ROYAL HOLLOWAY UNIVERSITY OF LONDON

CSPRC 2015

Abstracts

15th Computer Science Postgraduate Research Colloquium

Juan José Cáceres Silva Horacio Caniza Vierci

Royal Holloway University of London
Department of Computer Science

11th May 2015

Contents

Poster presentations

Preface

Welcome to the 15th Computer Science Postgraduate Research Colloquium (2015).

This colloquium serves as a forum for interaction between the faculty, staff and students engaged in the various different disciplines of research in Computer Science at Royal Holloway.

As last year, this year's event is the result of a staff-student collaboration, and has been coordinated by the students. The work presented showcases the innovative ideas, diversity and excellence of postgraduate research that is conducted here, within the Department of Computer Science at Royal Holloway University of London.

This year, for the first time, we have an invited speaker, Dr. Misha Dashevskiy. Dr. Dashevskiy is a former PhD student of our Department who currently works as an Analyst in the City of London for UBS. We hope to keep inviting alumni and make this a tradition, in order to provide a forum for past PhD students to share their experiences and provide us with feedback about how their time at Royal Holloway contributed to their working life.

We have organized the colloquium into three themes:

- Algorithms and Bioinformatics
- Machine Learning
- Types, Logic and Language Engineering

We hope you enjoy the day, and we thank you for taking part.

11th May 2015
Royal Holloway

Juan
Horacio

Invited Talk

Quantitative Analyst in Electronic Trading

Dr. Misha Dashevskiy

Misha Dashevskiy is a quantitative analyst, working on Foreign Exchange, Rates and Credit Electronic Trading desk at UBS investment bank. He holds a PhD in Machine Learning from Royal Holloway, University of London and a degree in Mathematics from Moscow State University.

Parameterized and Approximation Algorithms for the Load Coloring Problem

Bin Sheng

How to cope with NP-Complete problems? The Fixed Parameter Tractability theory came up as an alternative approach in the past decades. In the FPT setting, an instance comes with a parameter k, it is said that the problem is fixed parameter tractable if there is an algorithm running in time $f(k)\,poly(n)$. It is well-known that a parameterized problem L is FPT if and only if it is decidable and admits a kernelization. Due to applications, low degree polynomial size kernels are of main interest. The FPT method has achieved great success. Here we apply it to the Load Coloring problem.

Let c be a positive integer. Given a graph $G = (V, E)$ and a nonnegative integer k, the c-Load Coloring Problem (denoted c-LCP) asks whether there is a c-coloring ϕ: $V \to [c]$ such that for every $i \in [c]$, there are at least k edges with both end vertices colored i. Gutin and Jones (IPL 2014) studied this problem with $c = 2$. They showed 2-LCP to be fixed-parameter tractable (FPT) with parameter k by obtaining a kernel with at most $7k$ vertices. In this paper, we extend the study to any fixed c by giving both a linear-vertex and a linear-edge kernel. In the particular case of $c = 2$, we obtain a kernel with less than $4k$ vertices and less than $8k$ edges. These results imply that for any fixed $c > 2$, c-LCP is FPT and the optimization version of c-LCP (where k is to be maximized) has an approximation algorithm with a constant ratio.

PDB-Hadoop: A framework for parallelising user applications on the Protein Databank using Apache Hadoop

Jamie Al-Nasir

We present a framework that facilitates parallel execution of protein structure analysis tools to be carried out on the entire (or subsets of) the Protein Databank (PDB) using the Apache Hadoop platform. Our design enables structural Biologists to use the Hadoop platform without having to explicitly write Map-Reduce code. It is easily scalable and uses a mapper architecture that functions on a stand-alone basis or can be extended to include further Map-Reduce operations.

Hadoop is widely available, increasingly stable and is available on cloud computing platforms such as AWS. It is also optimised for the analysis of large data sets made up of semi-structured data such as the PDB. Nonetheless, the Map-Reduce formalism, typically implemented in Java, is not trivial to understand for non Computer Scientists. Hence, there is a need to facilitate structural Biologists to use Hadoop without having to write bespoke code to make use of it.

