中国计算机学会《学科前沿讲习班》

The CCF Advanced Disciplines Lectures

第 31 期

主题　自然语言处理与机器学习

——理论、方法与应用

2012 年 10 月 31 日－11 月 2 日 北京

随着计算机网络和移动通讯技术的快速发展和普及，面向网络文本理解和知识挖掘的自然语言处理技术正在成为学术界和产业界关注的热点。一方面，来自各类用户的海量信息充斥着巨大的网络空间，如 Facebook、QQ、Twitter、新浪微博等社交网站及 Skype、GTalk、MSN 等通信工具每天为我们记录或传送了数以十亿计用户的所思、所言，而这些数据 80％以上为自然语言文本，面对如此庞大的非结构化和半结构化的动态数据，如何针对不同需求建立有效、快捷的处理模型和实现方法？另一方面，机器学习理论和方法在近几年来发展迅速，若干模型和算法已经在自然语言处理的各个方向上发挥了重要作用。那么，面对自然语言处理的复杂任务，如何进一步研究和开发新的机器学习理论和方法？这是众多学者和企业家共同关注的问题。

本期 CCF 学科前沿讲习班将围绕面向互联网的自然语言处理和机器学习相关理论、技术及应用方法，邀请学术界和工业界的著名专家、学者系统地讲授相关学术理论知识和应用创新经验，面向有志于从事自然语言处理和机器学习基础理论研究和应用技术研发的青年学者和学生提供三天学习和交流的机会，让参加者全面了解并掌握该领域的基本概念和主要研究内容，把握研究热点和前沿，提高学术水平和研究实践能力。讲习班采用小班授课方式（60 人），中文讲解，利于答疑解惑，同时提供参加 NLP&CC2012 举办的微博情感分析与词汇语义关系评测研讨会的机会。

共同学术主任：宗成庆　　中国科学院自动化研究所

周明　　微软亚洲研究院

协办单位：微软亚洲研究院（MSRA）

日程安排、特邀讲者及报告题目：

◆ 10 月 31 日：**机器学习与机器翻译**

上午第一讲：概率图模型：表示、推断与学习

主讲人：王立威　　北京大学

下午第二讲：机器翻译及多语理解

主讲人：张　民　　新加坡信息通信研究所（I^2R）

◆ **11 月 1 日：微软亚洲研究院专题**

上午第三讲：互联网创新

组织者：周　明　　微软亚洲研究院

下午第四讲：搜索引擎基础和实战

主讲人：兰东俊　　微软亚洲互联网工程院

第五讲：如何做研究（参观、讲座和讨论）

组织者：马　歆　　微软亚洲研究院

◆ **11 月 2 日：面向互联网的自然语言处理技术与应用**

上午第六讲：信息抽取和问答系统

主讲人：赵　军　　中科院自动化所研究员

下午第七讲：文本挖掘的概率主题模型

主讲人：翟成祥　　美国伊利诺伊大学副教授

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（或研究生）900 元，非会员 1200 元
2. 9 月 30 日前缴费：会员（或研究生）1000 元，非会员 1300 元
3. 10 月 31 日缴费（含当天缴费）：会员 1200 元，非会员 1500 元

优惠办法：

1. 同一单位一次有 5 人报名者，第六个人免注册费（无论会员与否，仅对提前注册者有效，当天不予受理）。
2. 2011 年参加过 2 次讲习班的 CCF 会员可优惠 100 元。
3. 2012 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. 同时满足以上多项优惠条款时，只能选择一项。
6. 学员可免费参加 NLP&CC2012 举办的微博情感分析与词汇语义关系评测研讨会（MSA/LSR Workshop 2012。

食宿自理

缴费方式：

邮寄：北京 2704 信箱，邮编：100190　收款人：中国计算机学会，

银行转账：开户行：北京银行北京大学支行；户名：中国计算机学会

帐号：01090519 5001 201 097 020 28

请务必注明：姓名 TCCI-ADL

现场：报到时缴纳

报名方式：

即日起至 2012 年 9 月 30 日，报名者请填写附表并发送至：zlu@nlpr.ia.ac.cn，按报名先后录取。学会秘书处将与邮寄联系确认。自 10 月 1 日起不再接受通讯报名，只接受 10 月 31 日现场报名。

