Institute for Big Data Analytics

To create knowledge and expertise in
the field of Big Data Analytics by facilitating
fundamental, interdisciplinary and collaborative
research, advanced applications, advanced training
and partnerships with industry.

Research - Training - Outreach

Newsletter

Autumn 2014

Message From the Director

We are glad to share with you the first issue of the newsletter of the Institute for Big Data Analytics. We would like to use this publication to stay in touch and to share with you Institute news, projects, and accomplishments. We are proud to be 20 months old, and – as befits this age – we are not crawling anymore, but we walking and being more and more adventurous in our research and outreach. We are happy to be able to report a number of successes of our colleagues and students. We are very proud of the excellent performance of our team in the Microsoft's Entity Recognition and Disambiguation Challenge – congratulations to Marek and Arash! We are also very happy to announce that we have won an award in the NSERC CREATE competition for our proposal titled TRIBE: Training in Big Text Data (we have partnered with Simon Fraser University and Université de Montréal) – we were one of fifteen proposals chosen from the initial pool of 120. We will soon be announcing details of this graduate program and recruitment information on the Institute webpage www.bigdata.dal.ca. We have attracted significant interest from the media and from different levels of government, as well as from the private sector in Atlantic Canada and beyond. We are also expanding our collaboration with a variety of research groups within Dal – from marine ecology to anesthetics and post-operative care. Lest but not least, we a were able to secure access to a number of exciting BIG data sets, eg in the area human mobility analytics.

We will be in touch again in the Spring of 2015. Please do not hesitate to contact us with anything related to big data at bigdata@cs.dal.ca.

Marek Lipczak and Arash Koushkestani's Tulip Takes First Prize in Data Challenge

When the research labs of Google, Microsoft and Yahoo team up to organize a challenge to build a better system for Entity Linking you know that researchers around the world are going to take note. In April of this year the "Entity Recognition and Disambiguation Challenge" went out and 36 teams from North America, Europe and Asia signed up for the competition. There were two strands: one in search queries and the other in web documents. The Tulip system, built by Marek Lipczak and Arash Koushkestani beat out the competition and was awarded first prize in the web documents strand at the SIGIR conference in July.

Entity Recognition and Disambiguation is also known as "entity linking", the goal of which is to find mentions of entities (like Barack Obama) in text and linking them to an external knowledge base, typically Wikipedia. Although the goal is straightforward, the problem is not a simple one for either human beings or machines, and has been addressed by other researchers in the past. Human beings are limited by both the amount of data they can process and the tendency in reading to skim over and miss entities. Machines can

DALHOUSIE UNIVERSITY
INSTITUTE FOR
BIG DATA ANALYTICS

News

cope with large data sets but struggle with language ambiguity and an overwhelming number of options. The challenge in overcoming this problem is to use context in a smart way. Rather than bringing an existing system to the problem Marek and Arash started from scratch with fresh eyes and fresh ideas, and created a system which is both simple and fast (200 news articles per minute) and achieves an accuracy of between 0.7 and 0.8 (F1 score). Up against other teams who were using existing systems Marek and Arash felt that the fresh approach was one of the sources of their competitive advantage - their Tulip system represents a new stream in this field of research, solving the problem in a new way and creating more opportunities for future development. At the start of the challenge, Tulip was scoring near the bottom of the pack in its performance but their modifications and developments propelled the system rapidly up the leader board, overtaking its competitors in just a few months.

Although the prize money was not huge – a $500 cheque from Google – it is their success in developing a new idea to outperform teams from from around the world that puts a smile on Marek and Arash's faces. They have presented a paper on their work at the workshop following the challenge, and have plans for further publications. Their intentions are to develop their system as open-source technology.

