THE IEEE

Intelligent Informatics

BULLETIN

IEEE COMPUTER SOCIETY

IEEE Computer Society
Technical Committee
on Intelligent Informatics

December 2013 Vol. 14 No. 1 (ISSN 1727-5997)

Profile

Feature Articles

Announcements

On-line version: http://www.comp.hkbu.edu.hk/~iib (ISSN 1727-6004)

IEEE Computer Society Technical Committee on Intelligent Informatics (TCII)

Executive Committee of the TCII:

Chair: Jiming Liu
Hong Kong Baptist University, HK
Email: jiming@comp.hkbu.edu.hk

Vice Chair: Chengqi Zhang
(membership, etc.)
University of Technology, Sydney, Australia
Email: chengqi@it.uts.edu.au

Jeffrey M. Bradshaw
(conference sponsorship)
Institute for Human and Machine Cognition, USA
Email: jbradshaw@ihmc.us

Nick J. Cercone
(early-career faculty/student mentoring)
York University, Canada
Email: ncercone@yorku.ca

Pierre Morizet-Mahoudeaux
(curriculum/training development)
University of Technology of Compiegne, France
Email: pmorizet@hds.utc.fr

Toyoaki Nishida
(university/industrial relations)
Kyoto University, Japan
Email: nishida@i.kyoto-u.ac.jp

Vijay Raghavan
(TCII Bulletin)
University of Louisiana- Lafayette, USA
Email: raghavan@louisiana.edu

Past Chair: Ning Zhong
Maebashi Institute of Technology, Japan
Email: zhong@maebashi-it.ac.jp

The Technical Committee on Intelligent Informatics (TCII) of the IEEE Computer Society deals with tools and systems using biologically and linguistically motivated computational paradigms such as artificial neural networks, fuzzy logic, evolutionary optimization, rough sets, data mining, Web intelligence, intelligent agent technology, parallel and distributed information processing, and virtual reality. If you are a member of the IEEE Computer Society, you may join the TCII without cost at http://computer.org/tcsignup/.

The IEEE Intelligent Informatics Bulletin

Aims and Scope

The IEEE Intelligent Informatics Bulletin is the official publication of the Technical Committee on Intelligent Informatics (TCII) of the IEEE Computer Society, which is published once a year in both hardcopies and electronic copies. The contents of the Bulletin include (but may not be limited to):

1) Letters and Communications of the TCII Executive Committee

2) Feature Articles

3) R&D Profiles (R&D organizations, interview profile on individuals, and projects etc.)

4) Book Reviews

5) News, Reports, and Announcements (TCII sponsored or important/related activities)

Materials suitable for publication at the IEEE Intelligent Informatics Bulletin should be sent directly to the Associate Editors of respective sections.

Technical or survey articles are subject to peer reviews, and their scope may include the theories, methods, tools, techniques, systems, and experiences for/in developing and applying biologically and linguistically motivated computational paradigms, such as artificial neural networks, fuzzy logic, evolutionary optimization, rough sets, and self-organization in the research and application domains, such as data mining, Web intelligence, intelligent agent technology, parallel and distributed information processing, and virtual reality.

Editorial Board

Editor-in-Chief:

Vijay Raghavan
University of Louisiana- Lafayette, USA
Email: raghavan@louisiana.edu

Managing Editor:

William K. Cheung
Hong Kong Baptist University, HK
Email: william@comp.hkbu.edu.hk

Assistant Managing Editor:

Xin Li
Beijing Institute of Technology, China
Email: xinli@bit.edu.cn

Associate Editors:

Mike Howard (R & D Profiles)
Information Sciences Laboratory
HRL Laboratories, USA
Email: mhoward@hrl.com

Marius C. Silaghi
(News & Reports on Activities)
Florida Institute of Technology, USA
Email: msilaghi@cs.fit.edu

Ruili Wang (Book Reviews)
Inst. of Info. Sciences and Technology
Massey University, New Zealand
Email: R.Wang@massey.ac.nz

Sanjay Chawla (Feature Articles)
School of Information Technologies
Sydney University, NSW, Australia
Email: chawla@it.usyd.edu.au

Ian Davidson (Feature Articles)
Department of Computer Science
University at Albany, SUNY, USA
Email: davidson@cs.albany.edu

Michel Desmarais (Feature Articles)
Ecole Polytechnique de Montreal, Canada
Email: michel.desmarais@polymtl.ca

Yuefeng Li (Feature Articles)
Queensland University of Technology Australia
Email: y2.li@qut.edu.au

Pang-Ning Tan (Feature Articles)
Dept of Computer Science & Engineering
Michigan State University, USA
Email: ptan@cse.msu.edu

Shichao Zhang (Feature Articles)
Guangxi Normal University, China
Email: zhangsc@mailbox.gxnu.edu.cn

Publisher: *The IEEE Computer Society Technical Committee on Intelligent Informatics*
Address: *Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (Attention: Dr. William K. Cheung; Email:william@comp.hkbu.edu.hk)*
ISSN Number: *1727-5997(printed)1727-6004(on-line)*
Abstracting and Indexing: *All the published articles will be submitted to the following on-line search engines and bibliographies databases for indexing—Google(www.google.com), The ResearchIndex(citeseer.nj.nec.com), The Collection of Computer Science Bibliographies (liinwww.ira.uka.de/bibliography/index.html), and* ***DBLP*** *Computer Science Bibliography (www.informatik.uni-trier.de/~ley/db/index.html).*

