中国计算机学会《学科前沿讲习班》第 59 期

大数据时代的自然语言处理：理论、算法与架构

2015 年 10 月 9 日-11 日 南昌

在大数据时代，互联网、移动互联网和社会媒体的飞速发展为自然语言处理相关技术的发展带来了前所未有的机遇和挑战。这些技术包括信息检索、文本挖掘、知识发现、社会媒体分析、面向自然语言处理的机器学习，特别是深度学习等。本期 CCF 学科前沿讲习班《大数据时代的自然语言处理：理论、算法与架构》将邀请学术界和工业界的著名专家、学者对大数据环境下自然语言处理的基础理论、重要算法、系统架构以及其他热点问题进行系统的讲解。目的是为青年学者和学生提供一个三天的学习、交流机会，快速了解本领域的基本概念、研究内容、方法和发展趋势。

学术主任

黄萱菁，复旦大学教授
刘亦群，清华大学副教授

协办单位

江西师范大学

CCF Advanced Disciplines Lectures (ADL 59) & NLPCC 2015 Tutorials

Natural Language Processing in the Era of Big Data:

Theory, Algorithm and Architecture

Nanchang, October 9-11, 2015

The rapid development of the internet, mobile internet and social media has brought unprecedented opportunities and challenges for the research of natural language processing technologies in the big data era. Such technologies include information retrieval, text mining, knowledge discovery, social media analysis, as well as machine learning, especially deep learning for natural language processing. This Advanced Disciplines Lectures “Natural Language Processing in the Era of Big Data: Theory, Algorithm and Architecture” will invite the famous experts and professors from

Academia and Industry to give systematic lectures on the fundamental theories, algorithms, architecture and other hot-spot issues on natural language processing under big data environment. The objective of the lectures is to provide a three-day learning and communication platform for the young scholars and students to rapidly understand the element concepts, research contents, approaches and growing tendencies in the area.

Program（日程安排）

==========

October 9, 2015

8:30-9:00 Opening Ceremony, Group Photo

Lecture 1: Search and Discovery for Big Data

Xueqi Cheng, Institute of Computing Technology, Chinese Academy of Sciences

Lesson 1: 09:00-09:50 Top-k learning-to-rank and short-text topic modeling,
Lesson 2: 10:10-11:00 Multi-context recommendation and social network analysis
Lesson 3: 11:20-12:00 Scalable influence maximization, popularity prediction and collective behavior analysis

Lecture 2: Key Technologies behind a Live Social Observatory System
Tat-Seng Chua, National University of Singapore

Lesson 1: 14:00-14:50 Live social observatory system developed at NExT
Lesson 2: 15:10-16:00 Key research efforts to tackle the challenges in social media analysis
Lesson 3: 16:20-17:00 Summarization and QA

==========

October 10, 2015

Lecture 3: Construction and Mining of Text-Rich Heterogeneous Information Networks
Jiawei Han, University of Illinois at Urbana-Champaign

Lesson 1: 09:00-09:50 Phrase mining and concept discovery from massive text data
Lesson 2: 10:10-11:00 Entity recognition and typing for construction of text-rich heterogeneous information networks
Lesson 3: 11:20-12:00 Mining text-rich heterogeneous information networks

Lecture 4: Big Data Intensive Computation: concepts, Research Issues and some Solutions
Jianzhong Li, Harbin Institute of Technology
Lesson 1: 14:00-14:50 Big Data Intensive Computation: concepts
Lesson 2: 15:10-16:00 Big Data Intensive Computation: research Issues
Lesson 3: 16:20-17:00 Some Solutions from Harbin Institute of Technology
==========

October 11, 2015

Lecture 5: Learning to Process Natural Language in Big Data Environment
Hang Li, Noah's Ark Lab, Huawei Technologies