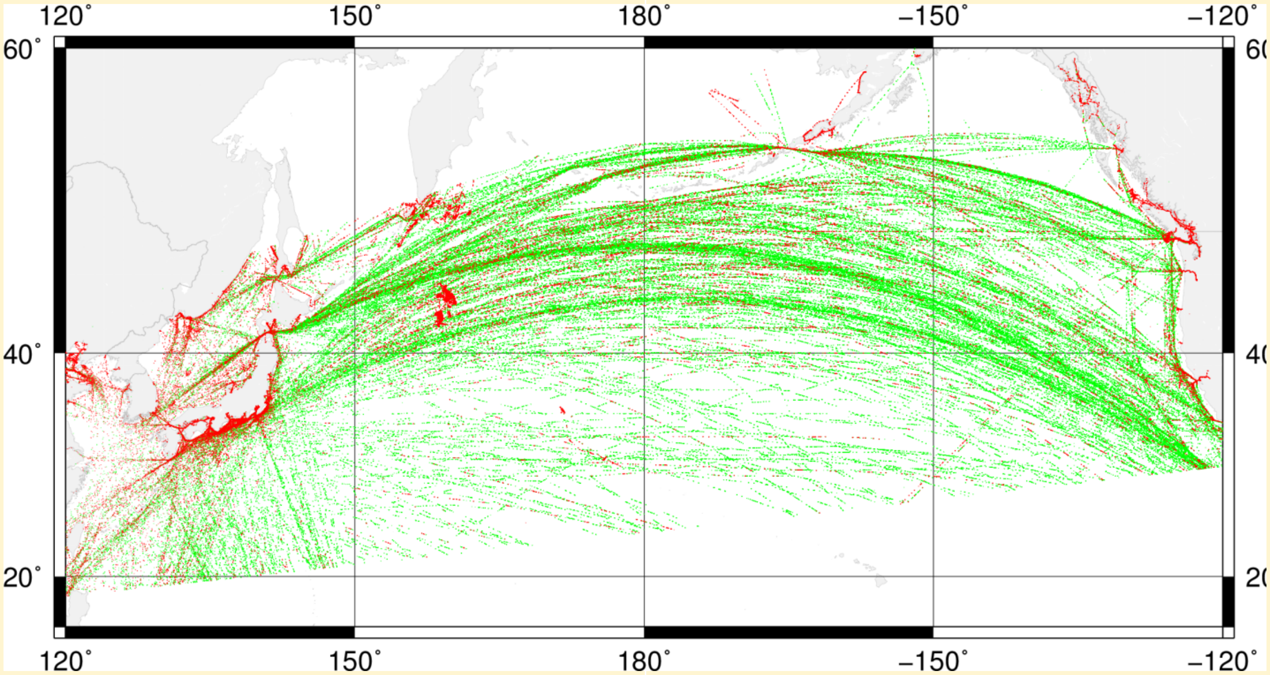
Marek Lipczak holds a PhD in computer science from Dalhousie and is currently working as a research associate with Evangelos Milios. Arash Koushkestani is student in the MCS program, supervised by Dr. Milios.

Big Data May Reveal Fishery Patterns

In March of this year, the Institute for Big Data Analytics assembled a group of people from a variety of research areas across the university who might have an interest in studying data about ship movements. One of the outcomes of those meetings was an exploratory collaboration with researchers in the Biology Department who have an interest in commercial fisheries. The collaboration aims to study data from the Automatic Identification System (AIS), which is an automatic tracking system used on ships for identifying the location of vessels. This AIS data should reveal whether vessels are engaged in fishing activity as compared to other activities, based on characteristic patterns of movement. In analyzing this data, the study findings may reveal patterns of overfishing which would be of interest to environmental institutions aimed at regulating fishery activities around the world.

Some preliminary results, focusing on the North Pacific look promising (see the diagram below), although the team still needs to validate the data. If they can confirm that the picture of red dots seen in the image below do indeed correspond to fishing activities then the next step will be will be to build a model for the whole globe.

The team involved in this study are Bo Liu, Erico N de Souza (PhD), and Behrouz Haji Soleimani from the Big Data Analytics Institute and Kristina Boerder from the Biology Department. The diagram below presents preliminary results of the Machine Learning algorithm developed by the Big Data Analytics Institute.



Fishing patterns detection. Green lines indicate non-fishing activity.
Red areas indicate fishing activity.

Training in Big Text Data Through NSERC's CREATE Program

The Institute for Big Data Analytics is pleased to announce a limited number of fully-funded, competitive M.Sc. and Ph.D. scholarships for work in Big Text Data in an industrial stream graduate training and research programme created jointly by Dalhousie, Université de Montréal and Simon Fraser University.

