中国计算机学会《学科前沿讲习班》
第五十二期

面向网络文本处理的统计学习方法

2014年12月5日-7日 深圳

互联网与移动网络的飞速发展为网络文本处理技术的发展带来了前所未有的机遇和挑战。网络文本处理涉及信息检索、信息抽取、观点挖掘、语义搜索等任务，统计学习是其中的重要方法。本期ＣＣＦ学科前沿讲习班《面向网络文本处理的统计学习方法》将邀请学术界和工业界的著名专家、学者对信息检索、信息抽取、观点挖掘和语义搜索中统计学习的基础理论、方法和应用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

赵军，中科院自动化所研究员
张敏，清华大学副教授

协办单位

哈尔滨工业大学深圳研究生院

CCF Advanced Disciplines Lectures (ADL 52)
& NLPCC 2014 Tutorials

Statistical Learning Methods for Natural Language Processing on the Internet

Shenzhen, December 5-7, 2014

The rapid development of the internet and mobile internet has brought unprecedented opportunities and challenges for the research of natural language processing for the internet corpora. Natural language processing tasks for the internet corpora include information retrieval, information extraction, opinion mining, semantic search, etc, where statistical learning is one of the important methods. This Advanced Disciplines Lectures “Statistical Learning Methods for Natural Language Processing on the Internet” will invite the famous experts and professors from Academia and Industry to give systematic lectures on the fundamental theories, methods, applications and other hot-spot issues on statistical learning in information retrieval, information extraction, opinion mining, semantic search, etc. 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）

===

December 5, 2014

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

Lecture 1: Wikification and Beyond: The Challenges of Entity and Concept Grounding

Heng Ji, Rensselaer Polytechnic Institute, USA

Lesson 1: 09:00-10:00 Motivation and Task Definition
A Skeletal View of Wikification Systems
Lesson 2: 10:20-11:20 Key Challenges and Recent Advances
New Tasks, Trends and Applications
Lesson 3: 11:40-12:00 QA

Lecture 2: Big Learning with Bayesian Methods

Jun Zhu, Tsinghua University

Lesson 1: 13:30-15:30 Basics, Big Learning Challenges,
and Regularized Bayesian Inference
Lesson 2: 15:50-16:50 Online Learning, Large-scale Topic Graph Learning
and Visualization
Lesson 3: 17:10-17:30 QA

===

December 6, 2014

Lecture 3: Sentiment Analysis: Mining Opinions, Sentiments and Emotions

Bing Liu, the University of Illinois at Chicago, USA

Lesson 1: 09:00-10:00 Sentiment analysis essentials
Lesson 2: 10:20-11:20 Advanced topics

Lesson 3:   11:40-12:00   QA

Lecture 4: Semantic Matching in Search

Jun Xu, Institute of Computing Technology, Chinese Academy of Sciences

Lesson 1: 13:30-14:30 Semantic Matching between Query and Document
Lesson 2: 14:50-16:50 Approaches to Semantic Matching in Search
Lesson 3: 17:10-17:30 QA

==

December 7, 2014

Lecture 5:   From simple search to search intelligence: the evolution of search engines

Jianyun Nie, University of Montreal, Canada

Lesson 1:   09:00-10:00   Traditional IR models, query and document expansion
Lesson 2:   10:20-11:20   Advanced methods of intelligent IR: mining relations in documents and query logs, mining search intents
Lesson 3:   11:40-12:00   QA

Lecture 6: Machine Learning for Search Ranking and Ad Auction

Tieyan Liu, Microsoft Research Asia

Lesson 1: 13:30-15:30   Machine learning for Web search
Lesson 2: 15:50-16:50   Machine learning for computational advertising
Lesson 3: 17:10-17:30   QA