© 2014 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the ***IEEE****.*

The Cognitive Anteater Robotics Laboratory (CARL) at the University of California, Irvine

I. Introduction

Within our grasp is a deep scientific understanding of how the brain's mechanisms give rise to perception, cognition, emotion, action and social engagement with others. Such an understanding will have a revolutionary impact on science, medicine, economic growth, security, and social wellbeing. One way to understand this complicated system is through the construction of working models. Developing neural models that follow the architecture and dynamics of brain networks, combined with building robotic systems that physically ground these models, has great potential to solve one of the Grand Challenges posed by the United States National Academy of Engineering: *Reverse-Engineering the Brain*. Our laboratory attempts to meet this challenge in four complementary ways by: 1) Promoting the field of Cognitive and Brain-Based Robotics. 2) Developing adaptive action selection systems based on principles of vertebrate neuromodulation. 3) Data-mining neuroinformatic and gene expression databases. 4) Constructing large-scale, detailed models of cortical and subcortical processing on parallel computing platforms.

II. Cognitive and Brain-Based Robotics

For over 10 years, we have been promoting the field of Cognitive Robotics, or Neurorobotics. These brain-based robots are physical devices whose control systems have been modeled after aspects of brain processing. The goals of these robots are to better understand cognition through the construction of physical artifacts, and to create practical systems that demonstrate cognitive capabilities. Neurorobotics is built on the notion that the brain is embodied in the body, which is, in turn, embedded in the environment, and that this coupling is necessary for an intelligent system. The field is small, but growing, due to technological advances and increased interdisciplinary research. Our group has developed a series of neurorobotic models that have successfully demonstrated perceptual categorization and conditioning [1], visual binding and scene segmentation [2], texture discrimination with artificial whiskers [3], adaptive motor control [4], spatial memory and navigation [5, 6], and neuromodulation as a general-purpose robot control system [7, 8]. These algorithms have several important features for autonomous robot control in general, such as fluid switching of behavior, gating in important sensory events, and separating signal from noise. Our algorithms and models have been tested on several robotic platforms in our laboratory, and we are currently working with other robotics laboratories around the world to demonstrate their applicability.

As an extension of our previous neurorobotic work in spatial memory and navigation, we are developing cognitive robots capable of contextual learning. A main goal of this research is to create a robot capable of constructing a cognitive map of its environment while foraging for different valued resources under varying environmental conditions. The system should lead to a better understanding of how areas of the medial temporal lobe interact with cortical areas to create flexible episodic memory. Such a system would be a major step forward for autonomous navigation by artificial systems.

Another research direction of our lab, which builds upon our cognitive robotics work, is to deploy teams of cognitive robots. These robot teams, or swarms, can be fairly large in size, and as a result, an inexpensive robot with strong communication capabilities is favourable. To that end, we have developed an open source robotic platform that leverages smartphone technology as a control system [9, 10]. The computing, communication, and sensing capabilities of current smartphones affords an inexpensive yet highly capable robotic platform that can be used for education and research. The platform, called leCarl, consists of an Android phone, R/C car platform, IOIO interface board, and additional sensors (see Figure 1). In the near future, our action selection, learning, and cognitive mapping algorithms will be deployed on a leCarl swarm in a Search and Rescue task.

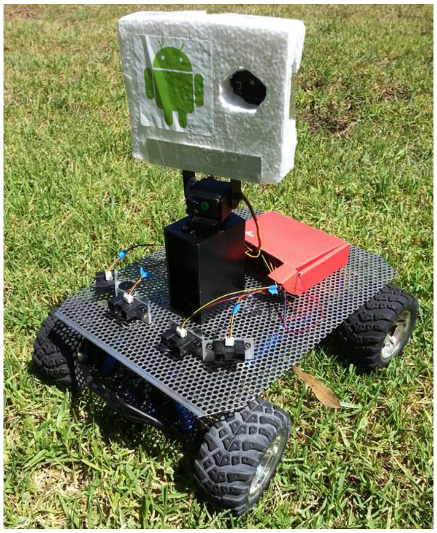

Figure 1. Android based robotic platform. The Android phone serves as the computing and sensing device. The IOIO provides an interface to add additional sensors, such as IR range finders. The base is installed on the chassis of a R/C truck. The robotic head is composed of a rectangular tube, two servos for the pan and tilt unit, and a phone holder made of foam. Adapted from [9].

III. Adaptive Action Selection Systems Based on Principles of Vertebrate Neuromodulation

The vertebrate neuromodulatory system plays a key role in regulating decision-making and responding to environmental challenges. In particular, the serotonergic system underlies control of stress, social interactions, and risk-taking behavior. The dopaminergic system has been implicated in the prediction of rewards and incentive salience. The cholinergic and noradrenergic systems are thought to play important roles in attention and judging uncertainty. We suggested that the behavior of an autonomous system modeled after the vertebrate neuromodulatory system, might demonstrate the complexity and flexibility associated with higher order animals by monitoring its surroundings, adapting to change, and responding decisively to important environmental events [11]. Since the publication of this paper, our group has demonstrated how these systems can modulate attention in uncertain environments [12], shape decision-making in social situations [13, 14], and be used as an adaptive controller for autonomous robots [8, 15]. Our attentional study showed how the noradrenergic and cholinergic systems interact with each other, and suggested how this could lead to behavioral adaptation in the face of uncertainty [12]. We suggested that basal forebrain activity tracks expected uncertainty and that this shapes attentional search. We also suggested that the locus coeruleus tracks unexpected uncertainty, and this leads rapid responses to changes in the environment.