The talk will present comparisons between running equivalent jobs using the OpenLava batch scheduler with PDB-Hadoop on the same cluster. This will highlight

performance increases when using PDB-Hadoop for structural calculations jobs and molecular docking of a putative oligopeptide ligand with entries in the protein databank.

PDB-Hadoop is an efficient and scalable framework for the concurrent execution of code utilising Apache Hadoop which does not require the users to re-write their applications according to the Map-Reduce formalism. We believe performance gains observed are a result of the efficient use of concurrency by YARN (Yet Another Resource Negotiator).

Measuring similarity of Mendelian diseases through ontological analysis

Horacio Caniza

Over the past few decades advances in molecular biology greatly increased our understanding of disease. Our perspective on them has now evolved from simple causal relationships linking a single gene to a disease to a more complex relationship linking biological network modules (possibly containing many genes) to diseases. In this context a disease is consequence of a perturbation in the underlying biological networks. For about 30% of hereditary diseases no disease gene is currently known. For these orphan diseases, very little (often nothing!) is known about their molecular basis.

In this talk, I will present a method to quantify similarity between heritable diseases at molecular level. In other words, our method provides a number that accurately quantifies the distance between any pair of disease modules in the interactome. This measure can help pinpoint the location of the molecular perturbations for the thousands of orphan diseases, and more generally, enable the transfer of knowledge between similar diseases. It provides hypotheses for causal genes discovery and even suggestions for drug repositioning.

The measure brings together information that is scattered across the vast corpus of biomedical literature. Therefore it is applicable to virtually every heritable disease which has been described so far in the literature. Our method exploits the terms and the structure of the MeSH ontology. We prove that sets of MeSH terms are highly effective at producing descriptive and well-structured annotations for hereditary diseases in OMIM, and that the structure of MeSH can be exploited to accurately quantify disease similarities.

Through a a machine learning approach we show that our measure is capable to accurately predict molecular-level relationships between diseases. We also provide a highly illustrative set of examples from recent medical literature that showcase the accuracy of the method. Importantly, by comparing the current version of OMIM with a two-year old one, we show that our measure can be used effectively for the prediction of candidate disease genes.

We have also developed a web application to query more than 28.5 million relationships between 7,574 hereditary diseases (96% of OMIM) based on our similarity measure.

A logic-independent study of Herbrand's Fundamental Theorem

Ionuţ Ţuţu

The Fundamental Theorem of Herbrand is a central result in proof theory that deals with the reduction of provability in first-order logic to provability in propositional logic. Its importance in the context of automated theorem proving was realized in the early 1960s, when, in combination with the theory of Horn-clause logic, it played a key role in establishing the mathematical foundations of logic programming. In the conventional setting of relational first-order logic, Herbrand's theorem states that, given a set of Horn clauses (a logic program), the answers to an existential query can be found simply by examining a term model the Herbrand model instead of all possible models. Over the last three decades, the original result has been generalized to a variety of other logics, including Horn-clause logic with equality, hidden algebra, and category-based constraint logic, culminating with an investigation of the theorem in an arbitrary institution a categorical formalization of the intuitive notion of logical system put forward by Goguen and Burstall in the late 1970s.

Thanks to its generality, the institution-based approach to Herbrand's theorem enabled the development of logic programming over a wide range of formalisms (e.g. hybridized logics). Even so, recent developments have shown that the institutional approach cannot capture constructions that arise when service-oriented computing is presented as a form of logic programming. This prompted the need for a new perspective on the theorem founded instead upon a concept of generalized substitution system, which extends institutions by allowing for direct representations of variables and substitutions.

In this talk, we survey the connection between the institution- and the substitution-system-based approach to logic programming by investigating a number of features of institutions, like the existence of a quantification space or of representable substitutions, under which they give rise to suitable generalized substitution systems. Building on these results, we further show how the original institution-independent versions of Herbrand's theorem can be obtained as concrete instances of a more general substitution-system-based result.