Lesson 1: 09:00-09:50 Top-k learning-to-rank and short-text topic modeling,
Lesson 2: 10:10-11:00 Multi-context recommendation and social network analysis
Lesson 3: 11:20-12:00 Scalable influence maximization, popularity prediction and collective behavior analysis

Lecture 6: Deep Learning for Chinese Information Processing
Chao Liu, Sogou

Lesson 1: 14:00-14:50 Basic embedding techniques
Lesson 2: 15:10-16:00 Feed-forward and recurrent neural networks for NLP applications
Lesson 3: 16:20-17:00 Self-built computing platform

17:00-17:20 Closing Ceremony

==========

讲者介绍

Lectures and Lecturers

==========

Xueqi Cheng
Title: Search and Discovery for Big Data
Speaker: Xueqi Cheng, Professor, Institute of Computing Technology, Chinese Academy of Sciences (CAS), Director, CAS Key Laboratory of Network Data Science and Technology
Abstract: Big data opens a new era for data-driven scientific discovery and data-driven services. It is revolutionizing paradigm of science and poses several critical issues for modern research. Firstly, the rate of data accumulation outperforms the rate of the improvement of computational power. We have to develop faster

algorithms or to design new methods that only work on part of the whole data. Secondly, we lack theory for data complexity that could guide us design good methods balancing data-complexity and task-quality. Finally, prediction-oriented tasks gradually dominate multiple scientific fields, requiring the capability to model the behavior of complex interaction behavior or mechanisms underlying complex systems. These issues motivate us to rethink our current research, fostering several new ideas or methods. In this talk, I will introduce our recent works on Web search and data mining. Specifically, this talk will cover top-k learning-to-rank, short-text topic modeling, multi-context recommendation, social network analysis, scalable influence maximization, popularity prediction, collective behavior analysis. In these specific research works, I will particularly introduce how big data transforms our research and what we could response to the challenges raised by big data.

Short bio:

Dr. Xueqi Cheng is a professor in the Institute of Computing Technology, Chinese Academy of Sciences (CAS), and the director of the CAS Key Laboratory of Network Data Science and Technology. His main research areas include Web search and data mining, data science, big data system, and social media analytics.

He is the general secretary of CCF Task Force on Big Data, the vice-chair of CIPS Task Force on Chinese Information Retrieval. He is the associate editor of IEEE Transactions on Big Data, Editorial Board Member of Journal of Computer Science and Technology and Chinese Journal of Computer. He was the general co-chair of WSDM'15, Steering Committee co-chair of IEEE Conference on Big Data, and PC members of more than 20 conferences, including ACM SIGIR, WWW, ACM CIKM, ACL, IEEE ICDM, IJCAI, and ACM WSDM.

He has more than 100 publications, and was awarded the Best Paper Award in ACM CIKM'11, and the Best Student Paper Award in ACM SIGIR'12. He is the principal investigator of more than 10 major research projects, funded by NSFC and MOST. He was awarded the NSFC Distinguished Youth Scientist (2014), the National Prize for Progress in Science and Technology (2012), the China Youth Science and Technology Award (2011) et al.

==========

Tat-Seng Chua

Title: Key Technologies behind A Live Social Observatory System

Speaker: Tat-Seng Chua, Chair Professor, National University of Singapore, Co-Director of NExT Research Center

Abstract: Given the popularity of social networks, users are sharing information on multiple aspects of their life on a wide variety of social networks. For any given topic, there are wide varieties of both social and non-social information from multiple sources. The challenges in social media analysis are multi-fold. The first and most fundamental problem is the ability to gather "representative" data about the topic from multiple sources. This is particular challenging for hot topics with many live data streams. As the key difference between social media and Web retrieval is the presence of huge amount of noise in social media data streams, hence the second challenge is

