中图分类号：TP391.4 秘密☆永久

论 文 编 号：10006SY1706404

北京航空航天大學

硕 士 学 位 论 文

恶意网络爬虫检测与对抗技术的研究及实现

作 者 姓 名 张浩凌

学 科 专 业 网络空间安全

指 导 老 师 李舟军 教授

何跃鹰 处长

培 养 院 系 计算机学院

[– This page is a preset empty page –]

Research and implementation of malicious web crawler detection and confrontation technology

A Dissertation Submitted for the Degree of Master

Candidate : Zhang Haoling

Supervisors: Prof. Li Zhoujun
He Yueying

School of Computer Science and Engineering

Beihang University, Beijing, China

[– This page is a preset empty page –]

中图分类号：**TP391.4** 秘密 ☆ 永久

论 文 编 号：**10006SY1706404**

硕　士　学　位　论　文

恶意网络爬虫检测与对抗技术的研究及实现

作者姓名	张浩凌	申请学位级别	工学硕士
指导老师姓名	李舟军	职　　称	教授
学科专业	网络空间安全	研究方向	网络空间安全
学习时间自	2017 年　09 月　01 日	起至	2020 年　01 月　31 日止
论文提交日期	2019 年　11 月　25 日	论文答辩日期	年　　月　　日
学位授予单位	北京航空航天大学	学位授予日期	年　　月　　日

[– This page is a preset empty page –]

关于学位论文的独创性声明

本人郑重声明：所呈交的论文是本人在指导教师指导下独立进行研究工作所取得的成果，论文中有关资料和数据是实事求是的。尽我所知，除文中已经加以标注和致谢外，本论文不包含其他人已经发表或撰写的研究成果，也不包含本人或他人为获得北京航空航天大学或其它教育机构的学位或学历证书而使用过的材料。与我一同工作的同志对研究所做的任何贡献均已在论文中作出了明确的说明。

若有不实之处，本人愿意承担相关法律责任。

学位论文作者签名：_____ 日期： 年 月 日

学位论文使用授权

本人完全同意北京航空航天大学有权使用本学位论文（包括但不限于其印刷版和电子版），使用方式包括但不限于：保留学位论文，按规定向国家有关部门（机构）送交学位论文，以学术交流为目的赠送和交换学位论文，允许学位论文被查阅、借阅和复印，将学位论文的全部或部分内容编入有关数据库进行检索，采用影印、缩印或其他复制手段保存学位论文。

保密学位论文在解密后的使用授权同上。

学位论文作者签名：_____ 日期： 年 月 日

指导教师签名：_____ 日期： 年 月 日

[– This page is a preset empty page –]

摘　要

随着各类基于大数据和 AI 的应用的兴起，能够快速廉价地获取大量有效数据的能力，成为互联网时代企业和个人竞争力的体现。因此，网络爬虫在数据收集收集方面的重要性逐渐凸显出来。但是，恶意爬虫同样给互联网用户和互联网服务提供者带来巨大的困扰，这些爬虫或者多线程高并发耗尽服务器的带宽和计算资源 [1]，或者爬取个人敏感信息、高价值的商业数据用于不法用途。

因此，构建一个成熟有效的反爬虫系统，成为一个亟待解决的问题。但遗憾的是，传统的反爬虫方式过于保守和被动，在漏检率和误检率居高不下的情况下，往往只能通过单一的反制手段来限制爬虫（如 IP 封锁，访问频率限制，虚假数据等），在反爬虫的战争中收效甚微。为此，本文提出了一种新型的反爬虫的系统，融合爬虫检测技术，爬虫行为分析和溯源技术，并通过动态符号执行、模糊测试以及污点分析等二进制漏洞挖掘方法，挖掘爬虫使用的框架、处理脚本，无头浏览器驱动程序驱动程序的漏洞，并通过返回恶意的攻击载荷，对运行恶意爬虫的主机进行反向攻击，最终诱使恶意爬虫进程奔溃，甚至可能获取到恶意爬虫主机上的敏感数据以及系统权限。

此外，本文将遵循上述的设计思路，实现该反爬虫系统的原型系统 Crawler-Net(捕虫网)，并将 Crawler-Net 部署在模拟的业务系统上，使用具有不同请求策略的爬虫流量混合正常的用户访问进行测试，从漏检率、误检率、系统性能损失等多方面的指标来评估反爬虫系统的性能。

关键字：恶意爬虫，反爬虫机制，无头浏览器，爬虫检测，爬虫溯源，动态符号执行，模糊测试

Abstract

With the flourish of the Applications based on the Big Data and Artificial Intelligence, it has aroused our attention in how to collect numerous valid data rapidly at the least cost, which could be regarded as an aspect of competitive competence both for the individuals and the companies. And thus, the power of the crawlers in collecting data has been addressed. However, there are plenty of malicious crawlers filling in the cyberspace, try to deplete the servers' resource with endless requests concurrently, or theft the sensitive individual's information or commercial data for illegal use.