Game theory can be a powerful tool for testing models and discovering the neural correlates of decision-making in cooperative and competitive situations. In a set of human robot interaction studies using socioeconomic game theory, specifically the Hawk-Dove game, we showed that adaptive agents, whose behavior is guided by simulated dopaminergic and serotonergic systems, could evoke changes in strategy, reward/cost tradeoffs, and reciprocal behavior in subjects [14]. We also showed that division into two groups best described subjects' responses during these games [13]. Lowering subjects' serotonin levels through Acute Tryptophan Depletion caused some subjects to be more aggressive (as expected), but others to be less aggressive (unexpected). We suggest that individual variation, possibly due to genetic differences in serotonin and dopamine action, may be influencing this variability. To further understand this relationship, we turned to another socioeconomic game, called the Stag Hunt, which focuses on cooperation. In the Stag Hunt, subjects can either hunt a low valued hare on their own or form a social contract with another player to hunt a highly valued stag (see Figure 2). We constructed an adaptive agent, based on the interaction between the dopaminergic and serotoninergic systems, which learned to play Stag Hunt and develop strategies based on the human player's tendencies [16]. In this study, we tested the performance of 40 subjects playing against five opponent types (the adaptive agent, and four other set strategies) in a spatiotemporal version of the Stag Hunt game. Subjects put more thought in their movements and in considering the movements of the agent when playing against the adaptive agent. Similar to our Hawk-Dove study, we observed differences between subjects on the individual level, with several responding to the adaptive agent by almost always cooperating, and several others remaining nearly exclusively uncooperative. In future work, we are interested in both the development of the agent strategy and the subjects' reaction to adaptive agents. Moreover, we plan to further investigate the neural correlates of these behaviors through brain imaging, pharmacological manipulations and genetic screening.

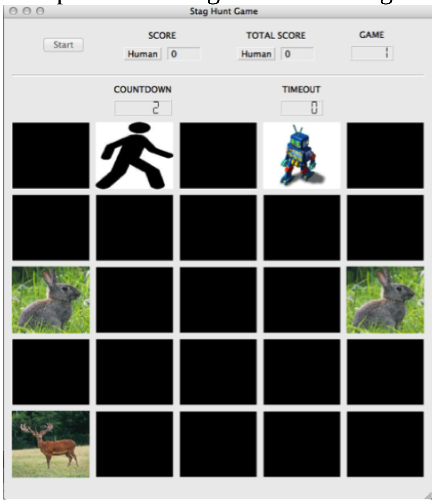

Figure 2. Screenshot of Stag Hunt game board. The game board included a 5x5 grid of spaces upon which the player (stick figure image), agent (robot image), stag (stag image), and hare (hare image) tokens resided. The screen included a button to start the experiment, the subject's score for the round, the subject's overall score for the experiment, the game number within the round, a 3-second countdown to the start of the game, and a 10-second counter monitoring the game's timeout. At the beginning of each game, the locations for the stag, player, and agent tokens were randomly placed along either the top row, bottom row, or middle column at least one square away from each other. The initial positions of the hares were fixed in the locations shown above for all games. The player and agent could move one square at a time towards their goal at the start of the game, while the targets remain fixed. Adapted from [16].

IV. Data-Mining Neuroinformatic and Gene Expression Databases.

In addition to our modeling work, we are taking a neuroinformatic approach to understanding cognitive function. Neuroinformatics is an emerging technique concerned with the management and sharing of neuroscience data. In recent work, we performed an exploratory survey of receptor gene expression associated with classical neuromodulatory systems (i.e., cholinergic, dopaminergic, noradrenergic, and serotonergic) within anatomical origins of these neuromodulatory systems, as well as in the amygdala [17]. Investigation of receptor gene expression in these regions was undertaken using the Allen Mouse Brain Atlas, a growing neuroinformatic resource that contains data sets of extensive mouse gene expression and neuroanatomical data. As a result, this type of exploratory analysis revealed many connectivity relations and receptor localization of these neuromodulatory systems that had not been previously reported (Figure 3). Currently, we are using this approach to understand the structural and functional underpinnings of reward processing by acquiring and analyzing expression data from dopamine and serotonin signaling genes across brain areas associated with the reward circuit.

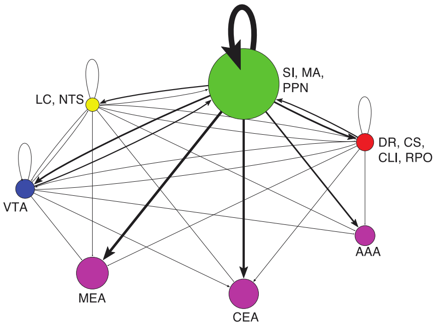

Figure 3. Network model showing overall expression of neuromodulatory receptors and their implied neuromodulatory projections to target areas. Vertices represent brain regions that are either standalone (purple = amygdala regions) or combined regions (yellow = noradrenergic, green = cholinergic, blue = dopaminergic, and red = serotonergic). Directed arcs represent projections going to and coming from a source. The pointed-arrow indicates the target location and the non-arrow end of the arc indicates the origin. The thickness of each arc, as well as the size of vertices,

is proportional to the amount of expression found in the target location. Adapted from [17].