Formalizing the Relation Between Subsumptive and Coercive Subtyping

Gerogiana Lungu

Subsumptive subtyping is widely used in Computer Science and Mathematics, naturally having the very important consequence that, for a subtyping relation between two types, wherever an object of the supertype is expected, one can use an object of the subtype. However, the subsumption rule stating that any object of a subtype is also an object of the supertype, is not an acceptable concept in some theoretical frameworks, particularly in a type theory with canonical objects over a logical framework. An alternative approach to obtain the consequence previously mentioned was to introduce coercive subtyping.

Intuitively it is straight forward that subsumptions are a particular case of coercions. The question we aim to answer in our work is whether we can formalize this intuition. We answer this question by developing a new calculus that deals with coercive subtyping judgements by means of signatures and we show that a calculus for subsumptive subtyping can be faithfully embedded in it, hence offering a means to understand this concept and its effects in a setting used for example in Martin Löf Type Theory or the Unifying Theory of dependent Types(UTT).

Defining Parser Combinators with Attribute Grammars *

Thomas van Binsbergen

Parser combinator libraries are popular tools for writing parsers in functional programming languages. Parsers written using parser combinator libraries are easily integrated in projects and offer strong type guarantees on the semantic actions applied to parse results. In addition, combinator libraries are modular, compositional and the parsers reusable. From elementary parser combinators more complicated parser combinators can be built that capture higher-level patterns common in, for example, EBNF specifications and patterns such as operator precedence and associativity.

Attribute Grammars (AGs) are a popular formalism for defining programming language semantics in a simple and modular fashion. A grammar is used to represent a set of trees. Tree nodes are augmented with attributes to collect all sorts of information on the tree. Finally a description of how the attributes are computed needs to be given, relying on other attributes and leaf nodes of the tree. Only those attributes that are crucial to the semantics of the language need to receive a definition. The definition of the logistical attributes can be inferred.

In this talk I connect these two topics by observing that the elementary parser combinators form a grammar for "the language of all grammars". I give semantics to this language such that the AG evaluator generated for the semantics behaves like a parser. In other words, I show that the definition of the elementary parser combinators can be generated from an AG description.

The talk concludes with examples of helper combinators and combinators for higher-level patterns, that are derived from previously defined combinators.

*****_This work is also presented as a poster._

The technology affordances of voted discussion forums: a way to provide more trustworthy health information online?

Jennifer Cole

Discussing and debating a subject or topic facilitates Social Learning and better absorption of knowledge. Rowntree (1995) described Message Boards and Bulletin Boards, the precursor of todays online discussion forums such as Reddit and Mumsnet, as a creative cognitive process of offering up ideas, having them criticised or expanded on, and being able to reshape ideas in the light of peer discussions. Thomas (2002) has also identified the Conversational Model of Learning to be applicable to online forums.

These observations are particularly relevant to the discussion of health issues in online forums. Internet users who search for health information online are more likely to communicate with doctors, talk with others, and seek more information on their condition (Menon et al, 2002; Huh et al, 2005), potentially leading to better health outcomes. In the UK, however, only 37 per cent of adults say they search online to find information about health related issues, compared with 66 per cent who shop online. Groselji (2014) suggests that online health forums in particular are an underused platform for obtaining health information.

The biggest challenge to health-seeking information online is lack trust in the quality of the information, due to lack of reliable indicators of quality. Difficulty in identifying accurate information makes internet users less likely to trust it, as the risks of acting on inaccurate health information are high (Luo and Najdawi, 2004). Academics writing as far apart in time as Impicciatiore (1997) and Whitelaw (2014) have suggested that rather than health information online being incorrect or inaccurate, it is more often the case that online information assessed to be of poor quality is incomplete or biased. In this regard, discussion forums, and in particular voted discussion forums, offer a technology affordance beyond that of other sites characteristics: posters can request further information where they feel it is needed, add to previous postings where they do not feel enough information has been given, and can actively counter information they believe to be incorrect or biased. In the latter case, as more posters join the debate, voting systems can enable the view with the strongest consensus to rise to the top.

This study is analyzing the health information posted in online discussion forums, and the discussions that take place around them, to determine if the process results in health information that is more likely than not to be accurate, complete and balanced.