Therefore, we are supposed to build up a feasible, active and robust anti-crawling system to stop those malicious crawlers, while most of the contemporary anti-crawling systems are using passive strategies. Those anti-crawling systems could only harness single mechanism to block the crawlers(such as IP blocking, limit the frequency of requests, fake data), and that does not seem to be effective in most cases. In this paper, we propose an anti-crawling system, which combines the crawler detection with the analysis of crawler's behaviors, then utilizes the dynamic symbolic execution, fuzzing and dynamic taint analysis , to excavate the vulnerabilities of the frameworks, the handling scripts and the headless web-driver binaries of the crawlers. Ultimately, malicious payloads would be generated to feed the crawler when it crawls our web pages, leading it to crash down or even obtain the sensitive data as well as the system privileges from the crawler's host.

In addition, at the end of the paper, I would build up the prototype of the anti-crawling system called "Crawler Net" according to the technologies and mechanisms mentioned before, with the deployment to the web application in a simulated production environment. Furthermore, several tests would be conducted with the requests generated by various types of crawlers and the requests from normal user to testify its ability to block the crawlers and the performance once deployed.

Key words: malicious crawler, anti-crawling system, headless browser, crawler detection, crawler trace, dynamic symbolic execution, fuzzing test

目　录

图 清 单

表 清 单

主要符号表

E 能量

m 质量

c 光速

P 概率

T 时间

v 速度

[– This page is a preset empty page –]

第一章　绪论

1.1　研究背景及意义

机器学习，数据挖掘等大量数据依赖型的技术在高速发展的同时，对于相关数据的质和量都提出了更高的要求。网络爬虫，又称网络蜘蛛或者网络机器人，在这样一个数据消费的时代，扮演着数据搬运者和传递者的角色。在其运行的生命周期中，往往会按照开发者预设的规则，爬取指定的 URL 地址或者 URL 地址列表，并将获取到的数据预处理成标准化的格式 [2]。

普通的网络爬虫并没有危害，相反，搜索引擎存储数据的来源都是基于大量分布式网络爬虫爬取得到的结果，网络爬虫在数据传递过程中起到至关重要的重要。但是多线程高并发的失控爬虫，针对网站隐私数据的窃密爬虫，以及不遵守爬虫道德规范的恶意爬虫，都对网络空间健康的生态环境提出了巨大的挑战。

爬虫失控往往是具有多线程操作的通用型爬虫，未能控制爬行时间间隔，或者因为未添加地址环回检测的处理逻辑，在处理特殊的地址链接时陷入死循环中。失控的爬虫通常给网站的性能资源和带宽资源带来巨大的消耗，甚至会影响正常人类用户的体验，这样的爬虫对网站的影响相当于是 DDOS 攻击 [1]。每年的三月份，是失控爬虫的高发期，原因正是因为大量的硕士在写论文时会爬取网站数据用于数据挖掘或者机器学习。

此外，部分互联网公司会爬取其他同行的优质数据用于商业用途，在给其竞争对手带来经济效益的损失的同时，还会促进产业内部恶性竞争的循环。例如，马蜂窝网站的用户评论数据涉嫌造假事件。甚至在一些情况下，恶意爬虫甚至会爬取敏感个人信息用于不法用途，例如某大数据公司非法爬取个人信息被起诉一案。

某些由黑客或者 APT 组织控制的爬虫，在爬取某些 CMS 系统或者 web 中间件的版本信息后，会使用相应的攻击向量攻击脆弱主机 [6]，给网站服务提供者及使用者造成巨大的损失。网络空间一直是爬虫与反爬虫战斗的前线，随着反反爬虫技术的不断迭代更新，传统的静态、单一、被动与非实时的反爬虫技术难以与之对抗，或者又因为部署成本和部署带来的性能损失而被束之高阁。

1.2 国内外研究现状

在数据需求不断增加的大数据时代，爬虫技术的发展也日新月异。与此同时，传统单一静态的爬虫识别技术已经无法满足现阶段的需求，研究相关领域的学者也一直在为反爬虫领域提供新的技术和新的思路。