V. Detailed Models of Cortical and Subcortical Processing on Parallel Computing Platforms

Despite recent increases in computer power, constructing a neural model that approaches the size of a human-brain will require several orders of magnitude increases in computation, communication, and memory capacity. Conventional computer hardware may not be the appropriate architecture for modeling a brain. Unlike a conventional computer, the brain is a massively parallel, analog, fault-tolerant, selective system that does not rely on programmed instructions. Alternative computer architectures and programming paradigms, which are neurobiologically inspired, are in need of investigation [18, 19]. Our group has been developing tools to incorporate these brain features into computer models. Specifically, we have constructed large-scale network models that capture the dynamics of neural signaling at the microcircuit (i.e., within brain areas) and macrocircuit (i.e., between brain areas) levels. We have developed a highly efficient implementation of Spiking Neural Networks (SNN) by leveraging the parallel computing power of Graphical Processing Units (GPUs). Our publically available software program, called CARLsim (http://www.socsci.uci.edu/~jkrichma/CARLsim), is a C/C++ based SNN simulator that runs on both generic x86 CPUs and standard off-the-shelf GPUs. With our optimizations, we have demonstrated roughly 25X speedups over cutting edge desktop computers. This simulation environment was released to the modeling community so that researchers would have easy access to large-scale SNN simulations [20]. It has been very popular among computer scientists, neuroscientists, and engineers. Our latest release of simulator software extended this prior model to include more biologically plausible descriptions of synaptic connections and learning rules [21]. In particular, this new simulation environment facilitates the development of very large-scale spiking neural networks that follow the brain's architecture. Using this simulator environment, we developed cortical models of visual form, color, and motion processing in which we replicated color opponency and motion perception results at both the psychophysical and neuronal level (see Figure 4). This simulation environment has also been used to replicate a recent and important finding on how basal forebrain activation can enhance cortical coding of natural scenes [22]. Our spiking neuron model, which included the basal forebrain, thalamus, and visual cortex, suggested that basal forebrain activation switches the firing mode of thalamic neurons, which in turn leads to an increase in within-cell reliability and a decrease in between-cell redundancy in LGN and visual cortex. In near future releases of our spiking simulator, we plan to introduce an automated parameter tuning framework, and a more extensive visual motion perception model. In addition, we are expanding our GPU-accelerated spiking neural network simulator (CARLsim) to run across many GPUs with the use of MPI. We believe this work in the spiking neural network domain will have a broad impact on the neuromorphic engineering community and will one day lead to practical applications deployed on specialized hardware.

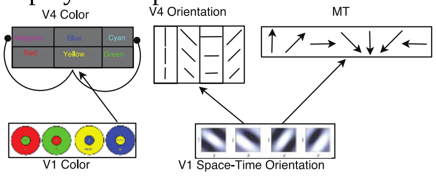

Figure 4. Architecture of the spiking neural network model of visual cortex. In the V1 color layer, there are four color opponent (center+/surround−) responses, which are combined in V4 to respond to six primary colors. The V1 motion energy model projects to edge detecting neurons in V4 and directionally selective neurons in cortical area MT Adapted from [21].

VI. Summary

By combining computational modeling and neuroinformatics with autonomous robots and parallel computing techniques, our group has created a multi-disciplinary approach to understanding the inner workings of the brain and cognition. It is our hope that this approach will continue to benefit both the neuroscience and computer science communities and move us closer to meeting the grand challenge of reverse-engineering the brain.

Contact Information

Derrik E. Asher dasher@uci.edu
Michael C. Avery averym@uci.edu
Michael Beyeler mbeyeler@uci.edu
Liam D. Bucci, lbucci@uci.edu
Kristofor D. Carlson, kdcarlso@uci.edu
Alexis B. Craig, acraig1@uci.edu
Nikil Dutt, dutt@ics.uci.edu
Jeffrey L. Krichmar, jkrichma@uci.edu
Nicolas Oros, noros@uci.edu
Emily L. Rounds, roundse@uci.edu
Andrew Zaldivar, azaldiva@uci.edu

Website:
http://www.socsci.uci.edu/~jkrichma

VII. References

[1] J. L. Krichmar and G. M. Edelman, "Machine Psychology: Autonomous Behavior, Perceptual Categorization, and Conditioning in a Brain-Based Device," *Cerebral Cortex*, vol. 12, pp. 818-830, 2002.

[2] A. K. Seth, J. L. McKinstry, G. M. Edelman, and J. L. Krichmar, "Visual Binding Through Reentrant Connectivity and Dynamic Synchronization in a Brain-based Device," *Cereb Cortex*, vol. 14, pp. 1185-1199, 2004.

[3] A. K. Seth, J. L. McKinstry, G. M. Edelman, and J. L. Krichmar, "Active sensing of visual and tactile stimuli by brain-based devices," *International Journal of Robotics and Automation*, vol. 19, pp. 222-238, 2004.

