中国计算机学会《学科前沿讲习班》
第四十六期

面向大数据的自然语言处理与机器学习

2013 年 11 月 15 日-17 日 重庆

简介

自然语言处理(Natural Language Processing, NLP)与机器学习(Machine Learning, ML)一直是计算机科学，尤其是人工智能研究方向的两个核心问题。近几年来，互联网、社交媒体与移动平台的迅猛发展为传统自然语言处理与机器学习技术带来前所未有的挑战和机遇，使其成为学术界和工业界的研究热点。一方面，传统的自然语言处理与机器学习在基础研究方面取得了长足进展，为工业应用提供了有效的技术支持；另一方面，不断涌现的真实问题和大规模数据又呼唤更加有效的自然语言处理与机器学习技术。

本期 CCF 学科前沿讲习班《面向大数据的自然语言处理与机器学习》将邀请学术界和工业界的著名专家、学者对自然语言处理基础理论、方法和应用，面向自然语言的机器学习技术，以及当前面向大数据时代的热点问题进行深入浅出的讲解。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。

本讲习班同时作为第二届 CCF 自然语言处理与中文计算会议（NLP&CC 2013）的 Tutorials，将围绕大会主题——“数据智能、知识智能与社会智能”，重点讲解自然语言处理基础、社会媒体和语言计算、ＮＬＰ工业应用和典型案例、面向ＮＬＰ的机器学习、Deep Learning 以及深层神经网络计算等内容。

学术主任

张民，苏州大学教授

李沐，微软亚洲研究院研究员

协办单位

重庆大学

苏州大学

微软亚洲研究院

日程安排

==========

2013年11月15日：自然语言处理和机器翻译

8:30-9:00　开班仪式、合影

第一讲　自然语言处理：基础技术与互联网创新　　万小军 北京大学 副教授
　第一课　09:00-10:20　自然语言处理基础
　第二课　10:40-11:40　语义计算和篇章分析
　第三课　11:40-12:00　互动问答

第二讲　大数据时代的机器翻译　　刘 洋　清华大学 副教授
　第一课　13:30-15:30　机器翻译概况、基于词和短语的方法
　第二课　16:00-17:30　基于句法的方法和未来发展趋势
　第三课　17:30-17:50　互动问答

==========

2013年11月16日：社会计算与自然语言处理应用

第三讲　社会网络计算及社会影响力分析　　唐杰　清华大学　副教授
　第一课　09:00-10:20　社会网络计算基础
　第二课　10:40-11:40　社会网络计算之社会影响力分析
　第三课　11:40-12:00　互动问答

第四讲　自然语言处理工业应用/开发实践
　第一课　13:20-15:20　大数据时代的智能问答和搜索　张阔　搜狗 研究员
　第二课　15:40-16:40　大数据时代的搜索广告查询分析　胡云华 阿里 研究员
　第三课　16:40-17:40　大数据下的广告排序技术及实践　蒋龙　阿里 研究员
　第四课　17:40-18:00　互动问答

==========

2013年11月17日：机器学习

第五讲　Statistical Machine Learning for NLP （统计机器学习与自然语言处理）
　　朱小槿　University of Wisconsin-Madison　副教授
　第一课　09:00-10:00　Basics of Statistical Learning （统计机器学习基础）
　第二课　10:20-11:10　Graphical Models （图模型）
　第三课　11:10-12:00　Bayesian Non-Parametric Models （贝叶斯非参数方法）
　第四课　12:00-12:20　互动问答

第六讲　Deep Learning - What, Why, and How　　俞栋　MSR　研究员

第一课　13:40-14:40　Deep Learning: Premise, Philosophy, and Its Relation to Other Techniques

第二课　14:40-15:40　Basic Deep Learning Models

第三课　16:00-17:15　Deep Neural Network and Its Application in Speech Recognition

第四课　17:15-17:35　互动问答

17:35-18:00　结业式

讲者介绍

==========

万小军　　北京大学　副教授

报告题目：自然语言处理：基础技术与互联网创新

摘要：随着互联网上文本数据的爆炸性增长，如何对这些数据进行智能分析、语义挖掘与深度利用是学术界和工业界所共同面临的重大挑战。自然语言处理技术则是应对这一挑战的关键基础技术，目前已在互联网搜索、智能问答、舆情与情报分析等系统中发挥重要的作用。本报告首先简要介绍自然语言处理基础概念与技术（包括词法、句法分析等），然后重点介绍自然语言语义与篇章分析的若干关键技术，以及面向互联网文本大数据的语义与篇章分析新技术与新应用，最后对该领域所面临的难点与机遇进行探讨。