Guo W 等人首先针对传统与单个 http 请求进行处理的爬虫识别算法提出新的改进，在他们的文章中，他们首先使用了 session 粒度的爬虫识别算法，重点关注人类访问 session 与爬虫访问 session 中对应的 url 请求资源类型（样式表，html，图片等）比例不同的特性，并对采集得到的相应的日志进行非实时的线下处理，提取出相关特征作为分类依据。这是最早的基于 session 的反爬虫机制。

Derek Doran 在他们的文章中使用了和 Guo W 等人类似的方法，他们在使用 url 请求资源类型比例作为重要特征的同时，引入了离散马尔科夫链的概率模型，并使用该模型得出的 log 概率来判定访问时来自于人类还是爬虫。但是总体而言，该文章提出的模型并没有过多的创新，而且马尔科夫链的概率模型的计算过程，需要消耗大量的计算资源，不能够应用在实时的爬虫检测上。

与此同时，为了将爬虫识别模型应用到真实的业务场景中，而不仅仅作为一种离线的验证算法。Andoena Balla 等人提出了实时的爬虫识别算法。在他们的文章中，他们还引入了如下的 session 粒度下的特征：（1）head request 的百分比（2）2xx 返回的比例（3）3xx 返回的比例（4）页面资源的访问比例（5）夜间访问的比例（6）访问两个页面之间的平均时间（7）其他二进制文件请求的比例。

Andoena Balla 的对 session 粒度使用了比较完备的特征，为后来基于 session 粒度的爬虫识别的相关研究，提供了有价值的参考。但对于如何进行 session 的分类以及如何处理 session 过长的问题，该文章并没有提供合格的解决方案。Yi Liu 在他们的文章中提出了一种解决 session 过长问题的方法。他们采用了滑动窗口的机制，对每一次处理的 http 请求做了相关的限制，对在最大程度上保证了处理的 http 请求的相关性，并在一个窗口内使用 SVM 来区分普通用户和爬虫。此外，他们还以他们的模型为基础，实现了一整套支持实时爬虫检测的系统。但是该文章在 session 粒度的特征方面，并没有充分利用 session 中的包含的信息。

Shengye Wan 在他的文章中综合了现有的反爬虫技术，提出一个名为 PathMarker 的

模型。PathMarker 会将 url 地址信息以及当前访问的用户信息加密，并替换掉原有的 url 地址。由此标注每个请求所对应的 session，并利用之前的研究中使用的 session 特征，来区分爬虫和人类。这种方案可以在很长的时间窗口内持续地追踪爬虫在某个网站的爬行轨迹，可以用于分析爬虫的爬行目标和爬行策略，并且在某种程度上可以追踪使用分布式 ip 池的爬虫群。PathMarker 在 session 分类良好的情况下能够产生较好的爬虫检测效果，但是其缺点也很明显，需要修改所有返回请求中的所有链接地址，在真实的生产环境下难以提供灵活的部署方案。此外，一旦访问的爬虫使用对抗性的爬虫策略，PathMarker 的 session 分类效果将不再准确，从而影响最终的爬虫识别效果。

总体而言，现有的反爬虫技术主要注重于爬虫识别技术的发展，发现可疑爬虫后的阻断方法，往往是单一的阻断或者使用 captcha 机制进行验证，无法对恶意爬虫的作者起到威慑的作用。在爬虫识别技术的主流技术中，在单粒度上识别的检测技术，一旦遇到采用字段变异的爬虫，便无法发挥应有的作用。而在 session 粒度上识别的检测技术，在实时性检测上，往往都有较大的性能消耗。除此之外，大部分的爬虫识别模型都没有考虑到爬虫可能采用的反反爬虫手段，对于一些有着特殊对抗策略的爬虫，其识别效果将大打折扣。

表 **1** 已有研究成果比较

论文题目	实时性	静态信息利用	动态信息利用	检测粒度	性能损耗	爬虫对抗
Protecting Web Contents Against Persistent Crawlers	Y	refer, user agent, cookies	N	session	高	N
Web robot detection techniques based on statistics of their requested URL resources	N	user agent,URL pattern	N	单粒度/session	高	N
RESEARCH ON AN ANTI-CRAWLING MECHANISM AND KEY ALGORITHM BASED ON SLIDING TIME WINDOW	Y	N	N	session	中	N
Detecting web robots using resource request patterns	N	N	N	session	高	N
Real-time Web Crawler Detection	Y	N	N	session	中	N
Our paper (Crawler-Net)	Y	user agent, http headers value, http headers key order	Y	单粒度/session	中	Y