[4] J. L. McKinstry, G. M. Edelman, and J. L. Krichmar, "A cerebellar model for predictive motor control tested in a brain-based device," *Proc Natl Acad Sci U S A*, vol. 103, pp. 3387-92, Feb 28 2006.

[5] J. L. Krichmar, A. K. Seth, D. A. Nitz, J. G. Fleischer, and G. M. Edelman, "Spatial navigation and causal analysis in a brain-based device modeling cortical-hippocampal interactions," *Neuroinformatics*, vol. 3, pp. 197-222, 2005.

[6] J. L. Krichmar, D. A. Nitz, J. A. Gally, and G. M. Edelman, "Characterizing functional hippocampal pathways in a brain-based device as it solves a spatial memory task," *Proc Natl Acad Sci U S A*, vol. 102, pp. 2111-6, Feb 8 2005.

[7] J. L. Krichmar, "A neurorobotic platform to test the influence of neuromodulatory signaling on anxious and curious behavior," *Front Neurorobot*, vol. 7, pp. 1-17, 2013.

[8] B. R. Cox and J. L. Krichmar, "Neuromodulation as a Robot Controller: A Brain Inspired Design Strategy for Controlling Autonomous Robots," *IEEE Robotics & Automation Magazine*, vol. 16, pp. 72-80, 2009.

[9] N. Oros and J. L. Krichmar, "Smartphone Based Robotics: Powerful, Flexible and

Inexpensive Robots for Hobbyists, Educators, Students and Researchers," *IEEE Robotics & Automation Magazine*, p. Under review, 2013.

[10] N. Oros and J. L. Krichmar, "Neuromodulation, Attention and Localization Using a Novel Android™ Robotic Platform," presented at the ICDL-EpiRob 2012 : IEEE Conference on Development and Learning and Epigenetic Robotics, San Diego, CA, 2012.

[11] J. L. Krichmar, "The Neuromodulatory System - A Framework for Survival and Adaptive Behavior in a Challenging World.," *Adaptive Behavior,* vol. 16, pp. 385-399, 2008.

[12] M. C. Avery, D. A. Nitz, A. A. Chiba, and J. L. Krichmar, "Simulation of Cholinergic and Noradrenergic Modulation of Behavior in Uncertain Environments," *Frontiers in Computational Neuroscience,* vol. 6, pp. 1-16, 2012-January-31 2012.

[13] D. E. Asher, S. Zhang, A. Zaldivar, M. D. Lee, and J. L. Krichmar, "Modeling individual differences in socioeconomic game playing," in *COGSCI 2012 - The Annual Meeting of the Cognitive Science Society*, Sapporo, Japan, 2012.

[14] D. E. Asher, A. Zaldivar, B. Barton, A. A. Brewer, and J. L. Krichmar, "Reciprocity and Retaliation in Social Games with Adaptive Agents," *IEEE Transactions on Autonomous Mental Development*, In Press.

[15] J. L. Krichmar, "A Biologically Inspired Action Selection Algorithm Based on Principles of Neuromodulation," in *IEEE World Congress on Computational Intelligence*, Brisbane, Australia, 2012, pp. 1916-1923.

[16] A. B. Craig, D. E. Asher, N. Oros, A. A. Brewer, and J. L. Krichmar, "Social contracts and human-computer interaction with simulated adapting agents," *Adaptive Behavior,* 2013.

[17] A. Zaldivar and J. Krichmar, "Interactions between the neuromodulatory systems and the amygdala: exploratory survey using the Allen Mouse Brain Atlas," *Brain Structure and Function,* pp. 1-18, 2012.

[18] J. L. Krichmar, N. Dutt, J. M. Nageswaran, and M. Richert, "Neuromorphic Modeling Abstractions and Simulation of Large-Scale Cortical Networks," in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, San Jose, CA, 2011, pp. 334-338.

[19] J. M. Nageswaran, M. Richert, N. Dutt, and J. L. Krichmar, "Towards reverse engineering the brain: Modeling abstractions and simulation frameworks," in *VLSI System on Chip Conference (VLSI-SoC), 2010 18th IEEE/IFIP*, 2010, pp. 1-6.

[20] J. M. Nageswaran, N. Dutt, J. L. Krichmar, A. Nicolau, and A. V. Veidenbaum, "A configurable simulation environment for the efficient simulation of large-scale spiking neural networks on graphics processors," *Neural Netw,* vol. 22, pp. 791-800, Jul-Aug 2009.

[21] M. Richert, J. M. Nageswaran, N. Dutt, and J. L. Krichmar, "An efficient simulation environment for modeling large-scale cortical processing," *Frontiers in Neuroinformatics,* vol. 5, pp. 1-15, 2011-September-14 2011.

[22] M. Avery, J. L. Krichmar, and N. Dutt, "Spiking Neuron Model of Basal Forebrain Enhancement of Visual Attention," in *IEEE World Congress on Computational Intelligence*, Brisbane, Australia, 2012.