the removal of noise, both in data and user accounts. With the social media contents becoming increasingly multimedia, the third challenge is how to infer user signals from non-textual contents. The fourth challenge is the detection and tracking of sub-topics, along with insights on users' sentiments, interests and demographics. Finally, for most organizations, they would like to know what social media posts or sub-events related to them are likely to become viral, and what actions they can take.
This tutorial is divided into 3 parts. The first part describes a live social observatory system that we have developed at NExT, a joint Center between National University of Singapore and Tsinghua University. The second part details the key research efforts to tackle the above five challenges, including our research to transform unstructured social media data into descriptive, predictive and prescriptive analytics. The third part looks into the future by examining our achievements in the last 5 years. In the coming years, the social media networks will evolve from mere communication tools to co-creation and co-invention platforms with more live data streams; and from more analysis towards more predictive and prescriptive analytics.

Short bio:

Dr Chua is the KITHCT Chair Professor at the School of Computing, National University of Singapore. He was the Acting and Founding Dean of the School during 1998-2000. Dr Chua's main research interest is in multimedia information retrieval and social media analysis. In particular, his research focuses on the extraction, retrieval and question-answering (QA) of text, video and live media arising from the Web and social networks. He is the Director of a multi-million-dollar joint Center between NUS and Tsinghua University in China to develop technologies for live media search. The project will gather, mine, search and organize user-generated contents within the cities of Beijing and Singapore. His group participated regularly in TREC-QA and TRECVID evaluations in early 2000.

Dr Chua is active in the international research community. He has organized and served as program committee member of numerous international conferences in the areas of computer graphics, multimedia and text processing. He is the conference co-chair of ACM Multimedia 2005, ACM CIVR (now ACM ICMR) 2005, ACM SIGIR 2008, and ACM Web Science 2015. He serves in the editorial board of: ACM Transactions of Information Systems (ACM), Foundation and Trends in Information Retrieval (NOW), The Visual Computer (Springer Verlag), and Multimedia Tools and Applications (Kluwer). He is the Chair of steering committee of ICMR (International Conference on Multimedia Retrieval) and Multimedia Modeling conference series; and as member of International Review Panel of a large-scale research project in Europe. He is the co-Founder of two technology startup companies and an independent Director of a publicly listed company in Singapore.

==========

Jiawei Han

Title: Construction and Mining of Text-Rich Heterogeneous Information Networks

Speaker: Jiawei Han, Abel Bliss Professor of Computer Science, University of Illinois at Urbana-Champaign

Abstract: Massive amounts of data are natural language text-based, unstructured, noisy, untrustworthy, but are interconnected, potentially forming gigantic, interconnected information networks. If such text-rich data can be processed and organized into multiple typed, semi-structured heterogeneous information networks, organized knowledge can be mined from such networks. Most real world applications that handle big data, including interconnected social networks, medical information systems, online e-commerce systems, or Web-based forum and data systems, can be structured into typed, heterogeneous social and information networks. For example, in a medical care network, objects of multiple types, such as patients, doctors, diseases, medication, and links such as visits, diagnosis, and treatments are intertwined together, providing rich information and forming heterogeneous information networks. Effective analysis of large-scale, text-rich heterogeneous information networks poses an interesting but critical challenge.

In this talk, we present an overview of recent studies on construction and mining of text-rich heterogeneous information networks. We show that relatively structured heterogeneous information networks can be constructed from unstructured, interconnected, text data, and such relatively structured, heterogeneous networks brings tremendous benefits for data mining. Departing from many existing network models that view data as homogeneous graphs or networks, the text-based, semi-structured heterogeneous information network model leverages the rich semantics of typed nodes and links in a network and can uncover surprisingly rich knowledge from interconnected data. This heterogeneous network modeling will lead to the discovery of a set of new principles and methodologies for mining text-rich, interconnected data. We will also point out some promising research directions and provide arguments on that construction and mining of text-rich heterogeneous information networks could be a key to information management and mining.

