PhD in Information Technology and Electrical Engineering

Università degli Studi di Napoli Federico II

PhD Student: Mauro Garofalo

XXIX Cycle

Training and Research Activities Report – First Year

Tutor: Giorgio Ventre

Information

Il sottoscritto Mauro Garofalo, laureato in Ingegneria Informatica presso l'Università di Napoli Federico II, iscritto al primo anno del Dottorato in Information Technology and Electrical Engineering – ITEE- (XXIX ciclo) sotto la supervisione del Prof. Giorgio Ventre, dichiara di aver partecipato alle attività di formazione e svolto le attività di ricerca esposte nel seguito.

Study and Training activities

Courses

- Theory and Applications of Piecewise Smooth Dynamical Systems, John Hogan, Giugno 2014
- Corso di EuroProgettazione, Gianpaolo Varchetta, Ottobre 2014
- Project Management per la Ricerca, Guido Capaldo, Febbraio 2015
- The Entrepreneurial Analysis of Engineering Research Projects, Luca Iandoli, Febbraio 2015
- Three core issues for the Internet: things, security and economics, Henning Schulzrinne, Febbraio 2015

Seminars

- High-Dimensional Pattern Recognition Via Sparse Representation, Prof Allen Yang, Giugno 2014
- Control System Design Using Energy Properties of Physical Systems, Alejandro Donarie, Giugno 2014
- Reliability and Availability Modelling in Practice - Kishor S Trivedi Ottobre 2014
- Capacity Planning for Infrastructure-as-a-Service Cloud - Kishor S Trivedi – Ottobre 2014
- Developmental Robotics for Embodied Language Learning, Angelo Cangelosi, Ottobre 2014
- UML Profiles for the specification of non functional properties of software systems, Simona Bernardi, Novembre 2014
- Verification of Mobile Agents in Partially Known Environments, Sasha Rubin, Novembre 2014
- Smoothed Particle Machine Perception: a proposed method for sensor fusion and physical-spacial perception, Nick Hockings. Gennaio 2015
- Risk management meets model checking: fault tree analysis and model-based testing via games, Mariëlle Stoelinga, Gennaio 2015
- Joint location and design optimization for resource allocation in software defined virtual networks, Antonia Tulino e Claudio Sterle, Gennaio 2015

- State of the art in Power Converters for High Voltage DC transmission systems, Philippe Ladoux, Gennaio 2015
- Efficient service distribution in next generation cloud networks, P. Festa, S. Avallone, R. Canonico, G. Di Stasi, S. Manfredi Febbraio 2015

Research activity

Title

"Analyzing BIG netFlow DATA for network anomaly detection: a distributed architecture"

Study

Negli ultimi anni c'è stata una costate crescita nell'utilizzo dei dispositivi mobili. Si calcola infatti che nel 2015 il numero di dispositivi mobili a livello mondiale supererà di un ordine di grandezza quello dei PC tradizionali [1]. Lo stesso trend si osserva anche nel traffico di rete mobile su scala mondiale (+69% nel 2014) [2]. Oltre a questi aspetti, la scarsa "cultura alla sicurezza" da parte degli utenti mobili [3] e la quantità di informazioni sensibili (personali e lavorative) all'interno di dispositivi ha reso gli smartphone vittime ideali dei software malevoli (malware).

Per identificare possibili infezioni si possono impiegare due principali strategie: (a) analizzare il dispositivo (ad esempio, il software in esecuzione, le prestazioni dell'hardware in termini di utilizzo della CPU, di consumo della batteria, etc.) [4], (b) analizzare il traffico di rete generato dal dispositivo [5]. Uno studio relativo alla caratterizzazione dei malware per la piattaforma Android [6] ha mostrato come la maggior parte (oltre il 90%) di questi software malevoli ha la capacità di ottenere il controllo remoto del dispositivo allo scopo di creare una rete di "vittime". Questo tipo di rete prende il nome di "Botnet". Sfruttando la numerosità e la distribuzione dei nodi di tale rete (detti zombie) si possono portare a termine differenti attività malevole come ad esempio: attacchi Distributed Denial of Service (DDoS), furto di informazioni sensibili, clickfraud o distribuzione di altro malware.

Gli Intrusion Detection System (IDS), sistemi capaci di rilevare la presenza di malintenzionati all'interno di una rete, si possono distinguere in base alla metodologia utilizzata in: Anomaly Based, Protocol Based, Behavior-Based e Rule-Based. Con il termine "Anomaly Based" ci si riferisce ad una metodologia che mira a scoprire, all'interno del traffico di rete, pattern che non sono conformi al comportamento normale previsto [12]. Gli Anomaly Based IDS, o più semplicemente Anomaly Detection System (ADS), sono stati proposti per la loro capacità di generalizzazione. Infatti, basandosi su algoritmi di Machine Learning, essi sono in grado di adattarsi alle continue modifiche di comportamento dei malware a differenza degli approcci classici (Protocol-Based e Rule-Based).

