The Center for Dynamic Data Analytics Quarterly Newsletter

Edition: Second Quarter, 2015 | Published May 8, 2015

Announcements

- **Big Data Regional Innovation Hubs**: The CDDA participated in the Northeast Regional Charrette on April 17th in Wakefield, MA. The purpose of the meeting was to help map solutions in the establishment of a national network of four Big Data Regional Innovation Hubs (Northeast, Mid-west, South and West). If you wish to participate in the online Hubs community, please visit bdhub.info or for a description of the Hubs initiative, click here or contact James Mielke.

- The CDDA **Spring 2015 Workshop and IAB Meeting** is scheduled for **Tuesday, May 12th** and will be held at Stony Brook University in Stony Brook, NY. Please register here.

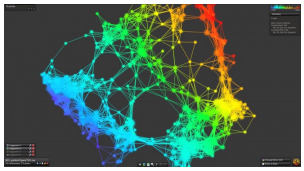
- The **University of Virginia** has submitted a formal Planning Grant proposal to join the CDDA. Dr. Peter Beling will serve as Director of the CDDA site at UVA. For more information, please contact Peter Beling (pb3a@virginia.edu) or James Mielke (james.mielke@rutgers.edu).

- **NIST** has issued a draft "Interoperability Framework," defining data analytics-related phrases, outlining management templates and describing common use cases for large data sets and other large amounts of information. NIST will be accepting comments thru May 21st. Learn more here.

Current Projects

- Tissue Quantification Project: PI, Dimitris Metaxas, IAB Collaborators, Colin Miller and Hui Jing Yu, BioClinica
- Anomaly Detection in Dynamic Networks: PI, Leman Akoglu, IAB Collaborator, Steve Cento, Northrop Grumman Aerospace Sector
- Remote Volume Rendering Pipeline for mHealth Applications: Researcher, Ievgeniia Gutenko, IAB Collaborator, Ron Cha, Samsung Research America
- Toward Automated Discovery of Artistic influence: PI, Ahmed Elgammal, Rutgers
- Big Graph Mining: PI, Tina Eliassi-Rad, Rutgers
- Privacy Preserving Data Mining: PI, Jaideep Vaidya, Rutgers
- The Reality Deck - 1.5 Gigapixel Display: PI, Arie Kaufman, Stony Brook
- 4D Cardiac Fluid Flow Modeling: PI, Dimitris Metaxas, Rutgers
- Optimal Bidding Strategies in Sequential Auctions: PI, Michael Katehakis, Rutgers
- Exploring the Role of Gaze Behavior in Video Annotation: PI, Dimitris Samaras, Stony Brook
- Volume-specific parameter optimization of 3D local phase features for improved extraction of bone surfaces in ultrasound: PI, Ilker Hacihaliloglu, Rutgers
- Scalable Parallel Processing Algorithms for Sequence Alignment and Assembly: PI, Song Wu, SB
- Rutgers Wellbeing Study: PI, Vivek Singh, Rutgers
- The SILK Project: Semantic Inferencing on Large Knowledge: PI, Paul Fodor, Stony Brook
- Crowd Simulation, Analysis, and Optimization: PI, Mubbasir Kapadia, Rutgers
- Behavioral Modeling and Prediction with Wearable and Mobile Devices: PI's, Chirag Shah and Vivek Singh, Rutgers

Contents



Iris, Ayasdi's data-visualization tool, finds connections in abstract data sets

"Many things which cannot be overcome when they are together, yield themselves up when taken little by little."
– (attribution disputed) Sertorius or Plutarch

"Geoflow" for Excel: 3D Big Data Visualization Built on Bing Maps

Publications

- Less is More: Building Selective Anomaly Ensembles
- Guilt-by-Constellation: Fraud Detection by Suspicious Clique Memberships
- Existence of Periodic Fixed Point Theorems in the Setting of Generalized Quasi-Metric Spaces
- Factorization of View-Object Manifolds for Joint Object Recognition and Pose Estimation
- Rules on the Web: From Theory to Applications
- Predictive sparse modeling of fMRI data for improved classification, regression, and visualization using the k-support norm
- Lightweight practical private one-way anonymous messaging
- Multi-armed Bandits under General Depreciation and Commitment
- Defeasibility in answer set programs with defaults and argumentation rules
- Unique in the Shopping Mall: On the reidentifiability of credit card metadata