17:40-18:00   Closing Ceremony

==

讲者介绍
Lectures and Lecturers

Heng Ji

Title: Wikification and Beyond: The Challenges of Entity and Concept Grounding

Abstract： Contextual disambiguation and grounding of concepts and entities in natural language text are essential to moving forward in many natural language understanding related tasks and are fundamental to many applications. The Wikification task aims at automatically identifying concept mentions appearing in a text document and linking them to (or “grounding them in”) concept referents in a knowledge base (KB) (e.g., Wikipedia). For example, consider the sentence, "The Times report on Blumenthal (D) has the potential to fundamentally reshape the contest in the Nutmeg State.". A Wikifier should identify the key entities and concepts (Times, Blumental, D and the Nutmeg State), and disambiguate them by mapping them to an encyclopedic resource revealing, for example, that “D” here represents the Democratic Party, and that “the Nutmeg State” refers Connecticut. Wikification may benefit both human end-users and Natural Language Processing (NLP) systems. When a document is Wikified a reader can more easily comprehend it, as information about related topics and relevant enriched knowledge from a KB is readily accessible. From a system-to-system perspective, a Wikified document conveys the meanings of its key concepts and entities by grounding them in an encyclopedic resource or a structurally rich ontology.

The primary goals of this tutorial are to review the framework of Wikification and motivate it as a broad paradigm for cross-source linking for knowledge enrichment. We will present and discuss multiple dimensions of the task definition, present the basic building blocks of a state-of-the-art Wikifier system, share some key lessons learned from the analysis of evaluation results, and discuss recently proposed ideas for advancing work in this area along with some of the key challenges. We will also suggest some research questions brought up by new applications, including interactive Wikification, social media, and censorship. The tutorial will be useful for both senior and junior researchers with interests in cross-source information extraction and linking, knowledge acquisition, and the use of acquired knowledge in natural language processing and information extraction. We will try to provide a concise roadmap of recent perspectives and results, as well as point to some of our Wikification resources that are available to the research communities.

Bio： Heng Ji is Edward P. Hamilton Development Chair Associate Professor in Computer Science Department of Rensselaer Polytechnic Institute. She received her B. A. and M. A. from Tsinghua University in 2000 and 2002 respectively, and Ph.D. in Computer Science from New York University in 2007. Her research interests include Natural Language Processing, Data Mining, Information Networks and Social Networks, and Security. She received Google Research Awards in 2009 and 2014, NSF CAREER award in 2009, Sloan Junior Faculty Award in 2012, IBM Watson Faculty Award in 2012 and 2014, PACLIC2012 Best Paper Runner-up, "Best of SDM2013" paper, "Best of ICDM2013" paper and "AI's 10 to Watch" Award by IEEE Intelligent Systems in 2013. She is the leader of the U.S. Army Research Lab projects on information fusion and knowledge networks construction. She coordinated the NIST TAC Knowledge Base

Population task in 2010, 2011 and 2014, served as the vice Program Committee Chair for IEEE/WIC/ACM WI2013, the Information Extraction area chair for NAACL2012, ACL2013, EMNLP2013 and NLPCC2014, Content Analysis Track Chair of WWW2015, and the Financial Chair of IJCAI2016. Her research is funded by U.S. National Science Foundation, U.S. Army Research Lab, U.S. Defense Advanced Research Projects Agency (DARPA), U.S. Air Force Research Lab, Google, Disney and IBM.

==

Jun Zhu

Title: Big Learning with Bayesian Methods

Abstract：Bayesian methods represent one important school of statistical methods for learning, inference and decision making. At the core is Bayes' theorem, which has been developed for more than 250 years. However, in the Big Data era, many challenges need to be addressed, ranging from theory, algorithm, and applications. In this talk, I will introduce some recent developments on generalizing Bayes' theorem to incorporate rich side information, which can be the large-margin property we like to impose on the model distribution, or the domain knowledge collected from experts or the crowds, and scalable online learning and distributed inference algorithms. The generic framework to do such tasks is called regularized Bayesian inference (RegBayes). I will introduce the basic ideas of RegBayes as well as several concrete examples with scalable inference algorithms.