Intelligent Web Crawling

(WI-IAT 2013 Tutorial)

Denis Shestakov

***Abstract*—Web crawling, a process of collecting web pages in an automated manner, is the primary and ubiquitous operation used by a large number of web systems and agents starting from a simple program for website backup to a major web search engine. Due to an astronomical amount of data already published on the Web and ongoing exponential growth of web content, any party that want to take advantage of massive-scale web data faces a high barrier to entry. We start with background on web crawling and the structure of the Web. We then discuss different crawling strategies and describe adaptive web crawling techniques leading to better overall crawl performance. We finally overview some of the challenges in web crawling by presenting such topics as collaborative web crawling, crawling the deep Web and crawling multimedia content. Our goals are to introduce the intelligent systems community to the challenges in web crawling research, present intelligent web crawling approaches, and engage researchers and practitioners for open issues and research problems. Our presentation could be of interest to web intelligence and intelligent agent technology communities as it particularly focuses on the usage of intelligent/adaptive techniques in the web crawling domain.**

***Index Terms*—web crawling, web crawler, intelligent crawling, adaptive crawling, collaborative crawling, Web ecosystem, Web structure, incremental crawling, focused crawling, deep Web**

I. Introduction

WEB crawling [1], [2], a process of collecting web pages in an automated manner, is the primary and ubiquitous operation used by a large number of web systems and agents starting from a simple program for website backup to a major web search engine. For example, search engines such as Google or Microsoft Bing use web crawlers to routinely visit billions of web pages, which are then indexed and made available for answering user search requests. In this way, the characteristics of obtained web crawls such as coverage or freshness directly affect on the quality of web search results served to users. Besides web search, the web crawling technology is central in such applications as web data mining and extraction, web monitoring, social media analysis, digital preservation (i.e., web archiving), detection of web spam and fraudulent web sites, web application testing, finding unauthorized use of copyrighted content (music, videos, texts, etc.), identification of illegal and harmful web activities (e.g., terrorist chat rooms), and virtual tourism.

Due to an astronomical amount of data already published on the Web and ongoing exponential growth of web content, any party (be it an individual, company, government agency, non-profit or educational organization) that want to take advantage of massive-scale web data faces a high barrier to entry. Indeed, only network costs associated with the downloading of web-scale size collection by themselves lead to expenses that are not affordable by the majority of potential players.

For those with flexible budgets, there is a next barrier: operating web-scale crawl, i.e. hundreds of millions of pages, is a challenging task that requires skills and expertise in distributed data retrieval and processing, not to mention large operational costs. Finally, for the parties who nevertheless manage to overcome the above obstacles but interested in specific subsets of web information, the results of crawl are often wasteful, as majority of retrieved pages do not match their criteria of interest.

In this paper, we will overview recent advances made in harvesting the information on the Web, in order to introduce the intelligent systems community to the challenges in this area, with particular stress on intelligent web crawling approaches using adaptive crawling agents as well as the underlying open issues and research problems. We will also address issues in building a spectrum of services and applications collecting and aggregating large amounts of web information, e.g., the role of web crawlers in the Web ecosystem, how intelligent crawling strategies can lead to a better overall quality of crawled data.

II. Web Crawling

This section will introduce the basics of web crawler operations and important web crawling applications, and provide relevant statistics on the Web link structure. Next we will describe the architecture of a web crawler and present a number of crawling strategies including three adaptive crawling approaches.

A. Overview

The underlying mechanism of crawling – namely, given an URL download a corresponding web page, extract all URL links from it and repeat the process for those links that were not visited yet – is naive and simple. However, due to a number of imposing restrictions and resource limitations under which crawlers operate, algorithms and techniques behind a large-scale web crawler are far more complicated than the trivial implementation. For example, in order not to be banned by a web server, a crawler has to avoid sending too many URL requests to a server within a short time period. Since the distribution of pages over web servers is non-uniform, a crawler faces a problem of downloading a large number of pages from only a relatively small number of web servers (comparing to their overall number on the Web).

D. Shestakov is with the Department of Media Technology, Aalto University, Finland e-mail: denis.shestakov@aalto.fi

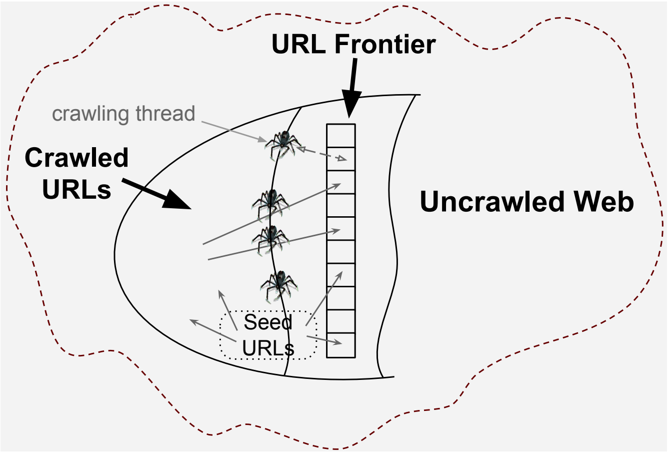

Fig. 1. URL Frontier in crawler's operations.

There are many applications with web crawlers playing a crucial role. The application spectrum ranges from visiting as many web pages as possible by web search engine or web archiver crawlers to the recently appeared trend of using crawlers for web application testing [3]. Needs of commercial web search engines are, however, the most important driving force in design and development of better crawler agents. With a few notable exceptions (e.g., see [4], [5]), academic crawling projects operate on a much smaller scale and apparently employ less sophisticated techniques.