Big Data News

- Big Data's Dark Side
- Learning to See Data
- Five Emerging Ways to Analyze Unstructured Data
- The Elements of Data Analytic Style
- IoT Is About Analysis, Not Things
- IBM Watson: 10 New Jobs For Cognitive Computing
- 7 Attitudes That Kill Big Data Projects

Upcoming Conferences

- 5/12/15: Stony Brook, NY—CDDA Spring 2015 Workshop and IAB Meeting
- 5/15/15: Austin, TX—DisrupTech: Analytics, Big Data & Beyond
- 5/18-20/15: Boston, MA—Alteryx Inspire 2015
- 5/19-21/15: Stanford, CA—XLDB 2015, 8th Extremely Large Databases Conf. & Wrkshp
- 5/20-21/15: Chicago, IL—Business Analytics Innovation Summit
- 5/26-27/15: Boston, MA—Re-Work Deep Learning Summit
- 6/8-11/15: Chicago, IL—Predictive Analytics World
- 6/15-16/15: New York, NY—14th Text Analytics Summit East
- 6/22-24/15: Bristol, UK—Dynamic Networks and Cyber-Security
- 7/6-11/15: Lille, France—International Conference on Machine Learning (ICML)

Featured Publication - Rain or Shine? Forecasting Search Process Performance in Exploratory Search Tasks

Dr. Chirag Shah

Abstract— Predicting how people perform in their information search processes is a hard problem. The prediction problem becomes even more complex when considering exploratory searches. Exploratory search is typically described as open-ended and multifaceted, where goals may be unclear and there may be none or multiple satisfying answers. Searchers engage in exploratory search when they commence researching a new topic, when they form a new idea, in problem identification, and other forms of information seeking for creative discovery.

While the modern information retrieval (IR) systems aim to support exploratory search, such systems are often unable to perform dynamic and timely predictions of their users' search performance. Exploratory search as a combination of browsing and focused search involves a variety of key factors such as context, intentions, motivations, prior knowledge, feelings, expectations, and strategies, which are often ignored by these IR systems. In the absence of this type of information, IR systems often rely on limited data (e.g. pages visited, queries used) and measurements (e.g., precision, recall) to perform evaluations of users' search processes. Therefore, predicting one's success or failure in an exploratory search situation is a challenging task. This may also result in missing out on opportunities for making recommendations that analyze the search process and/or recommend alternative search process instead of objects.

To overcome this limitation, this paper investigates whether by analyzing a searcher's current processes we could forecast his/her likelihood of achieving a certain level of success with respect to search performance in the future. A new machine-learning-based method is proposed to dynamically evaluate and predict search performance several time-steps ahead at each given time point of the search process during an exploratory search task. This prediction method uses a collection of features extracted from expression of information need and coverage of information. For testing, log data collected from four user studies, which included 216 users (96 individuals and 60 pairs), was used. The results show 80-90% accuracy in prediction depending on number of time-steps ahead. In effect, the work reported here provides a framework for evaluating search processes during exploratory search tasks and predicting search performance. Importantly, the proposed approach is based on user processes and is independent of any IR system, allowing one to apply and extend it to most forms of search and browse systems.

(To be published in Journal of Association for Information Science & Technology (JASIST).

To discuss possible project ideas based on this publication, please contact Dr. Chirag Shah at chirags@rutgers.edu

Featured Publication - Developing Troubleshooting Systems Using Ontologies

Dr. Reza Basseda

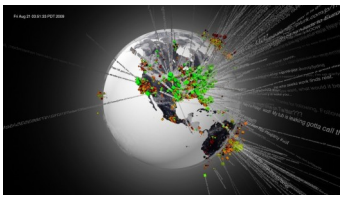
Abstract - Development of troubleshooting software is an attractive area of research for agent based system developers. In this project, we attempt to use ontologies extracted from different textual resources to automatically construct a troubleshooting virtual expert. In our solution, we verify the information about the structure of the system extracted from the textual document, then generate a conversation with the user in order to identify the problem and recommend appropriate remedies. To illustrate the approach, we have built knowledge base for a simple use case. A special parser generates conversations that can help the user solve software configuration problems.