联系人：陆征　E-Mail: zlu@nlpr.ia.ac.cn

电话：　010-82614468

地址：北京市海淀区中关村东路 95 号 中科院自动化所模式识别国家重点实验室

日程安排

2012 年 10 月 31 日：机器学习与机器翻译

◆ **上午： 8:30-9:00 开班仪式、合影**

9:00-12:00 第一讲：概率图模型：表示、推断与学习

主讲人：北京大学教授 王立威

第一课 09:00-10:20 机器学习基础 Elements of Machine Learning
休 息 10:20-10:40
第二课 10:40-12:00 概率图模型 Probabilistic Graphical Models

◆ **下午：13:30-18:00 第二讲：机器翻译及多语理解**

主讲人：新加坡信息通信研究所(I2R) 张 民

第一课 13:30-14:00 自然语言处理和机器翻译基础：预备知识、基本概念、相关学科和背景知识
第二课 14:30-15:30 机器翻译综述：前世和今生、主要研究方法和主要的公共模块
休 息 15:30-16:00
第三课 16:00-17:00 统计机器翻译：基于词、短语和句法的方法
第四课 17:00-17:30 机器翻译未来：下一步发展趋势、研究方法和产业化
第五课 17:30-18:00 Free QA（互动课）

2012 年 11 月 1 日：NLP2.0 — 互联网创新与应用

◆ **上午： 8:30-12:00 第三讲：互联网创新**

第一课 08:30-10:00 社会关系网的文本挖掘和应用 周明、刘晓华，韦福如
休 息 10:00-10:30
第二课 10:30-11:00 由文本到图像的变换 王欣靖
第三课 11:00-11:30 基于知识库的短文本概念化及其应用 宋阳秋
第四课 11:30-12:00 对象级别的互联网搜索及交互式知识挖掘 聂再清

◆ **下午： 13:30-15:30 第四讲：搜索引擎基础和实战**

主讲人：兰东俊

第一课 13:30-14:00 搜索引擎基本原理
第二课 14:00-14:30 搜索引擎的实战经验
第三课 14:30-15:00 搜索引擎的未来发展趋势和研究方向
休 息 15:30-16:00

16:00-18:00 第五讲：如何做研究（参观、讲座、讨论）

组织者：马歆、严峻、段楠

2012 年 11 月 2 日： 面向互联网的自然语言处理技术

◆ 上午：**8:30-12:00** **第六讲：信息抽取和问答系统**

主讲人：中科院自动化所研究员 赵 军

第一课	08:30-09:10	信息抽取
第二课	09:10-09:50	观点信息抽取
休 息	09:50-10:10	
第三课	10:10-10:50	问答系统
第四课	10:50-11:30	社区问答系统
第五课	11:30-12:00	互动课

◆ 下午：**13:30-15:30** **第七讲：文本挖掘的概率主题模型**

主讲人：美国伊利诺伊大学副教授 翟成祥

第一课	13:30-14:00	概率主题模型基本介绍
第二课	14:00-15:00	概率主题模型在文本挖掘中的重要应用
第三课	15:00-15:30	未来的研究展望

15:00-15:30 **结业式**

◆ 下午：**15:30-18:00** **NLP&CC2012 微博情感分析与词汇语义关系评测研讨会**

讲座与专家介绍

第一讲：概率图模型：表示、推断与学习

主讲人：王立威，北京大学 教授

主要内容：我们通过介绍机器学习的基本思想引入概率图模型。首先描述为何概率图模型适于表示机器学习问题，以及概率图模型的表示能力，包括有向图 Bayes 网和无向图 Markov 网。接下来我们转入如何利用概率图模型进行推断。报告将深入浅出地介绍概率图模型常用推断算法，包括著名的 belief propagation 算法，马尔科夫链蒙特卡罗(MCMC)方法等。同时，我们还将简要指出概率图模型推断的本质困难性以及近似的必要性。最后，我们介绍如何从数据中学习概率图模型，重点是图结构学习的常用算法。包括基于约束的算法和基于模型得分的算法等，并讨论它们各自的优点与不足。

王立威 北京大学信息学院智能科学系教授。分别于 1999 年、2002 年于清华大学电子工程系获本科和硕士学位。2005 年于北京大学数学学院获博士学位。自 2005 年起在北京大学信息学院任教。他的主要研究兴趣为机器学习理论与算法，对 boosting、主动学习等开展了深入研究。在机器学习顶级会议 NIPS, COLT, ICML 和顶级期刊 JMLR, IEEE Trans. PAMI 发表论文多篇。2010 年入选 AI's 10 to Watch。