1.3　研究目标和内容

本文的研究目标是提出并实现一种新型的针对恶意爬虫的检测和攻击的对抗性技术，用于在爬虫实时访问阶段，对爬虫进行实时监测与阻断，并针对识别出的、具有恶意探测行为和攻击行为的爬虫，在检测其相应的 webdriver 的类型和版本后，生成相对应的攻击载荷。

并在该技术的基础上，实现恶意爬虫识别与攻击系统。该系统的目标功能为：1）捕获到目标网站的所有访问请求，并实时地将其按照 session 粒度进行分类；2）通过不同粒度下的数据分析，检测和识别 HTTP 请求中出现的爬虫；3）根据设定阈值，从爬虫 session 中筛选出恶意行为的爬虫；4）利用事先生成的攻击载荷攻击恶意爬虫，并收集相应攻击返回。

基于以上研究目标，本文的主要研究内容包括以下几个部分：

1.3.1　对抗样本下 session 分类方法的研究

在普通的爬虫和正常的人类访问行为下，针对给定的 http request 序列进行 session 分类并不是一个困难的问题。通常我们在某一特定时间段内，将同一 ip 来源或者使用同一 cookie 的 http request 分类为同一个 session，但是在存在对抗行为的恶意爬虫面前，这样粗糙的分类方式通常难以获得令人信服的分类结果。参考利用 http 请求的静态特征以及利用 javascript 执行得到的动态特征，以此得到一个更加精确的 session 分类结果，是我们研究关注的重要内容之一。此外，针对一个真实生产环境下收集的实时数据，如何实现实时的 session 分类处理，以及各类参数和阈值（如 session 分类的时间长度和序列长度）的合理设定，也是我们研究的一部分。

1.3.2　基于 session 粒度的爬虫识别与检测

传统基于单个 http request 的爬虫识别方法，仅仅基于一些 http 字段的静态特征，很容易使用一些额外的工具和算法对 http 字段进行变异，从而逃避传统反爬虫程序的检测。因此，我们考虑到从 session 中额外提取爬虫的特征，并参考原先的静态特征，共同作用于爬虫识别。目前已经有一些论文讨论到基于 session 的爬虫识别，并给出了相关了 session 粒度下的爬虫特征 [4] [6] [7]，但是这些文章提出的一些特征存在一些问题：一部分特征实时检测的性能开销过大；一部分特征在数据集中并不显著，或者对恶意爬虫的支持不佳；因此，通过本次研究，我们还会得到一系列 session 的特征以及这些特

征的相关阈值设定。

1.3.3 基于隐式浏览行为的爬虫识别

正常的人类浏览行为下，除了访问频率和访问间隔于爬虫存在差异之外，还会存在一部分隐式的浏览行为 [9]。例如，在相关页面插入一些不在浏览器上渲染但是存在于 html 代码中的链接，在正常访问下不会被触发，但是遇到了爬虫的 html link parser，则可能会被访问，由此我们可以判定访问该链接的请求以及对应的 session 均是来自于爬虫的。如何能够利用人类隐式的浏览行为，来进行更为精细准确的爬虫识别，也是本研究的重点。

1.3.4 针对爬虫 webdriver 的自动化漏洞挖掘技术

一旦识别出恶意爬虫，在此基础使用 javascript 探测其 webdriver 版本以及可能存在的漏洞类型，并通过此漏洞生成相应的攻击载荷实现对爬虫的攻击，这是我们系统的最终的目的。如何通过已知的 webdriver 版本，来生成有效的攻击载荷，我们需要借助于已经成熟的自动化漏洞挖掘技术。虽然单个的 webdriver 程序体积很大，但是因为我们目标的 webdriver 往往是开源的。因此，本课题将研究在在获得源代码的前提下，如何对程序功能进行分割或者采用一些其他的简化技术，预防可能出现的路径爆炸问题，并尽可能多地找到一些漏洞。

1.4 本文的组织结构

本文的组织结构如下：第一章为绪论，总体介绍爬虫技术，爬虫检测技术等相关概念，及研究恶意爬虫对抗技术的重要意义。并总结国内外工业界和学术界针对反爬虫技术的研究现状，并指出目前已经工具及分析方法的局限性，提出本文的研究目标和内容。