Short bio:

Jiawei Han, Abel Bliss Professor of Computer Science, University of Illinois at Urbana-Champaign. He has been researching into data mining, information network analysis, database systems, and data warehousing, with over 700 journal and conference publications. He has chaired or served on many program committees of international conferences, including PC co-chair for KDD, SDM, and ICDM conferences, and Americas Coordinator for VLDB conferences. He also served as the founding Editor-In-Chief of ACM Transactions on Knowledge Discovery from Data and is serving as the Director of Information Network Academic Research Center supported by U.S. Army Research Lab, and Director of KnowEnG, a BD2K (Big Data to Knowledge) center supported by NIH. He is a Fellow of ACM and Fellow of IEEE. He received 2004 ACM SIGKDD Innovations Award, 2005 IEEE Computer Society Technical Achievement Award, 2009 IEEE Computer Society Wallace McDowell Award, and 2011 Daniel C. Drucker Eminent Faculty Award at UIUC. His book "Data Mining: Concepts and Techniques" has been used popularly as a textbook worldwide.

==========

Jianzhong Li

Title: Big Data Intensive Computation: concepts, Research Issues and some Solutions

Speaker: Jianzhong Li, Professor, Department of Computer Science and Engineering, Harbin Institute of Technology

Abstract: Recent years, big data intensive computation is becoming an important research area in response to the rapidly growing of big data and the need of high performance analyzing of big data. Big data intensive computation is different from conventional computation since they acquire and maintain continually changing big data and perform large-scale computations over big data. Big data intensive computation open up new opportunities to achieve great advances in science, biology and health care, industry and business efficiencies, and so on. This talk will discuss the concepts, motivation, challenges and research issues of DISCS. Some solutions from Harbin Institute of Technology are also presented.

Short bio:

Jianzhong Li is a professor in the Department of Computer Science and Engineering at Harbin Institute of Technology, China. He worked in the Department of Computer Science at Lawrence Berkeley National Laboratory in USA, as a scientist, from 1986 to 1987 and from 1992 to 1993. He was also a visiting professor at the University of Minnesota at Minneapolis, Minnesota, USA, from 1991 to 1992 and from 1998 to 1999. His research interests include massive data intensive computing and wireless sensor networks. He has published more than 200 papers in refereed journals and conference proceedings, such as VLDB Journal, Algorithmica, IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Parallel and Distributed Systems, SIGMOD, SIGKDD, VLDB, ICDE, INFOCOM. His papers have been cited more than 12000 times. He has been involved in the program committees of major computer science and technology conferences, including SIGMOD, VLDB, ICDE, INFOCOM, ICDCS, and WWW. He also served on the editorial boards for distinguished journals, including IEEE Transactions on Knowledge and Data Engineering, and refereed papers for varied journals and proceedings.

==========

Hang Li

Title: Learning to Process Natural Language in Big Data Environment

Speaker: Hang Li, Director, Noah's Ark Lab, Huawei Technologies

Abstract: With big data and deep learning (DL), natural language processing (NLP) is entering into a new era. In fact, DL stands out as the most effective and promising approach to learning of complicated models for building intelligent systems, among many machine learning techniques. The combination of big data and DL really provides tremendous new opportunities for making breakthroughs in NLP. Indeed NLP has observed significant progress in recent years, with advanced deep learning methods developed and big data utilized. In this lecture, I will give a survey on deep learning for natural language processing (DL4NLP), including some of the work done at Huawei Noah's Ark Lab. I will particularly focus on four major applications in NLP, namely search, question answering, natural language dialogue, and machine

translation. I will conclude my lecture by summarizing the challenges and opportunities in research on DL4NLP.

