GLOBAL INITIATIVE FOR ACADEMIC NETWORKS (GIAN)

Big Data Analysis and Mining in Deep Web Repositories

Overview:

Deep Web databases i.e., hidden databases only accessible by restricted web query interfaces, are widely prevalent on the Web. They represent an intriguing and prototypical instance of Big Data: huge, heterogeneous, not easily indexed, and inaccessible other than via restrictive and proprietary query interfaces such as keyword/form-based search and hierarchical/graph-based browsing interfaces. Examples include form-based web-accessible ecommerce databases (e.g., Amazon), keyword search-based document corpora (e.g., digital libraries), as well as graph-browsing based repositories (e.g., social media and collaborative websites such as Twitter, Yelp, etc.).

Efficient ways of exploring and mining contents in such hidden repositories are of increasing importance. There are two key challenges: one on the proper understanding of interfaces, and the other on the efficient exploration, e.g., crawling, sampling and analytical processing, of very large repositories. In this course, we focus on the fundamental developments in the field, including web interface understanding, crawling, sampling, and data analytics over web repositories with various types of interfaces and containing structured or unstructured data.

In the recent Information Technology Field Big Data applications are increasing day by day to perform the internet applications. Big Data Analysis is highly essential to perform Science and Engineering applications which involve huge data handling. Data Flow is highly crucial at the time of execution of any task. Parallel Task execution play vital role in Big Data applications to improve the performance of data flow in real world problems. Hadoop is a Big Data application by Apache it is most widely used tool for storing and managing of Big Data. Analyzing Big Data is a challenging task, as it involves large distributed file systems, which should be fault tolerant, flexible and scalable.

This course is organized in two modules that will be taken together. Classes will be held every day from 9:00 AM – 5:00 PM, Monday-Friday, for 05 days. Morning sessions will be devoted to lectures and afternoons for hands-on sessions with emphasis on problem solving and programming.

Course participants will learn these topics through lectures and hands-on experiments. Also case studies and assignments will be shared to stimulate research motivation of participants.

Modules	**A:Big Data Analysis and Mining in Deep Web Repositories (Morning Session) : Jan 08 - Jan 12,2018** **B: Tutorials (Afternoon Session) : Jan 08 – Jan 12,2018** **Number of participants for the course will be limited to fifty.**
You Should Attend If…	Faculty, Professionals and Research Scholars working in research areas like machine learning, Big Data Analytics. Anyone interested in learning how to extract actionable intelligence from large amounts of data, regardless of their field of specialization.
Venue	Central University of Rajasthan, Bandersindri, Kishangarh, Ajmer, Rajasthan, India.
Fees	**GIAN Portal registration** (Rs 500 fee is mandatory for all participants) Create login and password at http://www.gian.iitkgp.ac.in/GREGN/index Login and complete the Registration Form and select Course(s) Confirm application and pay Rs. 500/- (non-refundable) through online payment gateway. Download "pdf file" of the application form and email it to nagaraju@curaj.ac.in. **Central University of Rajasthan Course Registration Fee (**exclusive of GIAN Portal Registration Fee) Will be informed very soon.

The Faculty

Gautam Das is Professor and Head of the Database Exploration Laboratory (DBXLAB) at the CSE department of UT-Arlington. Prior to joining UTA in Fall 2004, Dr. Das has held positions at Microsoft Research, Compaq Corporation and the University of Memphis. He graduated with a B.Tech in computer science from IIT Kanpur, India, and with a Ph.D in computer science from the University of Wisconsin, Madison.

Dr. Das has broad research interests in all aspects of Big Data Exploration, including databases, data analytics and mining, information retrieval, and algorithms. His current research is focused on data management and algorithmic problems in the deep web, social networks and collaborative media, as well as ranking, search, and analytics problems in databases. His research has resulted in over 180 papers, many of which have appeared in premier data mining, database and algorithms conferences and journals.

His work has received several awards, including the IEEE ICDE 10-Year Influential Paper award received in 2012, ACM SIGKDD Doctoral Dissertation Award (honorable mention) in 2014 for his recent student, Best Student Paper Award of CIKM 2013, VLDB Journal special issues on Best Papers of VLDB 2012 and VLDB 2007, Best Paper of ECML/PKDD 2006, and Best Paper (runner up) of ACM SIGKDD 1998. He has been a keynote speaker on several occasions such as at ExploreDB 2015, IEEE APWC 2014, WebDB 2012, DBRank 2012, and presented invited lectures, tutorials and courses at various universities, research labs, and conferences.