Bio:

Dr. Jun Zhu is an associate professor at the Department of Computer Science and Technology in Tsinghua University. He received his Ph.D. in Computer Science from Tsinghua in 2009. Before joining Tsinghua in 2011, he did post-doctoral research at the Machine Learning Department in Carnegie Mellon University. His current work involves both the foundations of statistical learning, including theory and algorithms for probabilistic latent variable models, sparse learning in high dimensions, Bayesian nonparametrics, and large-margin learning; and the application of statistical learning in social network analysis, data mining, and multi-media data analysis.

Prof. Zhu has published over 50 peer-reviewed papers in the prestigious conferences and journals, including ICML, NIPS, KDD, JMLR, PAMI, etc. He is an associate editor for IEEE Trans. on PAMI. He served as area chair/senior PC for about 10 top-tier conferences, including ICML (2014, 2015), IJCAI (2013, 2015), UAI 2014, and NIPS 2013. He was a local co-chair of ICML 2014. He is a recipient of the CCF Distinguished PhD Thesis Award (2009), Microsoft Fellowship (2007), IEEE Intelligent Systems "AI's 10 to Watch" Award (2013), NSFC Excellent Young Scholar Award (2013), and CCF

Young Scientist Award (2013). His work is supported by the "221 Basic Research Plan for Young Talents" at Tsinghua.

===

Bing Liu

Title: Sentiment Analysis: Mining Opinions, Sentiments and Emotions

Abstract: Sentiment analysis or opinion mining is the computational study of people's opinions, sentiments, and emotions toward entities, events and their attributes. Opinions are important because they are key influencers of our behaviors. Our beliefs and perceptions of reality, and the choices we make, are to a considerable degree conditioned on how others see and evaluate the world. For this reason, when we need to make a decision we often seek out the opinions of others. This is true not only for individuals but also for organizations. In the past decade, sentiment analysis attracted a great deal of attentions from both academia and industry due to many challenging research problems and a wide range of applications. Sentiment analysis can be seen as a special NLP semantic analysis task. It touches every core issue of NLP, yet it is confined because a sentiment analysis system does not need to fully "understand" each sentence or document. It only needs to comprehend some aspects of it, e.g., positive or negative opinions and emotions. While general natural language understanding is perhaps far from us, we may be able to solve the sentiment analysis problem. Sentiment analysis offers an excellent platform for NLP researchers to potentially make major breakthroughs on many fronts of NLP and even machine learning. In this tutorial, I will define the problem, introduce the current state-of-the-art and discuss potential contributions that sentiment analysis can make to the general NLP and vice versa.

Bio： Bing Liu is a professor of Computer Science at the University of Illinois at Chicago (UIC). He received his PhD in Artificial Intelligence from the University of Edinburgh. Before joining UIC, he was a faculty member at the National University of Singapore. His current research interests include sentiment analysis and opinion mining, data mining, machine learning, and natural language processing (NLP). He has published extensively in top conferences and journals. He is also the author of two books: "Sentiment Analysis and Opinion Mining" (Morgan and Claypool) and "Web Data Mining: Exploring Hyperlinks, Contents and Usage Data" (Springer). In addition to research impacts, his work has also made important social impacts. Some of his work has been widely reported in the press, including a front-page article in The New York Times. On professional services, Liu has served as program chairs of many leading data mining related conferences of ACM, IEEE, and SIAM: KDD, ICDM, CIKM, WSDM, SDM, and PAKDD, as associate editors of several leading data mining journals, e.g., TKDE, TWEB, DMKD, and as area/track chairs or senior technical committee members of numerous NLP, data mining, and Web technology conferences. He currently also serves as the Chair of ACM SIGKDD, and is an IEEE Fellow.