Short bio:

Hang Li is director of the Noah's Ark Lab of Huawei Technologies. His research areas include information retrieval, natural language processing, statistical machine learning, and data mining. He graduated from Kyoto University in 1988 and earned his PhD from the University of Tokyo in 1998. He worked at the NEC lab in Japan during 1991 and 2001, and Microsoft Research Asia during 2001 and 2012. He joined Huawei Technologies in 2012. Hang has published three technical books and more than 100 scientific papers at top international journals and conferences, including SIGIR, WWW, WSDM, ACL, EMNLP, ICML, NIPS, and SIGKDD. He and his colleagues' papers received the SIGKDD'08 best application paper award, the SIGIR'08 best student paper award, and the ACL'12 best student paper award. Hang worked on the development of several products such as Microsoft SQL Server 2005, Microsoft Office 2007 and Office 2010, Microsoft Live Search 2008, Microsoft Bing 2009, Bing 2010. He has more than 35 granted US patents. Hang has also been very active in the research communities and is serving top international conferences as PC chair, Senior PC member, or PC member, including SIGIR, WWW, WSDM, ACL, EMNLP, NIPS, SIGKDD, ICDM, and top international journals as associate editor, including CL, IRJ, TIST, JASIST, JCST.

==========

Chao Liu

Title: Deep Learning for Chinese Information Processing

Speaker: Chao Liu, Chief Scientist and the General Manager of the Data Science Department, Sogou

Abstract: This tutorial surveys the recent developments of Chinese Information Processing, as powered by the advancement of deep learning techniques. We focus on the paradigm shift from traditional statistical models to uniform deep neural networks, and demonstrate the resultant differences on various applications. We start by explaining the basic five embedding techniques that map words, characters, and radicals into numerical vectors, and illustrate how to build feed-forward and recurrent neural networks on top of vectors for different applications, such as Chinese word segmentation, part-of-speech and entity tagging, machine translation and search ranking etc. Finally, we briefly review our self-built computing platform consisting of hundreds of GPUs, which restlessly powers all the work as discussed above.

Short bio:

Chao Liu is the Chief Scientist, and the General Manager of the Data Science Department in Sogou Inc. Dr. Liu is elected to China "Top 1000 Young Talents", the highest honor for young oversea researchers returning to China. Before coming back to China, he was a researcher and manager of the Data Intelligence Group in Microsoft Research at Redmond. His research has been focused on Web search/ads and data mining, with about 40 conference/journal publications and many research

results transferred to Microsoft Bing, Tencent Soso, and Sogou search engines. Dr. Liu has been on the program and organizing committees of many conferences, including SIGIR, SIGKDD, WWW, etc., and actively campaigns for the mutualism between academia and industry. Dr. Liu earned his PhD in Computer Science from the University of Illinois at Urbana-Champaign in 2007, and B.S. in Computer Science from Peking University in 2003.

学术主任

ADL Chairs

==========

Xuanjing Huang is a Professor of the School of Computer Science, Fudan University, Shanghai, China. She received her PhD degree in Computer Science from Fudan University in 1998. From 2008 to 2009, she is a visiting scholar in CIIR, UMass Amherst. Her research interest includes text retrieval, natural language processing, and data intensive computing. She has published dozens of papers in several major conferences including SIGIR, ACL, ICML, IJCAI, AAAI, CIKM, ISWC, EMNLP, WSDM and COLING. In the research community, she served as the organizer of WSDM 2015, competition chair of CIKM 2014, tutorial chair of COLING 2010, SPC or PC member of past IJCAI, ACL, SIGIR, WWW, EMNLP, COLING, CIKM, WSDM and many other conferences.

Yiqun Liu received his B.S. and Ph. D degrees in Tsinghua University in 2003 and 2007, separately. He is now working as associate professor and vice chair at the Department of Computer Science and Technology in Tsinghua University. His major research interests is in Web Search, especially in search user behavior modeling, search performance evaluation and Web data quality estimation. He is also a Principle Investigator (PI) of a joint Center (named NExT) between National University of Singapore and Tsinghua University to develop technologies for live media search. He published around 30 papers at top-tier academic conferences and journals such as SIGIR, WWW, CIKM, WSDM, AAAI, IJCAI, ACM TWeb and JIR. He serves in the editorial board of the Information Retrieval Journal (Springer) and also as the task co-leader for the NTCIR (NII Testbeds and Community for Information access Research) IMine tasks.