Learning War through Natural Language Processing

Andrej Žukov Gregorič

Thousands of news articles about conflict-related incidents are published each day. However, these reports on the deaths and injuries of civilians and combatants usually go unnoticed. This adds to the confusion surrounding every war. Accurate casualty reports are produced only years later - much too late to be of any use during conflict.

To establish such facts one must first extract relevant information from individual articles and then discover how these articles relate to the incident. We show how extracting the relevant information can be done using Conditional Random Fields coupled with domain-specific features we provide. We quickly review new advances in Natural Language Processing and propose other ways of improving such models irrespective of domain. We finish by exploring how articles may be related to individual incidents using clustering.

On-line learning with indoor environments

Khuong Nguyen

We spend most of our times indoor, where limited or no GPS service is available. Latest approaches that use built-in indoor infrastructure such as WiFi are struggling to cope with the harsh conditions of the indoor environment to provide fine-grained tracking.

Recently, we have found specific extra information from the indoor environment to counter such challenges. In this talk, we will discuss our latest experiments with the magnetometer measurements on the smart phones to provide robust additional information for indoor positioning. At the heart of our system is machine learning, which predicts the user's intended path in advance based on his current place. Our system constantly learns from the real time data obtained by the user, and updates the training database accordingly. We show that such additional information greatly enhances the positioning accuracy of our system, compared to current state-of-art approaches using just the WiFi data.

Conformal Anomaly Detection of Trajectories with a Multi-Class Hierarchy

James Smith

Anomaly detection is a large area of research in machine learning and many interesting techniques have been developed to detect abnormal behaviour of objects. In this presentation we address the problem of anomaly detection in the maritime trajectory surveillance domain. Conformal predictors are a useful machine learning technique to produce reliable predictions, allowing us to control the false-positive rate. We propose the use of a multi-class hierarchy framework for anomaly detection using conformal predictors. This approach utilizes different class representations, attempting to address the question of which class representations work best. Experiments are conducted with real-world data taken from shipping vessel trajectories obtained through AIS (Automatic Identification System) broadcasts and the results are discussed.

Sample stage classification in lymphoma mutagenesis and gene analysis in a Human-Mouse interactome

Juan José Cáceres

This work presents an approach for the classification of lymphoma growth state and selection of putative genes that cause or sustain lymphomas through mutagenesis in mice. The execution of these tasks is part of a research collaboration with Anthony Uren, Head of Cancer Genomics group at the MRC Clinical Sciences Centre. The collaborator performs wet lab experiments and provides the data used in the analysis.

In order to identify the lymphoma growth state, some measures were established over the profile of the clonality among the most mutated Common Insertion Sites (CIS). This provided a way order the samples and select the most relevant CIS from the late stage tumours. Since the tumours are obtained through mutagenesis, the insertion clonality on the CIS, that contain genes involved in the lymphomagenesis are expected to be higher than those that are not involved in the process.

The following tasks were also completed in order to get interesting sets of genes for further analysis. First, lists of co-expressed and mutually exclusive genes were extracted and ranked according to interactome closeness. Second, the CIS genes whose clonality was statistically different, regarding their source genotype, were selected and ordered according to their distance from BCL2. Finally, to obtain a set of putative genes, a set of known cancer genes was selected in the interactome and the CIS genes were tested for statistical significance on their connectedness to those known cancer genes.

Hues of Satisfaction: Many-valued Logics for Constraint Specification

Claudia Chiriţă

Service-Oriented Computing (SOC) is a recent computational paradigm focusing on the development of software applications based on dynamically changing networks of systems. In short, SOC builds upon a need-fulfilment mechanism through which applications connect to external suppliers every time a need for services appears. The requester first needs to discover the system components that guarantee, through an interface, the fulfilment of its conditions, and then select and bind to a provider. The interfaces describe properties that do not depend on the actual implementation of the provided services, like functional properties of their input-output behaviour. These are often defined using algebraic specifications of abstract data types or temporal specifications. In addition, one could also specify constraints expressing preferences meant to be used for the selection of a best provider in terms of the maximisation of their satisfaction degrees.