第二章为主流爬虫技术的研究部分，通过对已有的爬虫技术的研究，对主流的爬虫依据其特性进行分类。并概述目前爬虫技术中使用的基本的爬行策略以及针对现有反爬虫技术发展而来的伪装策略，分析进行爬虫识别过程中的主要挑战和难点。最后重点分析具有动态处理功能的爬虫底层使用的无头浏览器技术及该技术的特性，讨论了针对无头浏览器的对抗技术实施的可能性。

第三章为恶意爬虫检测技术的研究部分，爬虫的检测技术主要从三个方面入手：（1）

单粒度模式下的检测技术（2）基于 session 粒度的检测技术（3）基于隐式浏览行为的爬虫识别。其中 session 分类与特征提取是本文的核心研究内容之一。并在第三章的最后，介绍了我们的爬虫检测技术在三个不同数据集上的检测效果。

第四章为恶意爬虫对抗技术的研究部分。重点介绍了四种对抗技术:（1）资源耗尽型的对抗技术（2）进程 crash 类型的攻击技术（3）数据窃取类型的攻击技术（4）爬虫数据流追踪技术。爬虫检测技术产生的结果将作为生成对抗策略的依据，对抗策略将决定对抗技术的种类以及相应攻击载荷的生成。

第五章为恶意爬虫对抗系统的架构研究和实现部分。通过对整个对抗系统运行流程的分析，设计了由四部分组成的具有较强健壮性的恶意爬虫对抗系统，并具体地介绍四部分间协同合作的运作方式。

第六章为恶意爬虫对抗系统的评估和分析，将使用多策略爬虫工具测试整个系统的识别率和误报率，以及在高并发情况下的性能损失比率，并进一步分析产生这样结果可能的原因。

最后是整篇文章的结论部分，总结全文的工作，对本文仍然存在的不足之处提出了更为深远地思考，同时包含了对未来爬虫、反爬虫技术互相迭代的展望。

第二章　说明

Again，这是北航论文 LaTeX 模板（CTeX-Based）BUAAThesis。

本 LaTeX 模板为北航研究生学位论文模板，适用于理工类博士、学术硕士和专业硕士。本 LaTeX 模板参考自 2015 年 8 月版北航《研究生手册》，具体要求请参见各自的《手册》，最终成文格式需参考学院要求及打印方意见。本模板中大量内容和说明直接摘抄自《手册》（2015 年 8 月版），基本覆盖了论文内容和格式方面的要求。

本模板已上传GitHub，该仓库中同时也包含了相应的 Word 模板。

2.1　宏包使用

请将以下文件与此 LaTeX 文件放在同一目录中：

buaa.cls	▷ LaTeX 宏模板文件
buaa_mac.cls	▷ LaTeX 宏模板文件 (For Mac with XeLaTeX)
GBT7714-2005.bst	▷ 国标参考文献 BibTeX 样式文件 2005
GBT7714-2015.bst	▷ 国标参考文献 BibTeX 样式文件 2015
logo-buaa.eps	▷ 论文封皮北航字样
head-doctor.eps	▷ 论文封皮北博士学位论文标题
head-master.eps	▷ 论文封皮北学硕学位论文标题
head-professional.eps	▷ 论文封皮北专硕学位论文标题
tex/*.tex	▷ 本模板样例中的独立章节

通过 \documentclass[<thesis>,<permission>,<printtype>,<ctexbookoptions>]{buaa} 载入宏包：

thesis ▷ 论文类型 (thesis)，可选值：

a) 学术硕士论文（master）[缺省值]

b) 专业硕士论文（professional）

c) 博士论文（doctor）

permission ▷ 密级 (permission)，可选值：

a) 公开（public）[缺省值]

b) 内部（privacy）

c) 秘密（secret=secret3）

— c.1) 秘密 3 年（secret3）

— c.2) 秘密 5 年（secret5）

— c.3) 秘密 10 年（secret10）

— c.4) 秘密永久（secret*）

d) 机密（classified=classified5）

— d.1) 机密 3 年（classified3）

— d.2) 机密 5 年（classified5）

— d.3) 机密 10 年（classified10）

— d.4) 机密永久（classified*）

e) 绝密（topsecret=topsecret10）

— e.1) 绝密 3 年（topsecret3）

— e.2) 绝密 5 年（topsecret5）

— e.3) 绝密 10 年（topsecret10）

— e.4) 绝密永久（topsecret*）

printtype ▷ 打印属性 (printtype)，可选值：

a) 单面打印（oneside）[缺省值]

b) 双面打印（twoside）

ctexbookoptions ▷ ctexbook 文档类支持的其他选项：

使用 ctexbookoptions 选项传递 ctexbook 文档类支持的其他选项。例如，使用 fontset=founder 选项启用方正字体以避免生僻字乱码的问题[1]。