The size and structure of the Web [6] are the most essential aspects that define several key requirements for a web crawler. The exponential growth of the Web suggests that no crawler can cope to cover all the information on the public Web [7], [8]. Moreover, the dynamism of web content guarantees that any collection of crawled documents is stale (not up-to-date) to a certain degree. As normally only limited resources are available, making crawls to be up-to-date involves a trade-off between freshness and coverage of the harvested documents. Similarly, the link structure of the Web [9] is crucial for understanding how crawlers can better prioritize their unseen URL lists.

B. Intelligent Web Crawling

A general-purpose web crawler typically operates in a distributed fashion, with multiple crawl threads that may run under different processes and often at different nodes. The architecture of a crawler [10], [11] includes a number of components, including the URL frontier. It keeps URLs to be visited in some order and returns the one with the highest score to a crawler thread when it seeks for a new URL. The URL frontier is schematically depicted in Figure 1.

There are a number of approaches to prioritize the URLs in the URL frontier. The main goal is to assign a URL some value that corresponds to the "importance" of a web page located at this URL. The URL prioritization strategies clearly depend on the crawling goals. E.g., if a crawler has no domain focus (general) or has to primarily focus on harvesting pages on a certain topic (topical). Another possible concern could be if a crawler should make a snapshot of a certain segment of the Web (batch) or should re-crawl previously visited pages (incremental). In general, one can categorize the existing approaches into six popular strategies used for both general and topical batch crawling: Breadth-First, Depth-First, Backlink count, Best-First, PageRank and Shark-Search [12], [13]. In essence, a crawling strategy defines the assignment of a priority value to a newly extracted URL. Depending on the strategy a number of factors can be taken into account – from a simple time-stamp of adding a link to the frontier to an inherited score value based on relevance scores of several ancestor pages pointing to a page with this link.

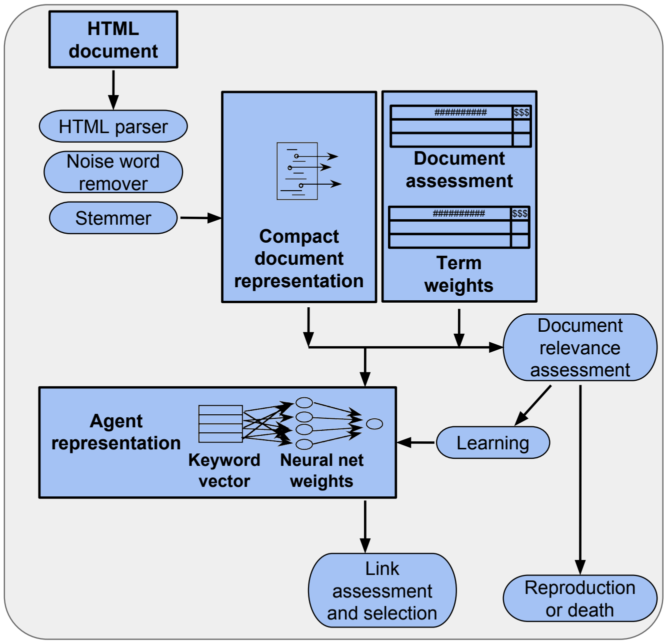

Fig. 2. Architecture of InfoSpider agent.

The abovementioned crawling strategies are static, in the sense that they do not learn from experience or adapt to the context of a topic in the course of crawl. In contrast to them, an intelligent crawler agent uses an adaptive learning model to assign priorities to the URLs in the frontier. In the literature, there exist at least three adaptive crawling approaches: InfoSpiders, ant-based crawling and HMM-supported crawling [14], [15], [16], [17]. While HMM-supported crawling utilizes Hidden Markov Models for learning paths leading to relevant pages, InfoSpiders and ant-based crawling are inspired by evolutionary biology studies and models of social insect collective behaviour correspondingly. Figure 2 shows the architecture of the InfoSpider agent, where agent's representation is supported by neural network.

III. Open Challenges

This section will briefly discuss the role of crawlers in the Web ecosystem and then present some open challenges in web crawling research, such as collaborative web crawling, crawling the deep Web and crawling multimedia content.

Being an important part of the Web ecosystem, crawler agents follow the pull model of resource access, under which a client has to first issue a request for a given resource (compared with the push model where a server can send (push) a content to a client without an earlier request from client-side). While the pull model has several advantages, it also leads to significant inefficiencies in crawlers' performance. The collaborative crawling or "crawling as a common service" approach [18] is the attempt to overcome some of these problems by supplementing a regular general crawler with a scalable filtering layer that allows other parties to crawl by setting conditions for documents of interest and obtaining relevant documents from the prime crawler.

The significant portion of the Web containing publicly-available information from myriads of online web databases (known as the deep Web [19]) is poorly accessible by crawlers. Accessing a deep web resource requires recognizing a search interface (search form) to a database and filling the recognized interface with meaningful values – both tasks are extremely challenging for conventional crawlers. In the literature, there are some relevant techniques for deep web crawling [20], [21], [22].

The Web has evolved from a huge textual repository to a fully-fledged multimedia platform serving web users all media types of content. Images, video, audio are now not just supplementing textual content of web documents but become integral part of many web resources. Most crawlers, however, do not adapt to this change and continue to operate as text harvesting systems. Thus, problems in crawling multimedia content [23], [24] are well-timed and of high importance.