第一讲：机器翻译及多语理解

主讲人：张　民，新加坡信息通信研究所（I2R）研究员

主要内容：实现不同自然语言之间的无障碍信息交流一直是人类的梦想。随着人类社会步入全球化时代和互联网以及社交网络的迅猛发展，这种需求尤为迫切。有鉴于此需求，近几年来，机器翻译和多语理解技术的研究和产业化越来越成为学术界和产业界的关注热点之一。本课程即在这一背景下，对机器翻译和多语理解技术进行系统、全面的介绍，包括其基本概念、需要解决的问题、研究背景和历史、相关学科、主要方法、最新研究进展、下一步发展趋势和产业化等等，使学生能够对这个学科的各个方面有较为系统的认识。

Dr. Min ZHANG is a research scientist at the Institute for Infocomm Research, Singapore and the Program Investigator of statistical machine translation team at the institute. His research interests include Machine Translation, Information Extraction, Information Retrieval and Machine Learning for Natural Language Processing. He has authored more than 120 papers in leading journals and conferences. He is the vice president of COLIPS, a steering committee member of PACLIC and a member of AFNLP and ACL. He supervises Ph.D students at the National University of Singapore and Harbin Institute of Technology. Dr. Min ZHANG joined the Institute in Dec. 2003. He received his Ph.D. degree from Harbin Institute of Technology in 1997. From Dec. 1997 to Aug. 1999, he worked as a postdoctoral research fellow in Korean Advanced Institute of Science and Technology in Korea. He began his academic and industrial career as a researcher at Lernout & Hauspie Asia Pacific (now Nuance) in Sep. 1999. He joined Infotalk Technology (Singapore) as a researcher in Jan 2001 and became a senior research manager in 2002.

第三讲：互联网创新

报告 1：社会关系网的文本挖掘和应用

主讲人：周　明、刘晓华、韦福如，微软亚洲研究院研究员

主要内容：最近几年，我们看到了研究人员、工程师和公司对利用社会网络来进行数据挖掘，商业智能，搜索和广告营销的兴趣越来越大。但是数据的爆炸以及社会网络的特殊的语言表达现象对这些努力构成了巨大的挑战。使用现有的面向标准书面语言的文本挖掘技术不能获得满意的结果。本讲座以推特为例，将介绍一组文本挖掘技术来从大规模的实时的推特中提取关键的信息来支持后续的更多的在社会网络中进行的数据挖掘和搜索任务。具体的讲座内容将包括推特和微博文本的预处理、命名实体的识别和情感分析。然后介绍再次基础上进行的搜索和文摘的工作。

周明，微软亚洲研究院自然语言组主任，高级研究员。1991 年哈工大博士毕业，1991-1993 年在清华任博士后，随后任副教授至 1999 年，其间 1996-1999 年在日本高电社领导中-日机器翻译的研发。1999 年加入微软研究院任研究员。2001 年起任自然语言组主任并曾于 2004 年短期兼任语音组主任。

刘晓华，微软研究院自然语言组研究员。作为主要研究人员曾参加微软英库和必应词典的研究工作。现从事社会网络的文本挖掘和搜索研究。发表 ACL、EMNLP、Coling、AAAI、IJCAI 文章 10 余篇。

韦福如，微软研究院自然语言组副研究员。从事社会网络的情感分析、摘要和自然语言问答的研究工作。发表 ACL、SIGIR、KDD, Coling、AAAI、IJCAI 文章 10 余篇。

报告 2：由文本到图像的变换

主讲人：王欣靖，微软亚洲研究院副研究员

主要内容：一图胜过千言万语。将文本可视化，即为一个词赋予一幅图像可以让词的意思一目了然，从而提升用户体验。本工作首先展示如何为单个词找出代表性的图像，并由此推广到大规模可视化互联网上的本体词汇(identity)。