模板已内嵌 LaTeX 工具包：ifthen，etoolbox，titletoc，remreset，remreset, geometry，fancyhdr，setspace，caption，float，graphicx，subfigure，epstopdf，booktabs，longtable，multirow，array, enumitem，algorithm2e，amsmath，amsthm，listings，pifont，color，soul, newtxtext, newtxmath。

[1]需要系统安装方正字体。

模板已内嵌宏：\highlight{text}（黄色高亮）。

请统一使用 UTF-8 编码。

2.2 选项设置

\refcolor ▷ 开启/关闭引用编号颜色，包括参考文献，公式，图，表，算法等

on：开启 [默认]

off：关闭

\beginright ▷ 摘要和正文从右侧开始

on：开启 [默认]

off：关闭

\emptypageword ▷ 空白页留字

\Listfigtab ▷ 是否使用图标清单目录

on：开启 [默认]

off：关闭

2.3 章节撰写

本模板支持以下标题级别标题级别：

\chapter{章}	▷ 第一章
\chapter*{无章号章}	▷ 无章号章
\chaptera{无章号有目录章}	▷ 无章号有目录章
\summary	▷ 总结
\appendix	▷ 附录
\achievement	▷ 攻读学位期间取得的成果
\acknowledgments	▷ 致谢
\biography	▷ 作者简介
\section{节}	▷ 1.1 节
\subsection{条}	▷ 1.1.1 条
\subsubsection{A}	▷ 1.1.1.1 A
\paragraph{a}	▷ 1.1.1.1.1 a
\subparagraph{a)}	▷ 1.1.1.1.1.1 a)

2.4　注意事项

- ▷ 中文斜体将转换为楷体；
- ▷ buaa.cls 采用包 newtxtext 和 newtxmath，**中文粗体**在 Windows 下转换为黑体（有可能是因为 newtx 包没安装好，By WeiQM），Linux 下正常（By QiaoJF）；
- ▷ buaa_mac.cls 采用包 times，**中文粗体**转换为黑体（By CaiBW）；
- ▷ \label{<text>} 中不能使用中文；
- ▷ 浮动体与正文之间的距离是弹性的；
- ▷ 命令符与汉字之间请注意加空格以避免 undefined 错误（pdfLaTeX 下好像一般不存在这个问题，主要在 XeLaTeX 编译环境下发生）；

2.5　ToDo

- ▷ 数学环境的行间隔；
- ▷ 参考文献的行间隔；

2.6 意见及问题反馈

E-mail：weiqm@buaa.edu.cn

GitHub：https://github.com/CheckBoxStudio/BUAAThesis/issues

第三章　示例

3.1　参考文献引用

3.1.1　数字标注

\cite{knuth86a} ⇒ [?]

\citet{knuth86a} ⇒ ?]

\citet[chap.~2]{knuth86a} ⇒ ? , chap. 2]

\citep{knuth86a} ⇒ [?]

\citep[chap.~2]{knuth86a} ⇒ [? , chap. 2]

\citep[see][]{knuth86a} ⇒ [see ?]

\citep[see][chap.~2]{knuth86a} ⇒ [see ? , chap. 2]

\citet*{knuth86a} ⇒ ?]

\citep*{knuth86a} ⇒ [?]

\citet{knuth86a,tlc2} ⇒ ? ?]

\citep{knuth86a,tlc2} ⇒ [? ?]

\cite{knuth86a,knuth84} ⇒ [? ?]

\upcite{knuth86a,knuth84} ⇒ [? ?]

\citet{knuth86a,knuth84} ⇒ ? ?]

\citep{knuth86a,knuth84} ⇒ [? ?]

\cite{knuth86a,knuth84,tlc2} ⇒ [? ? ?]

3.1.2　数字标注-上标形式

\upcite{knuth86a} ⇒ [?]

\upcite{knuth86a,knuth84,tlc2} ⇒ [? ? ?]

