in European cities make need for security a pressing issue.

Biometric personal identification systems are in some cases complex and computationally intensive. Improving their effectiveness often involves increasing the number of features and incorporating additional pre-processing steps. Raw biometric data may be large and when the database contains data on many individuals a single machine may not be powerful/capable enough to store the data and process requests for identification. In order to achieve faster results, the identification systems can be paralleled or/and distributed. Computing Grid architectures with their large computational power offer great solution for such high workload biometric identification systems. They are naturally suited for distributing complex stages of such systems among many parallel units. Delegating an identification task to distributed environments enable the usage of simple and comfortable hardware at the point of application. The problem may be that many algorithms used in identification systems do not directly support redistribution of data. Therefore, there may be a need for developing parallel versions or designing new algorithms suitable for parallel computing. If we have more storage/processing power, we could:

- store and process large quantities of data (numerous individuals and/or numerous features, high "resolution" of data);
- increase the precision of algorithms;
- try to merge different sources of data/features;
- use raw data, so that future algorithm will be able to use this data, instead of using preprocessed/extracted feature vectors (essence of data).

4. Building Dynamic Social Networks

A considerable amount of behavioral monitoring (Surveillance) research is dealing with Web content. Researchers are interested in new arising topics of common interest, formation of social structures; businesses want to discover the profiles of their (potential) customers etc.

Therefore, Web mining technologies are widely applied for behavioral monitoring. In particular, clustering and classification techniques are of high interest. On one hand, One can cluster Web users based on their Web traversing behavior (click stream analysis), on the other hand Web content creators can be clustered based on the content of Web pages they create and on the links they insert, or both (Web content and Web structure mining). The placement of either users on the content clusters or of creators on user clusters may reveal interesting features. The natural clusters emerging may be also of interest for characterizing objects (users or creators) based on features of objects with known characteristics, belonging to the clusters.

Web content analysis may provide with keyword characterization, summarization, or with pointing to a typical representative. Pages with similar contents may be identified. Also the spamming behavior of Web content creators may be identified – both on content and link structure basis. Web content classification methods (operating either on clusters, or on individual objects) can serve creation of models for assignment of labels to unlabeled objects, given a pre-defined labeling of some subset. An example of such application is pornography detector. One of the most important features needed for surveillance are predicting and filtering capabilities, that is recommending systems. They may be used to point to the most interesting behavioral changes in the supervised world, or to point to probable future developments, like pointing to objects, which will

likely cooperate in future, or are worth bringing together. This kind of behavioral monitoring may be of interest both for security officers and various business activities (like reconsideration of advertisement placement, identification of hot topics to be included in advertisement actions, etc.).

5. Data Integrity and Security

To implement security in a grid computational environment, without appreciable performance degradation in grids. A suitable alternative to the computationally expensive encryption could be developed, which uses a key for message authentication. Methods of secure transfer and exchange of the required key(s) are also to be developed [2].

In a grid environment, it is important to seek solutions that do not rely directly on cryptography. Furthermore, another issue that arises is that use of encryption requires a substantial amount of computational resources. This is particularly so since the size of the key has to increase to protect the encrypted data. We propose less resource consuming methods of secure authentication and secure data transfer. For authentication, we investigate the idea of temporal distribution of key information. The process that needs to execute at a dedicated Domain Resource Manager (DRM) must first authenticate itself with that DRM. This authentication procedure should be secure to prevent entry to unauthorized programs. We suggest different methods, based on exchange of key information to secure subsequent communication and data transfer. For the issue of security in the transfer of data, an encryption-less method that still offers similar level of security could be used. For the issue of security in the transfer of data, an encryption-less method that still offers similar level of security could be used and we plan to investigate the Winnowing and Chaffing approach.

Distributed Intrusion Detection and Prevention Systems (DIDPS) could consolidate intrusion detection information from many different individual sensors in the grid network and even multiple grids. We propose a DIDPS that will be encapsulated inside mobile agents and placed at grid nodes / network [3].

6. QOS Based Surveillance System

Privacy can be seen as an aspect of security — one in which trade-offs between the interests of one group and another can become particularly clear. The right against unsanctioned invasion of privacy by the government, corporations or individuals is part of many countries' privacy laws, and in some cases, constitutions. Different EU Nations have different laws, which in some way limit privacy. In some nations individual privacy may conflict with freedom of speech laws and some laws may require public disclosure of information, which would be considered private in other nations and cultures. To protect critical infrastructures, in some cases, the Government or policy makers might have to voluntarily sacrifice privacy in exchange for perceived benefits and very often with specific dangers and losses. Final aspect of this project is to develop an adaptive surveillance system, which could maintain and protect privacy factors depending on the regulations and constitution (quality of service) of the countries involved.

7. Conclusions

This paper detailed a framework for cyber surveillance of unlawful activities using a computational grid based environment, which is capable of distributed data mining. Proposed framework integrates several areas of computer science research namely data mining, computational grids, biometrics, social networks.

References

[1] LOBSTER project web page: http://www.ist-lobster.org/

[2] Sanyal S, Vasudevan R, Abraham A and Paprzycki M, Grid Security and Integration with Minimal Performance Degradation, Journal of Digital Information Management, Vol. 2, Number. 3, pp. 122-126, 2004.

[3] Haslum K, Abraham A. and Knapskog S, DIPS: A Framework for Distributed Intrusion Prediction and Prevention Using Hidden Markov Models and Online Fuzzy Risk Assessment, Third International Symposium on Information Assurance and Security, IEEE Computer Society press, USA, ISBN 0-7695-2876-7, pp. 183-188, 2007.

[4] J. Albusac, J.J. Castro-Schez, L.M. Lopez-Lopez, D. Vallejo, L. Jimenez-Linares, A supervised learning approach to automate the acquisition of knowledge in surveillance systems, Signal Processing, Volume 89, Issue 12, pp. 2400-2414, 2009.

[5] Kevin Aquilina, Public security versus privacy in technology law: A balancing act?, Computer Law & Security Review, Volume 26, Issue 2, pp. 130-143, 2010.

[6] Anthony C. Caputo, Security Integration and Access Management, Digital Video Surveillance and Security, pp. 281-306, 2010.