报告人简历：2006 年 7 月在北京大学信息科学技术学院获博士学位，2008 年入选教育部新世纪优秀人才支持计划与北京市科技新星计划。研究方向为自然语言处理与文本挖掘，主要研究兴趣包括文本摘要与生成、情感分析与观点挖掘、知识获取与推荐等。曾担任自然语言处理领域顶级国际会议 ACL2011 与重要国际会议 IJCNLP2011 领域主席，担任多个国际一流会议（ACL、WWW、SIGIR、CIKM、COLING、EMNLP 等）程序委员会委员，同时担任多个国际权威期刊（T-ALIP、T-KDE、T-ASLP、T-IST、T-SLP、TACL 等）审稿

人。获得 EMNLP2010 最佳审稿人奖。在相关领域顶级期刊与会议上发表论文 20 多篇(包括 SIGIR、ACL 、AAAI 、IJCAI、ACM TOIS、Computational Linguistics 等)。 获授权发明专利 10 余项。在 NTCIR-8 MOAT、TAC2010 RTE-6、TAC2011 Guided Summarization 等权威国际评测中取得多项国际第一名的优异成绩。

==========

刘 洋 清华大学 副教授

报告题目：大数据时代的机器翻译

摘要：随着全球经济一体化的迅速发展，世界上不同国家不同民族之间的交流越来越频繁，语言障碍问题日益凸显。机器翻译能够利用计算机将一种语言的文本自动转换成另外一种语言，是能够缓解语言障碍问题的有效技术手段之一。目前，互联网已步入大数据时代，为数据驱动的统计机器翻译方法提供了丰富的多语言文本数据，极大促进了机器翻译研究与应用的发展。本课程将介绍目前主流的统计机器翻译方法在过去二十年的发展历程，围绕基于词、基于短语和基于句法的方法介绍相关模型、算法、评测和开源工具等，并对未来的发展趋势展开探讨。

报告人简历：2007 年于中国科学院计算技术研究所获得博士学位，2011 年至今在清华大学工作。研究方向是自然语言处理，在顶级国际刊物和会议 Computational Linguistics、ACL、AAAI、EMNLP、COLING 上发表二十余篇论文，获 COLING/ACL 2006 Meritorious Asian NLP Paper Award，在 ACL 2010 讲授题为“Tree-based and Forest-based Translation”的 tutorial，将担任 ACL 2014 Tutorial 共同主席。个人主页：http://nlp.csai.tsinghua.edu.cn/~ly/.

==========

唐 杰　　清华大学　　副教授

报告题目：社会网络计算及社会影响力分析

摘要： In this talk, I am going to introduce the detailed background knowledge for social computing, including methodologies and tools for macro-level, meso-level and micro-level social network analysis. Then I will focus on social influence, an important phenomenon that occurs when one's opinions, emotion, or behaviors are affected by others. Many applications have been built based around the implicit notation of social influence between people, such as marketing, advertisement and recommendations. With the exponential growth of online social network services such as Facebook and Twitter, social influence can for the first time be measured over a large population. I will examples on Twitter, Weibo, ArnetMiner, Flickr, Gowalla, etc. to explain how to verify the existence of social influence, how to quantify social influence in social networks, and how to model the influence diffusion procedure.

首先从宏观、中观、微观三个方面介绍关于社会计算的基础知识。然后着重分析社会网络中的关键问题：社会影响力。社会影响力是社交网络演化的原动力，影响力分析有很多实际的应用，例如：广告、推荐等。将使用 Twitter、微博、学术、Flickr 等网络为例介绍如何证明社会影响力的存在、如何度量社会影响力以及如何对影响力传播进行建模。

报告人简历：Jie Tang is an associate professor at the Department of Computer Science and Technology, Tsinghua University. He was honored with the CCF Young Scientist Award, NSFC Excellent Young Scholar, IBM Innovation Faculty Award, and the New Star of Beijing Science & Technology. His research interests include social network analysis, data mining, and machine learning. He has published more than 100 journal/conference papers (in major international journals and conferences including: KDD, IJCAI, WWW, SIGMOD, ACL, Machine Learning Journal, TKDD, and TKDE) and held 10 patents. He also served as Workshop Co-Chair of SIGKDD'13, Local Chair of SIGKDD'12, Publication Co-Chair of SIGKDD'11, Program Co-Chairs of ADMA'11 and SocInfo'12, and also serves as the PC member of more than 50 international conferences. He is now leading the project Arnetminer.org for academic social network analysis and mining.