Dato il gran numero di dispositivi da monitorare, la velocità di risposta necessaria a rendere efficaci le contromisure e la varietà dei dati da trattare, possiamo definire questo un tipico problema di Big Data [8, 9].
Dal punto di vista dei dati da analizzare, gli approcci di analisi del traffico di rete si possono classificare in: port-based, packet-based (deep payload inspection o DPI) e flow-based. Il lavoro di ricerca si è concentrato su quest'ultimo approccio poiché può portare diversi vantaggi [13] in termini di:

- prestazioni, in quanto viene analizzata una quantità minore di dati rete rispetto agli altri approcci;
- rispetto della privacy degli utenti;
- maggiore applicabilità su larga scala.

Un flusso TCP/IP, sulla cui osservazione si basano gli approcci flow-based, è caratterizzato da una quintupla composta da: indirizzo sorgente, indirizzo destinazione, porta sorgente, porta destinazione, protocollo utilizzato. Gli approcci flow-based prevedono inoltre l'analisi di ulteriori informazioni come ad esempio la durata del flusso o il numero di byte e pacchetti trasmessi [10].

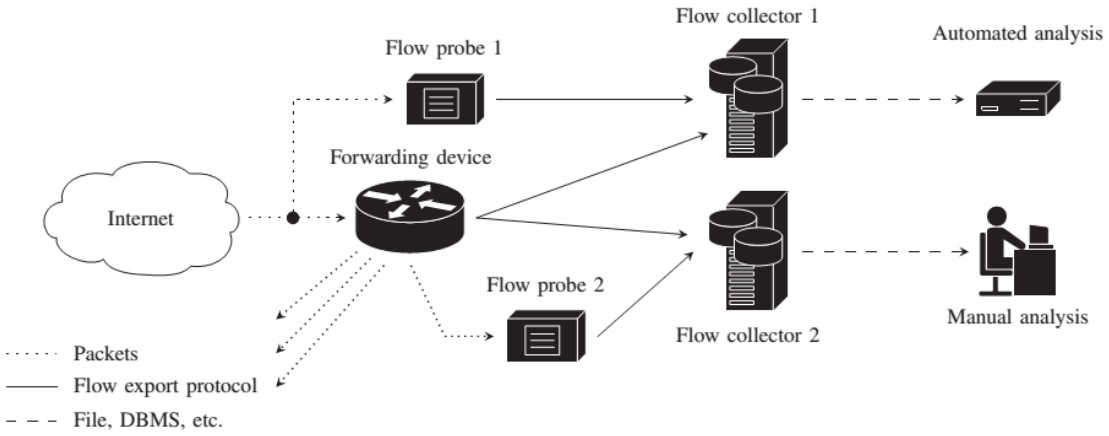

Figura 1 Architettura di riferimento [10]

L'obbiettivo della ricerca è quindi quello di sviluppare un framework di Anomaly Detection che analizzi i dati relativi ai flussi di rete, raccolti ad esempio dai router, sfruttando algoritmi di Machine Learning. Il sistema, a regime, dovrà essere in grado di risolvere problemi di analisi del traffico di rete allo scopo di identificare dispositivi mobili potenzialmente.

Un modello dell'architettura, rappresentato nella figura precedente, è in fase di definizione e di realizzazione prototipale. Tale architettura fa uso di tecnologie per l'analisi di Big Data (come Apache Hadoop e Apache Spark). Il proseguo del lavoro nel breve periodo mira a renderla operativa per poter condurre i primi test di validazione e misura delle performance al variare dei possibili algoritmi di Machine Learning.

Bibliography

[1] Gartner 2014 http://www.gartner.com/newsroom/id/2791017

[2] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update

[3] Norton 2014. Internet Security Threat Report 2014

[4] Hien Thi Thu Truong, Eemil Lagerspetz, Petteri Nurmi, Adam J. Oliner, Sasu Tarkoma, N. Asokan, Sourav Bhattacharya, "The Company You Keep: Mobile Malware Infection Rates and Inexpensive Risk Indicators", 2013, arXiv:1312.3245

[5] Taehyun Kim; Yeongrak Choi; Seunghee Han; Jae Yoon Chung; Jonghwan Hyun; Jian Li; Hong, J.W.-K., "Monitoring and detecting abnormal behavior in mobile cloud infrastructure," *Network Operations and Management Symposium (NOMS), 2012 IEEE* , vol., no., pp.1303,1310, 16-20 April 2012

[6] Yajin Zhou; Xuxian Jiang, "Dissecting Android Malware: Characterization and Evolution," Security and Privacy (SP), 2012 IEEE Symposium on , vol., no., pp.95,109, 20-23 May 2012

[7] Bhuyan, M.H.; Bhattacharyya, D.K.; Kalita, J.K., "Network Anomaly Detection: Methods,

Systems and Tools," Communications Surveys & Tutorials, IEEE , vol.16, no.1, pp.303,336, First Quarter 2014

[8] Zibin Zheng; Jieming Zhu; Lyu, M.R., "Service-Generated Big Data and Big Data-as-a-Service: An Overview," Big Data (BigData Congress), 2013 IEEE International Congress on , vol., no., pp.403,410, June 27 2013-July 2 2013