实现源码：\newcommand{\upcite}[1]{\textsuperscript{\cite{#1}}}。

3.1.3 著者-出版年制标

\cite{knuth86a} ⇒ **?**

\citet{knuth86a} ⇒ **?**

\citet[chap.~2]{knuth86a} ⇒ **?**, chap. 2

\citep{knuth86a} ⇒ (**?**)

\citep[chap.~2]{knuth86a} ⇒ (**?**, chap. 2)

\citep[see][]{knuth86a} ⇒ (see **?**)

\citep[see][chap.~2]{knuth86a} ⇒ (see **?**, chap. 2)

\citet*{knuth86a} ⇒ **?**

\citep*{knuth86a} ⇒ (**?**)

\citet{knuth86a,tlc2} ⇒ **??**

\citep{knuth86a,tlc2} ⇒ (**??**)

\cite{knuth86a,knuth84} ⇒ **??**

\citet{knuth86a,knuth84} ⇒ **??**

\citep{knuth86a,knuth84} ⇒ (**??**)

3.1.4 其他形式的标注

\citealt{tlc2} ⇒ **?**

\citealt*{tlc2} ⇒ **?**

\citealp{tlc2} ⇒ **?**

\citealp*{tlc2} ⇒ **?**

\citealp{tlc2,knuth86a} ⇒ **? ?**

\citealp[pg.~32]{tlc2} ⇒ **?** , pg. 32

\citenum{tlc2} ⇒ **?**

\citetext{priv.\ comm.} ⇒ [priv. comm.]

\citeauthor{tlc2} ⇒ **?**

\citeauthor*{tlc2} ⇒ **?**

\citeyear{tlc2} ⇒ **?**

\citeyearpar{tlc2} ⇒ [**?**]

3.2　浮动体

3.3　算法环境

模板中使用 algorithm2e 宏包实现算法环境。关于该宏包的具体用法请阅读宏包的官方文档。

————↓————Space Check————↓————

Data: this text
Result: how to write algorithm with LaTeX2e
initialization;
while *not at end of this document* **do**
    read current;
    if *understand* **then**
        go to next section;
        current section becomes this one;
    else
        go back to the beginning of current section;
    end
end

算法 1: A How to (plain).

————↑————Space Check————↑————

算法 2: A How to (ruled).

Data: this text
Result: how to write algorithm with LaTeX2e
initialization;
while *not at end of this document* **do**
    read current;
    if *understand* **then**
        go to next section;
        current section becomes this one;
    else
        go back to the beginning of current section;
    end
end

3.3.1　三线表

推荐使用三线表的方式，如表 3。

————┼————Space Check————┼————
————┼————Space Check————┼————

Data: this text
Result: how to write algorithm with LaTeX2e
initialization;
while *not at end of this document* **do**
  read current;
  if *understand* **then**
    go to next section;
    current section becomes this one;
  else
    go back to the beginning of current section;
  end
end

算法 3: A How to (boxed).

算法 4: A How to (boxruled).

Data: this text
Result: how to write algorithm with LaTeX2e
initialization;
while *not at end of this document* **do**
  read current;
  if *understand* **then**
    go to next section;
    current section becomes this one;
  else
    go back to the beginning of current section;
  end
end

表 **2** 表的标题

操作系统	**TeX** 发行版
所有	TeX Live
macOS	MacTeX
Windows	MikTeX

表 **3** 让我们看看一个长标题长什么样。还不够长？那我再多写一点。还是不够长？那我再多写一点点。**OK**，就是长这样的！

操作系统	**TeX** 发行版
所有	TeX Live
macOS	MacTeX
Windows	MikTeX

我们在这儿插入一行字；

我们在这儿再插入一行字；

我们在这儿插入一行字；

我们在这儿再插入一行字；

我们在这儿插入一行字；

我们在这儿再插入一行字；

我们在这儿插入一行字；

我们在这儿再插入一行字；

3.4　长表格

超过一页的表格要使用专门的 longtable 环境（表 4）。

————↓————Space Check————↓————

表 **4**　长表格演示

名称	说明	备注
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC

续下页

表 **4** 长表格演示（续）

名称	说明	备注
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC

————↑————Space Check————↑————

3.5 插图

————↓————Space Check————↓————

图 **1** 测试图片
第二行题注

————↑————Space Check————↑————

我们在这儿插入一行字；

我们在这儿再插入一行字；

我们在这儿插入一行字；

我们在这儿再插入一行字；

我们在这儿插入一行字；

我们在这儿再插入一行字；

我们在这儿插入一行字；

我们在这儿再插入一行字；

3.6　数学环境

3.6.1　数学符号

模板定义了一些正体（upright）的数学符号：

符号	命令
常数 e	\eu
复数单位 i	\iu
微分符号 d	\diff
arg max	\argmax
arg min	\argmin

更多的例子：

$$\mathrm{e}^{\mathrm{i}\pi} + 1 = 0 \tag{3.1}$$

$$\frac{\mathrm{d}^2 u}{\mathrm{d}t^2} = \int f(x)\,\mathrm{d}x \tag{3.2}$$

$$\arg\min_{x} f(x) \tag{3.3}$$

3.6.2　定理、引理和证明

定义 3.1. If the integral of function f is measurable and non-negative, we define its (extended) **Lebesgue integral** by