We advance a general technique for enriching logical systems formalised as institutions with soft constraints, leading to a concept of multi-valued institution suitable for specifying soft constraint satisfaction problems. In this way, we can formally describe preferences as soft constraints, magnitudes of satisfaction as constraint satisfaction, and finding best solutions (with respect to a set of preferences) as constraint optimisation. The approach to constraint satisfaction that we propose here generalises both the c-semiring approach (SCSP) and the valued CSP approach (VCSP), by making use of residuated lattices, which offer a unifying truth structure for both SCSP and VCSP. This allows us to base our methodology for selecting the most promising provider of a given resource in the context of service discovery on the concept of graded semantic consequence.

We further study the dynamics of constraint satisfaction that follows from the non-mereological composition of specifications in the context of networks of systems: we formalize the changing of preferences during the development of a system by adapting our framework to allow the composition of specifications based on different truth structures. As a case study, we show how the resulting framework can be used to model Free Jazz performances as complex systems.

Examining protein family diversity in metagenomics data

Golestan Sally Radwan

Metagenomics is the study of the genomes of entire microbial community samples obtained directly from the environment and with little knowledge as to what species might be included in each sample. Existing research on metagenomics focuses primarily on identifying known species within the sample and, in some cases, study their relative abundance under certain (normally artificial) stresses such as chemicals or other pollutants. It mostly uses existing tools and algorithms developed for other -omics disciplines, sometimes slightly modified to fit a metagenomics pipeline. This research takes a different view in many ways. First and foremost, it starts with an environmental/biological question that needs to be answered, namely: how do different stresses, natural or artificial, affect entire microbial communities, and what does this say about the resilience and adaptability of those communities? Secondly, it attempts to innovate in the way research is conducted, namely by discarding typical -omics workflows and focusing on finding so-called ”signatures” within protein families, which can then be applied to the sample at hand, helping to draw a ”functional profile” of the entire community without a need for species identification. This helps eliminate many of the sources of inaccuracy currently plaguing metagenomics data analysis. Finally, this research aims to find a standardised method which can be applied to different types of microbial communities, different types of stresses, and different levels of data quality and sequencing methods. Potential application areas are quite diverse and include research into: climate change, origin and evolution of life, astrobiology, antibiotics resistance, and personalised medicine, to name a few. This poster shows the current state of research after the equivalent of a few months full-time, and will continue to progress over the next few years. The data currently in use is high-quality, assembled data from the Human Microbiome project, curated by the NCBI. As the research progresses, more data samples will be used from diversified sources including, ultimately, sampling and sequencing in the field.

Extensible and Reusable Architecture for Software Product Lines

Ahmed Mohamud

The increases of software users and technological improvements have increased the expectations to handle more difficult problems on a large and growing scale. The complexity, safety and reliability of software have also increased and play a major role in many industries such as the aviation industry, the health care industry, the financial sector and many other industries. Increasing the complexity of a system, while ensuring system reliability and safety, forced modern system structure and its architecture to define the future success and evolution of their system.

For example, a successful health care service involves the exchange of the most accurate health information about the patient when needed. Providing a health care service requires communication and is crucial to the patient's safety and well-being. A simple patient care and management involves the contribution of different specialisms and processes with each their own sets of regulations and guidelines. Collaboration and integration are important to ensure the best care is offered by the different specialisms in a real-time patient-centric environment. However, health care and medical devices require rigours testing and verification. They are often provided by specialist firms and software houses focused on the delivery of products. For example, a laboratory may use industry standard machinery such as HiSeq 3000 from the firm Ilumina, for sequencing blood-results. The same laboratory may also use a software supplied by ClinSys Group to help analyse and track tissues and samples. This brings challenges in integrating and streamlining systems with other systems while ensuring the system architecture reflects **the extensibility and re-usability** of the system.

It is therefore crucial that the underlying architecture supports the separation between the data of the different heterogeneous systems and their interactions. Often regulatory reporting, clinical reviews and laws and legislations impact the way hospital systems are interacting and collaborating. This brings new challenges in the separation of decisions that affect the different system components and its underlying data. It is important that the software architecture supports the extensibility and re-usability of the system without the need to **rewrite an entire application, code or package**.