Introduction - Troubleshooting a complex systems is a logical and systematic search for the sources of problems. After finding the problem sources in a complex system, the troubleshooting system provides remedies to solve the problem, so the system can be made operational again. Troubleshooting techniques are used widely in different complex systems such as smart phone services and applications. A troubleshooting and diagnosis system runs a troubleshooting process based on its knowledge about the structure and behavior of a complex system. A troubleshooting process not only identifies malfunctions within a failed system but also requires confirmation that the solution restores the failed system to its working state. Information extracted from resources on the world wide web can be considered as a potential source of knowledge to build a diagnosis system for commonly used applications and technologies such as smart phone services. The availability of various information resources on the web emphasizes the role of consistency checking of the knowledge base components for troubleshooting systems. There are many feasible approaches based on the behavioral knowledge about the system under diagnosis. A diagnostic system based on a simple search of symptoms and causal models is presented in [1] . In [2], Portinale uses the formalism of Petri nets to provide a diagnostic model. A model-based diagnosis method for discrete event systems with an incomplete system model has been proposed in [3]. A similar model-based reasoning system with uncertain observations has been presented in [4]. In [5], Zhang et al. present a value propagation model and an algorithm for finding a minimal diagnosis. All of these approaches rely on behavioral knowledge about the system under diagnosis. In all of those methods, a behavioral model of the complex system... (for more, please visit here)

To discuss possible project ideas based on this publication, please contact Dr. Reza Basseda at rbasseda@cs.sunysb.edu

Collaboration Outreach

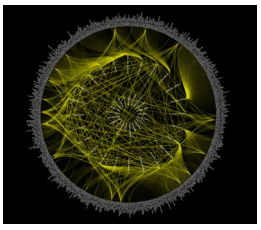
This section will feature requests for introductions to potential collaborators for all CDDA participants. Example: "Jane Doe from Rutgers is looking for collaborators in the Ads and Commerce Division of Google." or "John Doe from (CDDA Member Company) would like to discuss a possible collaboration with data scientists who have expertise in analytical chemistry." Listings will be anonymized upon request. Please contact james.mielke@rutgers.edu for postings



Visualization of people tweeting "Good morning" on August 21, 2009 — Jer Thorp

"Errors using inadequate data are much less than those using no data at all."

– Charles Babbage, Inventor and Mathematician



Patterns occurring in Oscar movies: The visualization shows the relationships between actors who have won Oscars, the directors they have worked with and all of the other actors they have worked with.

The Center for Dynamic Data Analytics

Rutgers Address:
617 Bowser Rd.
Piscataway, NJ
08854

Stony Brook Address:
1500 Stony Brook Rd.
Stony Brook, NY
11794

E-mail: james.mielke@rutgers.edu
Phone: 848-445-8824

E-mail: rzhao@cs.stonybrook.edu
Phone: 631-632-4627

"From Chaos to Knowledge"

cdda.rutgers.edu

About CDDA

The Center for Dynamic Data Analytics (CDDA) is a National Science Foundation (NSF) sponsored Industry and University Cooperative Research Program (I/UCRC) established between Rutgers University and the State University of New York (SUNY), Stony Brook.

The motivation for this center is the lack of scalable algorithms, methods and solutions for addressing the ever increasing amounts of industry-related data. The focus is on data sets that are massive, dynamic, complex and multidimensional, or what is commonly known as Big Data analytics. The goal of the center is to discover, develop and apply data analytics solutions to industry problems such that the chaotic data is transformed into knowledge and industry products.

NSF Factsheet—CDDA

CDDA Partners and Sponsors

Logos featured in this logo collage represent several partners and sponsors of the CDDA. All logos are property of their respective owners. Presence, position or size in the collage does not reflect center membership, specific significance or specific contribution to the CDDA.