Big Text Data involves collections of texts in natural language (e.g. article abstracts, emails, tweets, blogs, medical reports) that are so large and complex that they are difficult to process and digest using traditional text analytics techniques. On the one hand, Big Text represents additional challenges with respect to Big Data, as the information is unstructured and presented in (often simplified or incorrect) natural language. On the other hand, Big Text is ubiquitous in business, science, and everyday life – estimates indicate that 80% of Big Data is in fact text data. There is therefore a strong industrial interest in working on Big Text problems, and a significant need for Highly Qualified Personnel trained in Big Text Analytics techniques.

Our programme is based on five pillars: (1) a structured curriculum combining the necessary know-how, including "soft skills", in demand by industry, into a novel, specialized Graduate Certificate; (2) hands-on training in the form of an industrial internship; (3) national and international student mobility and exchange; (4) a bilingual and multi-lingual application and training environment; (5) respect for privacy as a value instilled in our students and in our approach to Big Data. The combination of these pillars will provide unique, industrially-relevant graduate training of HQP in an area of unmet high demand in Canada and globally. Students who will fulfil all the course and internship requirements will obtain a Graduate Certificate in Big Text Analytics.

The objective of the programme is to train a strong group of Canadian graduate students in the area of text mining, focusing on large data collections and streaming data. Training related to the specifics of text data is particularly important, as standard Big Data curricula do not prepare students to cope with the additional complexities and opportunities of text data. Text data is particularly important in the Canadian context, due to the economic effects globalization particularly affecting Canada; it is important to the export nature of our economy, and to the multi-cultural character of our society. The training will take place in a research-intensive environment at three large Canadian Computer Science departments; and in the form of internships with Canada's leading high-tech companies. Students will acquire research experience through their interaction with the teaching faculty who are all active researchers, and through their research project courses supervised by those faculty members.

As the focus is on text data in natural language, there is an opportunity to train at least some of the students in the use of bilingual and multi-lingual resources in select Big Text tasks – a truly unique and Canadian proposition. We mean by this exposing the students to the challenges and opportunities of work with bilingual text data by teaching text mining techniques specific to bilingual and multilingial text corpora. Several of the co-applicants are bilingual, and the participation of the Université de Montréal team further strengthens this aspect of the project. The multilingual aspect of the data and of the training will be further strengthened by a collaboration with a Brazilian University.

This programme is funded by the NSERC Collaborative Research and Training Experience (CREATE) initiative. Interested and qualified applicants should send their brief academic history, including list of courses, and marks obtained,

Federal Science Minister Ed Holder visits the Institute

The Institute for Big Data Analytics was privleged to receive a visit on October 17 from Mr. Ed Holder, Minister of State for Science and Technology. Mr Holder took a tour of the Institute, spent time talking to students about their research and listened to a presentation by Dr. Stan Matwin. In the picture he is discussing research with Eman Alyami, Interdisciplinary PhD student. Mr. Holder seemed impressed by the diversity of research projects and affirmed the importance of this field to innovation and economic prosperity. Although he had come to Halifax for a different meeting, the visit to the Institute was the one diversion in his trip that he insisted on making.

Director

Dr. Stan Matwin PhD, CRC

Executive Committee

Dr. Michael Bliemel
Faculty of Management (representing the Dean of Management)

Dr. Tom Marrie
Dean of Medicine

Dr. Evangelos Milios
Associate Dean, Research, Faculty of Computer Science

Dr. Andrew Rau-Chaplin
Faculty of Computer Science

Dr. Nur Zincir-Heywood
Faculty of Computer Science

Dr. Stan Matwin
Director, (ex officio), Faculty of Computer Science

Advisory Board

Sue Abu-Hakima Amika
President and CEO,
AmikaMobile, Ottawa, ON

Martin Davis
Vice-President of IT, JD Irving,
Moncton, NB

Peter Hickey
Co-founder, Oris4, Halifax, NS

David Kasik
Senior Technical Fellow,
The Boeing Company, Seattle,
Washington

Stephen Perelgut
IBM Canada University Relations
Manager, Markham, ON

Newsletter Editor:
David Langstroth
dll@cs.dal.ca

"Without big data analytics, companies are blind and deaf, wandering out onto the web like deer on a freeway."
- Geoffrey Moore, author and consultant