王欣靖，于 2005 年取得清华大学博士学位，目前为微软亚洲研究院互联网搜索与挖掘组的研究员，从事大规模网络图像理解方面的研究。

报告 3：基于知识库的短文本概念化及其应用

主讲人：宋阳秋, 微软亚洲研究院副研究员

主要内容：在互联网高速发展的时代，短文本处理技术越来越多地被应用在搜索、广告、图像标签和微博等数据当中。由于短文本缺乏统计信息，数据相对于长文更加稀疏、模糊且有更多的噪音。因此，我们需要建立知识库来令计算机更好的处理短文本。在本工作中，我们使用一个基于概率建模的知识库 Probase。Probase 拥有数以百万级的概念。这些概念是由计算机自动从数十亿网页中自动抽取出来。在此基础上，我们提出一个基于概率的框架，系统地对短文本中的实例和属性进行概念化。利用短文本概念化的结果，我们通过实际数据验证了本方法可以更好地帮助搜索、广告进行相关度匹配。另外，通过对查询日志和 Twitter 数据的聚类，我们进一步验证了该方法的有效性。

宋阳秋，于 2003 年和 2009 年分别获得清华大学自动化系本科和博士学位，并于 2010 年加入微软。研究方向为机器学习、数据挖掘、信息检索和可视化。

报告 4：对象级别的互联网搜索及交互式知识挖掘

主讲人：聂再清，微软亚洲研究院研究员

主要内容：互联网中蕴含着大量的关于现实世界对象(例如人物、机构、和地点)的结构化信息。我们在探索一种全新的搜索体验：抽取和集成网页上各式各样的对象信息，让用户能够进行对象级别的信息搜索和浏览。对象级别搜索的一个显著优点是可以利用对象的语义信息，采用直接或者聚合的结果来响应复杂查询。在本次讲座中，我将以人立方和微软学术搜索为例介绍互联网对象级别搜索的用户体验及其关键技术。

聂再清，于 2004 年 4 月加入微软亚洲研究院互联网搜索与挖掘组，负责对象级别互联网搜索引擎的研发工作包括数据抽取，集成和检索。人立方关系搜索和微软学术搜索是对象级别搜索技术的两个成功应用实例。聂再清于 1996 和 1998 年在清华大学计算机系获学士和硕士学位，2004 获美国亚利桑那州立大学计算机科学专业博士学位。

第四讲：搜索引擎基础和实战

主讲人：兰东俊，微软亚洲互联网工程院项目经理

主要内容：搜索引擎需要服务的网页数量高达百亿甚至千亿，而从用户拿到的输入只有搜索框中的几个查询词。从巨大的结果集中、根据很少的输入判断用户的意图，把和用户意图最相关的内容呈现给用户，同时这一切都必须要在几秒内完成，其挑战性和系统复杂性不难想见。在本讲座中， 我们将介绍搜索引擎的现状， 搜索引擎的基本架构，搜索引擎的技术难点及工程实践。通过以上的内容，听众将会了解搜索引擎的基本原理、常见方法和工程实践。

兰东俊，微软亚洲互联网工程院的项目经理。现在微软广告平台的广告相关性团队工作。在加入广告团队前，兰东俊在微软必应搜索引擎的基础设施团队工作，管理数万台机器和数百 PB 的数据，为微软在线服务部门提供分布式存储和计算平台。加入微软之前，兰东俊是 IBM 中国研究院的资深研究员。兰东俊毕业于清华大学电子工程系。

第五讲：如何做研究

主讲人：马歆、严峻、段楠，微软亚洲研究院

主要内容：每一位学生在踏上研究之路时，都希望能得到来自前辈或者学长的建议与指导。微软亚洲研究院的研究员在亲自指导学生以及自己做研究的过程中，也总结了一系列“如何做研究”的经验。在这个讲座中，几位报告者将为同学们分享自己的研究和成长经历，包括从认知自己和怎样在研究中发挥自己的优势到如何发表第一篇学术论文，从做研究的基本功训练到如何选题、制定方向和解题。

马歆，微软亚洲研究院学术合作部资深经理。2001 年加盟微软亚洲研究院。现负责制定和开展微软亚洲研究院与亚洲地区高校、学术机构在人才培养与合作的战略和相关项目，包括微软学者奖学金，“明日之星”实习生项目，联合培养博士生项目等。此外，还负责微软亚洲研究院亚太区文化遗产数字化保护研究计划，与北京故宫博物院，台北故宫博物院、敦煌研究院等多家文化遗产单位和博物馆建立合作研究项目。

严峻，毕业于北京大学数学系，获得博士学位。研究方向为模式识别和信号处理。目前是微

软亚洲研究院机器学习组研究员，研究方向为大规模数据挖掘、机器学习和计算广告学。已在 SIGKDD、 SIGIR、WWW、ICDM 和 TKDE 等国际会议发表了 50 余篇论文。