报名信息：

注册费： （含资料和 3 天的午餐）

1、 9 月 15 日前报名并缴费：会员 1100 元，非会员 1500 元

2、 9 月 15 日之后及现场缴费：1725 元

优惠办法：

1、 同一单位一次有 5 人报名者，第 6 位免注册费（无论会员与否，仅对提前注册者有效，当天不予受理）。

2、2014 年参加过 2 次讲习班的 CCF 会员可优惠 100 元。

3、2015 年参加 3 次讲习班的 CCF 会员，第 4 次参加时免交注册费。

4、往届学员推荐一名新学员时，推荐者当期注册费优惠 100 元。

5、同一单位一次参加 10 人（含）以上报名者，均按会员价注册。

6、预订全年活动 5 次（含）报名者，可享受 6 折优惠（900 元）。

7、同时满足以上多项优惠条款时，只能选择一项。

NLPCC 联合注册及优惠措施：

本次 ADL 同时做为 NLPCC 2015(CCF 国际自然语言处理与中文计算会议)的 Tutorials，提供 NLPCC 大会与 ADL 的联合注册方式。详细信息请访问 NLPCC 2015 网站：

http://tcci.ccf.org.cn/conference/2015/index.html

联合注册将给予 200 元注册优惠（不能与前述优惠条款同时使用）

缴费方式：

邮寄:江西省南昌市紫阳大道 99 号江西师范大学计算机信息工程学院,邮编:330022 ，收款人：徐凡

银行转账：

开户行：**中国银行南昌市南湖支行营业部**
户　名：**南昌恒天会务会展服务有限公司**
账　号：**1962 1436 2434**
请务必注明：**参会者姓名+NLPCC2015(ADL)**

报名方式：

即日起至 2015 年 9 月 15 日，报名者请填写附表并发送至联系人邮箱 xufan@jxnu.edu.cn，按报名先后录取（名额有限、先报先得）。学会秘书处将邮件联系确认。自 9 月 16 日起，只接受现场报名。

联系人：徐凡　　E-Mail: xufan@jxnu.edu.cn
电话：18870088082
地址：江西省南昌市紫阳大道 99 号江西师范大学计算机信息工程学院

CCF ADL59 报名表

《大数据时代的自然语言处理：理论、算法与架构（南昌）》

<table>
<tr><td>姓名</td><td></td><td>性别</td><td></td></tr>
<tr><td>任职单位</td><td colspan="3"></td></tr>
<tr><td>职称</td><td colspan="3"></td></tr>
<tr><td>是否 CCF 会员 1</td><td></td><td>会员号</td><td></td></tr>
<tr><td>手机</td><td></td><td>Email</td><td></td></tr>
<tr><td rowspan="3">住宿 2
（如需安排）</td><td colspan="3">入住时间：</td></tr>
<tr><td colspan="3">离开时间：</td></tr>
<tr><td colspan="3">单住：　　合住：</td></tr>
<tr><td>注册费缴纳方式
√</td><td colspan="3">☐邮寄；　☐银行转账；　☐现场缴费（仅限现场报名）</td></tr>
<tr><td>发票抬头 3</td><td colspan="3"></td></tr>
<tr><td>发票项目内容 4
√</td><td colspan="3">☐注册费　☐会议费　☐会务费　☐培训费</td></tr>
<tr><td colspan="4">参加本期讲习班的目的：</td></tr>
<tr><td colspan="4">信息来源：√（请注明）
□CCF 周刊 □CCF 网页 □《CCCF》 □熟人介绍 □单位通告 □其它_____</td></tr>
<tr><td colspan="4">我申请参加本期讲习班并承诺按主办单位的规定参加。</td></tr>
</table>