清华计算机系副教授，曾在康纳尔大学、伊利诺伊香槟分校、香港科技大学访问。主要研究兴趣包括：社会网络分析、数据挖掘和语义 Web，提出基于话题的社会影响力建模方法。主持国家级、部委级和国际合作研究项目 30 余项；发表论文 100 多篇(包括：SIGKDD，IJCAI, ACL, SIGMOD, SIGIR, WWW 等)，申请专利 10 项，荣获首届国家自然科学基金优秀青年基金，获 2012 中国计算机学会青年科学家奖、2011 年北京市科技新星、2010 年清华大学学术新人奖（清华大学 40 岁以下教师学术最高奖）、IBM 全球

创新教师奖以及 KDD’12 Best Poster Award、PKDD’11 Best Student Paper Runnerup 和 JCDL’12 Best Student Paper Nomination。研发了研究者社会网络 ArnetMiner 系统，吸引了 220 个国家和地区 432 万独立 IP 的访问。

==========

张阔　　搜狗　　研究员

报告题目：大数据时代的智能问答和搜索

摘要：传统的网页搜索以关键字的匹配为主要基础，将互联网上相关的网页返回给搜索引擎用户。然而传统的网页搜索没有知识的引导以及对于网页内容的深入整理，返回的网页结果不能精准的给出所需的信息，针对该问题，搜狗研发并发布了知立方产品，基于对互联网数据的整理及知识化、用户自然语言查询的结构化、图数据的索引及搜索，使得搜索引擎拥有推理计算的能力，帮助用户更快、更准的获取所需信息。

报告人简历：清华大学学士、工学博士，搜狗搜索科学家，研发总监。主要研究方向为信息检索、中文信息处理、机器学习。曾先后担任搜狗公司商务搜索研究部门、搜狗搜索研究部门总监。曾参加或负责多项国家重点科研项目、高校联合实验室科研项目。现担任科技部 863 重大项目中国云中《以公众汉语服务为主的搜索引擎研制》课题总负责人。在 SIGIR,WWW,CIKM, DASFAA 等会议发表论文 10 余篇，申请专利 20 余项。

==========

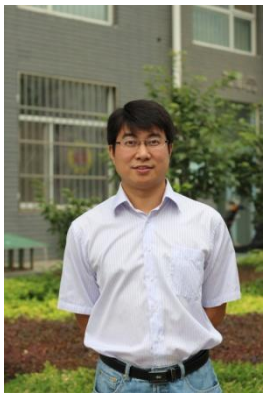
胡云华　阿里　研究员

报告题目：大数据时代的搜索广告查询分析

摘要：搜索广告中，查询（Query）分析为广告的触发和排序提供了至关重要的信息。如何充分利用大数据，实现对查询的精准、快速、多维度分析和理解，是搜索广告面临的重要挑战。本讲座将向大家介绍阿里妈妈的广告算法团队如何面临这种挑战，获得查询分析的突破。首先会对学术和产业界查询的分析和理解的研究现状、方法等进行简单概述，并介绍淘宝的数据特点，进而引入我们独创的一种基于大数据的查询分析框架；然后会介绍我们如何在统一框架下快速高效支持不同的自然语言处理任务；最后会介绍查询分析在各广告业务线中的贡献以及未来的发展。