===

Jun Xu

Title: Semantic Matching in Search

Abstract: Most of the tasks in natural language processing and information retrieval, including search, question answering, and machine translation, are based on matching between language expressions. This approach works quite well in practice; its limitation is also obvious, however. Sometimes mismatch can occur. 'Semantic matching' is an effective approach to overcome the challenge, that is to conduct more semantic analysis on the language expressions and perform matching between language expressions at semantic level. In this talk, major approaches to semantic matching in search will be introduced, including matching by query reformulation, matching with term dependency model, matching with translation model, matching with topic model, and matching with latent space model. Open problems and future directions of semantic matching will also be discussed in the talk.

Bio: Jun Xu is a researcher at Institute of Computing Technology, Chinese Academy of Sciences. He received his PhD in computer science from Nankai University in 2006. After that, he worked at Microsoft Research Asia and Huawei Noah's Ark Lab. In 2014, he joined Institute of Computing Technology, Chinese Academy of Sciences. Jun Xu's research interest focuses on web search and text mining. He has published extensively in prestigious conferences and journals including SIGIR, WWW, CIKM, JMLR, and TOIS etc. He is very active in the research communities and severed or is serving the top conferences and journals. He developed the learning to rank algorithms of IR-SVM and AdaRank, as well as the LETOR dataset. He released the AdaRank algorithm, RLSI algorithm, and LETOR dataset to the academic. He has also contributed to the development of Microsoft products including Microsoft Bing and Office.

===

Jianyun Nie

Title: From simple search to search intelligence: the evolution of search engines

Abstract: Traditional information retrieval (IR) methods rely on word matching, which is unable to cope with the great variability of natural language expressions. If early users were amazed by the great amount of information they could find using simple word-matching search, search users today are no longer content with such a simplistic approach. More and more, they expect search engines to exhibit some level of intelligence as a human being. For example, it should understand that "ping pong" means the same thing as "table tennis", and "next world championship" would refer to that of 2015. During the last 10 years, search engines have made tremendous progress in meeting the user's expectation. They do so by leveraging various resources such as query logs, Wikipedia and so on.

In this tutorial, I will review a series of approaches in IR aiming at incorporating some reasoning capability in search. These include the traditional query and document expansion. More recent approaches exploit different resources to understand user's query and information need, and to provide answers accordingly.

Bio: Jian-Yun Nie is a professor in University of Montreal. He obtained his PhD from University of Grenoble (France) on information retrieval. Since then, his research has always been focused on information retrieval and natural language processing. Among other topics, he has worked on IR models, cross-language IR, query expansion and query understanding. Jian-Yun Nie has published a number of papers on these topics and his papers have been widely cited. He published a monograph on cross-language information retrieval (Morgan and Claypool, 2010). He is on editorial board of several international journals, and is a regular PC member of the major conferences in these areas (SIGIR, CIKM, ACL, etc.). He has also been the general chair of SIGIR conference in 2011 held in Beijing.

===

Tieyan Liu

Tit le: Machine Learning for Search Ranking and Ad Auction

Abstract : In the era of information explosion, search has become an important tool for people to retrieve useful information. Every day, billions of search queries are submitted to commercial search engines. In response to a query, search engines return a list of relevant documents according to a ranking model. In addition, they also return some ads to users, and extract revenue by running an auction among advertisers if users click on these ads. This "search + ads" paradigm has become a key business model in today's Internet industry, and has incubated a few hundred-billion-dollar companies. Recently, machine learning has been widely adopted in search and advertising, mainly due to the availability of huge amount of interaction data between users, advertisers, and search engines. In this talk, we discuss how to use machine learning to build effective ranking models (which we call learning to rank) and to optimize auction mechanisms. (i) The difficulty of learning to rank lies in the interdependency between documents in the ranked list. To tackle it, we propose the so-called listwise ranking algorithms, whose loss functions are defined on the permutations of documents, instead of individual documents or document pairs. We prove the effectiveness of these algorithms by analyzing their generalization ability and statistical consistency, based on the assumption of a two-layer probabilistic sampling procedure for queries and documents, and the characterization of the relationship between their loss functions and the evaluation measures used by search engines (e.g., NDCG and MAP). (ii) The difficulty of learning the optimal auction mechanism lies in that advertisers' behavior data are strategically generated in response to the auction mechanism, but not randomly sampled in an i.i.d. manner. To tackle this challenge, we propose a game-theoretic learning method, which first models the strategic behaviors of advertisers, and then optimizes the auction mechanism by assuming the advertisers to respond to new auction mechanisms according to the learned behavior model. We prove the effectiveness of the proposed method by analyzing the generalization bounds for both behavior learning and auction mechanism learning based on a novel Markov framework.