IV. Supporting Materials

The material of this article was presented as a tutorial on the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT 2013) held in Atlanta, USA in November 2013. The tutorial slides are available at http://goo.gl/woVtQk; note that last part of tutorial provides relevant references to important crawl datasets and self-study materials. The bibliography for web crawling domain can be found in [1], [25].

References

[1] C. Olston and M. Najork, "Web crawling," *Found. Trends Inf. Retr.*, vol. 4, no. 3, pp. 175–246, 2010.

[2] D. Shestakov, "Current challenges in web crawling," in *Proc. ICWE 2013*, 2013, pp. 518–521.

[3] S. Choudhary, M. E. Dincturk, S. M. Mirtaheri, A. Moosavi, G. von Bochmann, G.-V. Jourdan, and I. V. Onut, "Crawling rich internet applications: The state of the art," in *Proc. CASCON 2012*, 2012, pp. 146–160.

[4] H.-T. Lee, D. Leonard, X. Wang, and D. Loguinov, "IRLbot: Scaling to 6 billion pages and beyond," *ACM Trans. Web*, vol. 3, no. 3, pp. 8:1–8:34, 2009.

[5] P. Boldi, A. Marino, M. Santini, and S. Vigna, "Bubing: Massive crawling for the masses," 2013. [Online]. Available: http://vigna.di.unimi.it/papers/P4.8.pdf

[6] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "The web as a graph: Measurements, models, and methods," in *Proc. COCOON 1999*, 1999, pp. 1–17.

[7] Z. Bar-Yossef and M. Gurevich, "Random sampling from a search engine's index," *J. ACM*, vol. 55, no. 5, pp. 24:1–24:74, 2008.

[8] D. Shestakov, "Sampling the national deep web," in *Proc. DEXA 2011*, 2011, pp. 331–340.

[9] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tompkins, and E. Upfal, "The web as a graph," in *Proc. PODS 2000*, 2000, pp. 1–10.

[10] V. Shkapenyuk and T. Suel, "Design and implementation of a high-performance distributed web crawler," in *Proc. ICDE 2002*, 2002, pp. 357–368.

[11] P. Boldi, B. Codenotti, M. Santini, and S. Vigna, "Ubicrawler: a scalable fully distributed web crawler," *Software: Practice and Experience*, vol. 34, no. 8, pp. 711–726, 2004.

[12] M. Najork and J. L. Wiener, "Breadth-first crawling yields high-quality pages," in *Proc. WWW 2001*, 2001, pp. 114–118.

[13] M. Hersovici, M. Jacovi, Y. S. Maarek, D. Pelleg, M. Shtalhaim, and S. Ur, "The shark-search algorithm. an application: tailored web site mapping," *Computer Networks and {ISDN} Systems*, vol. 30, no. 1-7, pp. 317–326, 1998.

[14] F. Menczer, G. Pant, and P. Srinivasan, "Topical web crawlers: Evaluating adaptive algorithms," *ACM Trans. Internet Technol.*, vol. 4, no. 4, pp. 378–419, 2004.

[15] A. Micarelli and F. Gasparetti, "Adaptive focused crawling," in *The Adaptive Web*, ser. Lecture Notes in Computer Science, 2007, vol. 4321, pp. 231–262.

[16] F. Gasparetti and A. Micarelli, "Swarm intelligence: Agents for adaptive web search," in *Proc. ECAI 2004*, 2004, pp. 1019–1020.

[17] S. Batsakis, E. G. Petrakis, and E. Milios, "Improving the performance of focused web crawlers," *Data & Knowledge Engineering*, vol. 68, no. 10, pp. 1001–1013, 2009.

[18] J. M. Hsieh, S. D. Gribble, and H. M. Levy, "The architecture and implementation of an extensible web crawler." in *Proc. NSDI 2010*, 2010, pp. 329–344.

[19] D. Shestakov, "Deep web: databases on the web," in *Encyclopedia of Database Technologies and Applications, 2nd edition*, 2009, pp. 581–588.

[20] J. Madhavan, D. Ko, L. Kot, V. Ganapathy, A. Rasmussen, and A. Halevy, "Google's deep web crawl," *Proc. VLDB Endow.*, vol. 1, no. 2, pp. 1241–1252, 2008.

[21] D. Shestakov, "On building a search interface discovery system," in *Proc. VLDB Workshops 2009*, 2010, pp. 81–93.

[22] Z. Behfarshad and A. Mesbah, "Hidden-web induced by client-side scripting: an empirical study," in *Proc. ICWE 2013*, 2013, pp. 52–67.

[23] A. Nanopoulos, D. Rafailidis, M. M. Ruxanda, and Y. Manolopoulos, "Music search engines: Specifications and challenges," *Information Processing & Management*, vol. 45, no. 3, pp. 392–396, 2009.

[24] J. Hare, D. Dupplaw, W. Hall, P. H. Lewis, and K. Martinez, "Building a multimedia web observatory platform," 2013. [Online]. Available: http://eprints.soton.ac.uk/352465/

[25] (2013) Web Crawling (Mendeley Group). [Online]. Available: http://www.mendeley.com/groups/531771/web-crawling/