报告人简历：阿里巴巴集团阿里妈妈事业部高级技术专家，负责广告算法的搜索、定向及品牌展示广告团队，致力于提高广告业务的算法水平。目前主要工作集中在基于大数据的查询分析、个性化搜索、广告触发、点击率预估、在线学习等核心技术方向。加入阿里前曾在微软亚洲研究院（MSRA）自然语言处理组（NLP）及网络搜索和挖掘组（WSM）工作过 5 年，先后从事企业搜索、学术搜索（Academic Search）以及搜索日志挖掘等工作。研究成果已在微软全球性的核心产品 Office 办公软件及 Bing 搜索引擎中成功应用。研究方向包括文本挖掘，机器学习，以及自然语言处理等。曾在 SIGIR，WSDM，CIKM，AAAI，JCDL 等知名国际会议以及 IPM 等一流国际期刊上发表学术论文 10 余篇，并任 TOIS，JCST，IPM，TKDD 等期刊的评委。

==========



蒋龙　阿里　研究员

报告题目：大数据时代的搜索广告查询分析

摘要：广告是当今互联网应用最重要的变现方式之一，是当今互联网能蓬勃发展的重要支柱之一。本次讲座会首先互联网广告的业界研发现状进行简单概述，回顾当今互联网广告的主要业务形式，分析其中的关键技术——广告点击率预估的内涵及作用。接着介绍业界常用的基于机器学习的广告点击率预估模型，特征，评估指标，然后分析大数据下广告点击率预估技术面临的挑战，以及在阿里巴巴我们如何利用 Hadoop 及 MPI 并行计算集群来完成海量数据下的特征处理，模型训练及自动评测的。

报告人简历：现任阿里巴巴集团阿里妈妈事业部高级技术专家，领导算法部门的基础研究和 rank model 团队，为阿里广告各个业务线提供 NLP 和机器学习技术支持，同时开展前瞻性的广告技术研究，主要工作包括大数据机器学习平台建设，广告点击率预估，转化率预估，推荐算法融合，Query 分析，相关性计算等。加入阿里之前在微软亚洲研究院(MSRA)从事自然语言处理研究，主要经历包括研发了微软对联系统，研究机器翻译技术并随队在 2008 年 NIST MT 评测中获得中英翻译第一名，研究网络数据挖掘为微软 Engkoo 项目（现为必应词典）提供双语数据和名词术语翻译，研究微博的搜索排序和情感分析。在国际知名会议（ACL，IJCAI，COLING，EMNLP,SIGIR,KDD,CIKM 等）累计发表 10 余篇文章，并拥有多项美国专利。曾多次做客大学讲授高级自然语言处理研究生课程。

==========

朱小槿　University of Wisconsin-Madison　副教授

报告题目：Statistical Machine Learning for NLP （统计机器学习与自然语言处理）

摘要：Statistical machine learning offers a rich and rigorous formalism for many tasks in natural language processing. This tutorial provides a systematic introduction to the basics of machine learning and fruitful connections with NLP. In part 1, I give an overview of statistical learning, introducing the notion of models, parameters, estimation, and decision theory. In part 2, I go over probabilistic graphical models, including directed and undirected graphical models for structured learning, belief propagation and Gibbs sampling, and parameter learning. In part 3, I review a few advanced probabilistic models including Chinese Restaurant Process and Indian Buffet Process. The tutorial will be delivered in Chinese and is accessible to anyone with basic knowledge of probability and statistics

统计机器学习给自然语言处理提供了一个严谨的数学体系。我们系统地讲解机器学习的基础知识，以及它与自然语言处理的联系。第一课介绍统计学习概论，包括统计模型，参数，估计，和决策理论。第二课介绍概率图模型，包括有向和无向图模型，belief propagaion, Gibbs sampling, 以及参数学习。第三课介绍几个前沿的贝叶斯非参数模型，包括中国餐馆过程和印度自助餐过程。本讲座假定听众有本科概率与统计课程基础。

报告人简历： Xiaojin Zhu is an Associate Professor in the Department of Computer Sciences at the University of Wisconsin-Madison. Dr. Zhu received his B.S. and M.S. degrees in Computer Science from Shanghai Jiao Tong University in 1993 and 1996, respectively, and a Ph.D. degree in Language Technologies from Carnegie Mellon University in 2005. He was a research staff member at IBM China Research Laboratory from 1996 to 1998. Dr. Zhu received the National Science Foundation CAREER

Award in 2010, and Best Paper awards at ICML, ECML/PKDD, and SIGSOFT. His research interest is in machine learning, with applications in natural language processing, cognitive science, and social media.