[9] Zaslavsky, A.; Perera, C.; Georgakopoulos, D., "Sensing as a Service and Big Data", 2013, arXiv preprint arXiv:1301.0159

[10] Hofstede, R.; Celeda, P.; Trammell, B.; Drago, I.; Sadre, R.; Sperotto, A.; Pras, A., "Flow Monitoring Explained: From Packet Capture to Data Analysis With NetFlow and IPFIX," Communications Surveys & Tutorials, IEEE , vol.16, no.4, pp.2037,2064, Fourthquarter 2014

Collaborations

Nell'ambito della valutazione di possibili architetture di calcolo ad alte prestazioni per l'analisi dei dati, si è avviata una collaborazione con il Dipartimento di Fisica dell'Università di Napoli Federico II, il California Institute of Technology e l'Osservatorio Astronomico di Capodimonte, che fanno parte del gruppo di ricerca DAME (Data Mining & Exploration). Il gruppo si occupa dell'analisi di Massive Data Sets con metodi di Machine Learning utilizzando architetture sia di calcolo distribuito (GRID), sia di calcolo parallelo (GPU).

Products

Publication

Journal: Publications Of The Astronomical Society Of The Pacific, Vol. 126, P. 783-797, ISSN: 0004-6280

Title: DAMEWARE: A Web Cyberinfrastructure for Astrophysical Data Mining

Authors: Brescia M, Cavuoti S, Garofalo Mauro, et al. (2014)

Abstract

Astronomy is undergoing through a methodological revolution triggered by an unprecedented wealth of complex and accurate data. The new panchromatic, synoptic sky surveys require advanced tools for discovering patterns and trends hidden behind data, which are both complex, and of high dimensionality. We present DAMEWARE (DAta Mining & Exploration Web Application REsource): a general purpose, web-based, distributed data mining environment developed for the exploration of large datasets, and finely tuned for astronomical applications. By means of graphical user interfaces, it allows the user to perform classification, regression or clustering tasks with machine learning methods. Salient features of DAMEWARE include its capability to work on large datasets with minimal human intervention, and to deal with a wide variety of real problems such as the classification of globular clusters in the galaxy NGC1399, the evaluation of photometric redshifts and, finally, the identification of candidate Active Galactic Nuclei in multiband photometric surveys. In all these applications, DAMEWARE allowed to achieve better results than those attained with more traditional methods. With the aim of providing potential users with all needed information, in this paper we briefly describe the technological background of DAMEWARE, give a short introduction to some relevant aspects of data mining, followed by a summary of some science cases and, finally, we provide a detailed description of a template use case.

List those in preparation

Title: A survey on Big Data and Networking

Title: An architecture for flow-based anomaly detection based on Big Data analysis tools

Conferences and Seminars

Conference Paper

Conference: Perspectives of GPU computing in Physics and Astrophysics. Università la Sapienza di Roma, Roma 15-17 Settembre 2014

Title: Acceleration of Machine Learning Models based on GPGPU technology for fast data mining in multidisciplinary physical environments.

Authors: Garofalo Mauro, Guarino D, Brescia M, Cavuoti S, Pescapè A, Longo G, Ventre G

Abstract

The continued growth of data and their use to solve complex problems in all fields of science needs advanced tools to analyze and process them. These tools require more and more high computing power. GPUs are very attractive for acceleration of general-purpose application, due to their widespread and affordability, becoming in some case crucial for computationally demand algorithms. Based on our recent experiences, we present and discuss several GPU based solutions to accelerate time-consuming machine learning models, such as pure genetic algorithms (GA), hybrid GAs mixed with feed-forward neural networks and Support Vector Machine, all models that have been developed in house. We approached and compared different acceleration technologies, ranging from Thrust (a high-level CUDA library) to OpenACC a compilation tool based on directives, through CUDA C, reaching a satisfying level of performance optimization, up to 200x of speedup in some case. We also discuss their scientific validation, successfully assessed on a series of multidisciplinary problems in Astrophysics (classification of globular clusters, photometric redshift estimation), Medicine (early prediction of Alzheimer disease from hippocampus volume classification through magnetic resonance image of human brain), Computer Networks (network traffic classification). Finally, we highlighting the advantages and disadvantages of development approaches and some future evolutions and challenges.

CS Summary

	Credits year 1								Credits year 2								Credits year 3									
		1	2	3	4	5	6			1	2	3	4	5	6			1	2	3	4	5	6			
	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Total	Check
Modules	20		3		3		6	**12**	18							**0**	0							**0**	**12**	30-70
Seminars	7		0,6		2	0,9	2,2	**5,7**	6							**0**	0							**0**	**5,7**	10-30
Research	33	10	6,4	10	5	9,1	1,8	**42,3**	36							**0**	60							**0**	**42,3**	80-140
	60	10	10	10	10	10	10	**60**	60	0	0	0	0	0	0	**0**	60	0	0	0	0	0	0	**0**	60	180