段楠，2011 年博士毕业，系天津大学和微软亚洲研究院联合培养的第一名博士。博士期间在微软亚洲研究院从事统计机器翻译的研究。现为微软亚洲研究院自然语言计算组博士后研究员，从事自动问答和搜索的研究。在 ACL、EMNLP、COLING 等自然语言处理会议中发表 10 余篇学术论文。

第六讲： 信息抽取和问答系统

主讲人：赵 军，中国科学院自动化研究所研究员

主要内容：问答系统被认为是下一代搜索引擎的重要形态，而信息抽取是支撑问答系统等互联网应用的关键技术之一。本课程将围绕信息抽取和问答系统两个研究方向，系统介绍其中的基本概念、主要方法、最新研究进展、需要解决的问题和发展趋势，使听者能够对信息抽取和问答系统研究领域的重点问题和主要方法有较为系统的了解。

赵军，研究员，博士生导师。1998 年在清华大学计算机科学与技术系获得博士学位。1998 年—2002 年在香港科技大学计算机科学系做访问学者。2002 年 5 月至今在中国科学院自动化研究所模式识别国家重点实验室工作。研究方向为自然语言处理、网络信息抽取和问答系统等。主持多项国家自然科学基金、863 计划、中国出版集团科技项目等的研究工作。在 ACL、SIGIR、CIKM、IJCAI、EMNLP、CoNLL 等顶级国际会议上发表一系列学术论文。
主页：http://www.nlpr.ia.ac.cn/cip/jzhao.htm。

第七讲：文本挖掘的概率主题模型

主讲人：翟成祥，美国伊利诺伊大学副教授

Abstract：Statistical Topic Models (also known as probabilistic topic models, or just topic models) have recently been successfully applied to many text mining problems. They can be used to naturally model the topics in unstructured/semistructured text collections, and extract various types of topical patterns from text. A great deal of recent work have shown that topic models not only have a solid theoretical foundation, but also offer solutions to many practical text mining tasks. This lecture will systematically review the recent progress in applying statistical topic models to text mining. We will first introduce the basic probabilistic topic models, and then discuss a number of extensions of the basic models and their applications in text mining. In particular, we will discuss in depth how to use topic models for contextual text mining where context variables such as time, location, authors, and sources are considered when analyzing topics in text. Sample results on a wide range of applications such as spatiotemporal topic trend analysis, opinion integration and summarization, and event impact analysis will be presented.

Chengxiang Zhai is an Associate Professor of Computer Science at the University of Illinois at Urbana-Champaign, where he also holds a joint appointment at the Institute for Genomic Biology, Statistics, and the Graduate School of Library and Information Science. He received a Ph.D. in Computer Science from Nanjing University in 1990, and a Ph.D. in Language and Information Technologies from Carnegie Mellon University in 2002. He worked at Clairvoyance Corp. as a Research Scientist and a Senior Research Scientist from 1997 to 2000. His research interests include information retrieval, text mining, natural language processing, machine learning, and bioinformatics. He is an Associate Editor of ACM Transactions on Information Systems, and Information Processing and Management, and serves on the editorial board of Information Retrieval Journal. He is a

program co-chair of ACM CIKM 2004 , NAACL HLT 2007, and ACM SIGIR 2009. He is an ACM Distinguished Scientist, and received the 2004 Presidential Early Career Award for Scientists and Engineers (PECASE), the ACM SIGIR 2004 Best Paper Award, an Alfred P. Sloan Research Fellowship in 2008, and an IBM Faculty Award in 2009. More details about Dr. ChengXiang Zhai can be found at his personal website http://www.cs.uiuc.edu/homes/czhai/

CCF ADL 报名表
《自然语言处理与机器学习》

<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 colspan="2"></td><td>Email</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>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">我申请参加本届研究峰会并承诺按主办单位的规定参加。</td></tr>
</table>

说明：

1、 会员号：不填写会员号，按非会员处理，

2、 仅需要组织者代位安排住宿是，填写“安排住宿”一栏。

3、 发票：“发票付款单位”如空，则认为同“任职单位”。

4、 发票项目如不选择，则认为是“会议费”。

5、对会员优惠 400 元，仅对开班前三个月前入会者有效。如果正在申请入会，请填写“正在办理”，享受 200 元优惠。