==========

俞 栋 Microsoft Research, Redmond （微软雷德蒙研究院 高级研究员）

报告题目：Deep Learning – What, Why, and How（深度学习 – 是什么，为什么，如何实现）

摘要：In the recent years, deep learning techniques have been successfully applied to large scale real-world applications such as large vocabulary speech recognition (大词汇量语音识别), computer vision, and natural language processing. In this tutorial, I will explain what deep learning is, what the basic ideas behind it are, what the most popular models are, how to apply deep learning to real-world applications, and why it works so well in applications such as speech recognition. Although I will mainly use speech recognition as the example, the insights and techniques shall be applicable to many other applications.

最近几年，深度学习技术在包括大词汇量语音识别、计算机视觉、和自然语言理解等实际应用中获得了成功地应用。本课程讲解什么是深度学习、深度学习背后的基本思路是什么、最常用的深度学习模型有哪些、如何应用深度学习解决实际问题、以及为什么深度学习在语音识别等应用中取得很好的结果。虽然我会用语音识别作为主要例子，从中得到的认识和技术可以引用到很多实际问题中。

课程将会分为三部分:

1. 深度学习：根据、基本原理、以及与其他技术的关系
2. 基本的深度学习模型
3. 深层神经网络及其在语音识别中的应用

报告人简历： Dr. Dong Yu joined Microsoft Corporation in 1998 and the Microsoft Speech Research Group (now expanded to Conversational Systems Research Center) in 2002, where he currently is a senior researcher. His research interests span speech processing, robust speech recognition, discriminative training (区分性训练), and machine learning. He has published over 100 papers in these areas and is the inventor/coinventor of more than 50 granted/pending patents.

His most recent work focuses on deep learning and its application in large vocabulary speech recognition. The context-dependent deep neural network hidden Markov model (CD-DNN-HMM 上下

文相关深层神经网络隐马尔可夫模型) he co-proposed and developed has been seriously challenging the dominant position of the conventional GMM (高斯混合模型) based system for large vocabulary speech recognition.

Dr. Dong Yu is a senior member of IEEE. He is currently serving as a member of the IEEE Speech and Language Processing Technical Committee (2013-) and an associate editor of IEEE transactions on audio, speech, and language processing (2011-). He has served as an associate editor of IEEE signal processing magazine (2008-2011) and the lead guest editor of IEEE transactions on audio, speech, and language processing - special issue on deep learning for speech and language processing (2010-2011).

俞栋博士于 1998 年加入微软公司并于 2002 年加入微软研究院语音研究组。他目前是微软研究院高级研究员。他的研究兴趣包括语音处理、区分性训练和机器学习。他已发表了 120 多篇论文并且是 50 多项专利的共同发明人。他最近的工作集中在深度学习及其在语音识别中的应用。他引领发展的上下文相关深层神经网络隐马尔可夫模型严重地挑战了传统的基于高斯混合模型的系统在大词汇量语音识别中的地位。

俞栋博士是 IEEE 高级会员。他目前担任 IEEE 语音和语言处理专业委员会委员及 IEEE 音频、语音和语言处理会刊的副编辑，并曾于 2008-2011 年担任 IEEE 信号处理杂志的副编辑。

报名信息：

注册费： （含资料和 3 天的午餐）

1、 9 月 30 日前报名并缴费：会员 1100 元，非会员 1500 元
2、 10 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、2012 年参加过 2 次讲习班的 CCF 会员可优惠 100 元。
3、2013 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。

缴费方式：

银行转账：

开户行：建行重庆沙坪坝支行

户名：重庆大学

账号：50001053600050005883

请务必注明：姓名+ADL 重庆

报名方式：

即日起至 2013 年 9 月 30 日，报名者请填写附表并发送至：yang@cqu.edu.cn，按报名先后录取（名额有限、先报先得）。学会秘书处将与邮寄联系确认。自 10 月 1 日起，只接受现场报名。

联系人：杨小春　李宽　E-Mail：yang@cqu.edu.cn
电话：023-65106748 / 13638307036
地址：重庆市沙坪坝区沙正街 174 号 重庆大学 A 校区

CCF ADL46 报名表

《面向大数据的自然语言处理与机器学习（重庆）》

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： 单住：　　合住：		
注册费缴纳方式 √	□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 √	□注册费　□培训费　□会议费　□会务费		
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。			