Bio: Tie-Yan Liu is a senior researcher and research manager at Microsoft Research. His research interests include machine learning (learning to rank, online learning, statistical learning theory, and deep learning), information retrieval, and algorithmic game theory. He is well known for his pioneer work on learning to rank for information retrieval. He has authored the first book in this area, and published tens of highly-cited papers (with over 6500 citations in the past few years) on both algorithms and theorems of learning to rank. He has also published extensively on other related topics. In particular, his paper on graph mining won the best student paper award of SIGIR (2008); his paper on video shot boundary detection won the most cited paper award of the Journal of Visual Communication and Image Representation (2004-2006); and his work on Internet economics won the research break-through award of Microsoft Research Asia (2012). Tie-Yan is very active in serving the research community. He is a program committee co-chair of ACML (2015), WINE (2014), AIRS (2013), and RIAO (2010), a local co-chair of ICML 2014, a tutorial co-chair of WWW 2014 and SIGIR 2016, a doctoral consortium co-chair of WSDM 2015, a demo/exhibit co-chair of KDD (2012), and an area/track chair of many conferences including ACML (2014), SIGIR (2008-2011), AIRS (2009-2011), and WWW (2011, 2014). He is an associate editor of ACM Transactions on Information System (TOIS), an editorial board member of Information Retrieval Journal and Foundations and Trends in Information Retrieval. He has given keynote speeches at ECML/PKDD (2014), CCML (2013), CCIR (2011, 2014), and PCM (2010), and tutorials at SIGIR (2008, 2010, 2012), WWW (2008, 2009, 2011), and KDD (2012). He is a senior member of the IEEE, the ACM, and the CCF. He is currently an adjunct professor/Ph.D. supervisor of the Nankai University, the University of Science and Technology of China, and the Sun Yat-Sen University; and an Honorary Full Professor of the University of Nottingham.

报名信息：

注册费： （含资料和 3 天的午餐）

1、 10 月 30 日前报名并缴费：会员 1100 元，非会员 1500 元
2、 11 月 15 日前缴费：会员 1300 元，非会员 1500 元
3、 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、2013 年参加过 2 次讲习班的 CCF 会员可优惠 100 元。
3、2014 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 同时作为 NLPCC 2014（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 2014 网站：

http://tcci.ccf.org.cn/conference/2014/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：广东省深圳市南山区西丽大学城 哈工大深圳研究生院 C 栋 303B 室，邮编：518055 ，收款人：李莉

银行转账：

开户行：**平安银行深圳华新支行**
户名：**哈尔滨工业大学深圳研究生院**
账号：**0142 1003 2763 8**
请务必注明：**姓名+ADL52**

报名方式：

即日起至 2014 年 11 月 15 日，报名者请填写附表并发送至：lily@hitsz.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 11 月 16 日起，只接受现场报名。

联系人：李莉 E-Mail: lily@hitsz.edu.cn
电话：0755-26033182 / 15889532935
地址：广东省深圳市南山区西丽大学城 哈工大深圳研究生院 C 栋 303B 室

CCF ADL52 报名表

《面向网络文本处理的统计学习方法（深圳）》

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