Course Co-ordinator

Dr.A.Nagaraju
Co-ordiantor & Assistant Professor
Department of Computer Science
nagaraju@curaj.ac.in
+91-7568841375

Department of Computer Science
School of Mathematics , Statistic and Computational Science
Central University of Rajasthan,
NH-8, Bandarsindri, Kishangarh,
District-Ajmer (Rajasthan), 305817,
India

.....

http://www.gian.iitkgp.ac.in/GREGN

http://www.curaj.ac.in

He is on the Editorial Board of the journals ACM TODS and IEEE TKDE. He has served as General Chair of ICIT 2009, Program Chair of COMAD 2008, CIT 2004 and SIGMOD-DMKD 2004, Best Paper Awards Chair of ACM SIGKDD 2006, as well as in program committees of numerous conferences. He has served as a Guest Editor for the ACM TKDD special issue devoted to the best papers of ACM SIGKDD 2006.

INSTRUCTION FOR REGISTRATION:

Please follow the steps below for registering in the GIAN Programme on "**Big Data Analysis and Mining in Deep Web Repositories**:-

1. Register at "the GIAN portal on the link http://www.gian.iitkgp.ac.in/ by clicking on 'Course Registration/Participant Login'
2. It shall state – 'Registration to the portal is one time affair and will be valid for lifetime of GIAN. Once registered in the portal, an applicant will be able to apply for any number of GIAN courses as and when necessary. One time Non-refundable fee of Rs. 500/- is to be charged for this service. Please also note that mere registration to the portal will not ensure participation in the courses'.
3. Once done with registration, please select the course **"Big Data Analysis and Mining in Deep Web Repositories",** From the list of courses.
4. Send the soft copy of registration details from GIAN website to the following email; nagaraju@curaj.ac.in

For payment please consider any of the options;

1. DD/multicity cheque payable at Bandar Sindri , Kishangarh ,Ajmer in the name of Central University of Rajasthan.
2. Bank Transfer at

Central University of Rajasthan Ac/No	666110210000003
Bank	Bank of India
State	Rajasthan
District	Ajmer
Branch	Central University of Rajasthan
IFSC Code	BKID0006667
MICR Code	305013027
Branch Code	6667
Swift Code	BKIDINBBJPR
Address	Central University of Rajasthan, NH-8, Bandersindri, Kishangarh, District-Ajmer, Rajasthan (India), 305817

Last date for registration is 15/12/2017. Kindly complete all formalities by then. In case of any queries, please feel free to contact the Course coordinators.

COURSE TITLE: Big Data Analysis and Mining in Deep Web Repositories

Day-Wise Detail Topic

S No	Day	Topic
1	Day 1	Lecture 1: Introduction to surface web and deep web Lecture 2: Introduction to mining web repositories Tutorial 1: Problem solving session with examples of surface and deep web repositories, both hidden as well as data accessible repositories.
2	Day 2	Lecture 3: Deep web approach and mining hidden databases. Lecture 4: Resource discovery, interface understanding, and schema matching. Tutorial 2: Problem solving session with examples of resource discovery, interface understanding, and schema matching techniques.
3	Day 3	Lecture 5 : Challenges of crawling, sampling, and data analytics. Lecture 6 : Crawling over hidden databases and over search and browsing interfaces. Tutorial 3: Problem solving session with examples of crawling techniques over hidden databases, as well as search and browsing interfaces.
4	Day 4	Lecture 7: Overview of sampling techniques in approximate query processing. Lecture 8: Sampling-based data analytics over hidden databases. Tutorial 4: Problem solving session with examples of sampling based data analytics over. form/keyword and graph-browsing based deep web databases.
5	Day 5	Lecture 9: Overview of Big Data, Big Data Analytics, Exploring the use of Big Data in Business Context. Lecture 10: The Map Reduce Frame work, Techniques to Optimize Map Reduce Jobs. Tutorial 5: Hands on Practice on R which Covers Exploring R, Reading Datasets and Exporting Data From R , Manipulating and Processing Data in R