$$\int f = \sup_{g} \int g, \tag{3.4}$$

where the supremum is taken over all measurable functions g such that $0 \leq g \leq f$, and where g is bounded and supported on a set of finite measure.

例 3.1. Simple examples of functions on $\mathbf{R}^d$ that are integrable (or non-integrable) are given by

$$f_a(x) = \begin{cases} |x|^{-a} & \text{if } |x| \leq 1, \\ 0 & \text{if } x > 1. \end{cases} \tag{3.5}$$

$$F_a(x) = \frac{1}{1+|x|^a}, \qquad \text{all } x \in \mathbf{R}^d. \tag{3.6}$$

Then f_a is integrable exactly when $a < d$, while F_a is integrable exactly when $a > d$.

引理 3.1 (Fatou)**.** Suppose $\{f_n\}$ is a sequence of measurable functions with $f_n \geq 0$. If $\lim_{n\to\infty} f_n(x) = f(x)$ for a.e. x, then

$$\int f \leq \liminf_{n\to\infty} \int f_n. \tag{3.7}$$

注. We do not exclude the cases $\int f = \infty$, or $\liminf_{n\to\infty} f_n = \infty$.

推论 3.2. Suppose f is a non-negative measurable function, and $\{f_n\}$ a sequence of non-negative measurable functions with $f_n(x) \leq f(x)$ and $f_n(x) \to f(x)$ for almost every x. Then

$$\lim_{n\to\infty} \int f_n = \int f. \tag{3.8}$$

命题 3.3. Suppose f is integrable on $\mathbf{R}^d$. Then for every $\epsilon > 0$:

1. There exists a set of finite measure B (a ball, for example) such that

$$\int_{B^c} |f| < \epsilon. \tag{3.9}$$

2. There is a $\delta > 0$ such that

$$\int_E |f| < \epsilon \qquad \text{whenever } m(E) < \delta. \tag{3.10}$$

定理 3.4. Suppose $\{f_n\}$ is a sequence of measurable functions such that $f_n(x) \to f(x)$

a.e. x, as n tends to infinity. If $|f_n(x)| \leq g(x)$, where g is integrable, then

$$\int |f_n - f| \to 0 \qquad \text{as } n \to \infty, \tag{3.11}$$

and consequently

$$\int f_n \to \int f \qquad \text{as } n \to \infty. \tag{3.12}$$

证明. Trivial. □

3.6.3　自定义

Axiom of choice. Suppose E is a set and E_α is a collection of non-empty subsets of E. Then there is a function $\alpha \mapsto x_\alpha$ (a "choice function") such that

$$x_\alpha \in E_\alpha, \qquad \text{for all } \alpha. \tag{3.13}$$

Observation 3.1. Suppose a partially ordered set P has the property that every chain has an upper bound in P. Then the set P contains at least one maximal element.

A concise proof. Obvious. □

Observationvar2 3.2. Suppose a partially ordered set P has the property that every chain has an upper bound in P. Then the set P contains at least one maximal element.

A concise proof. Obvious. □

我们在这儿插入一行字；

我们在这儿再插入一行字；

我们在这儿插入一行字；

我们在这儿再插入一行字；

我们在这儿插入一行字；

我们在这儿再插入一行字；

我们在这儿插入一行字；

我们在这儿再插入一行字；

我们在这儿插入一行字；

我们在这儿再插入一行字；

我们在这儿插入一行字；

我们在这儿再插入一行字；

我们在这儿插入一行字；

我们在这儿再插入一行字；

我们在这儿插入一行字；

我们在这儿再插入一行字；

我们在这儿插入一行字；

我们在这儿再插入一行字；

结 论

学位论文的结论单独作为一章，但不加章号。如果不可能导出应有的结论，也可以没有结论而进行必要的讨论。

* 嗯，这就是你的论文了 *

附　录

下列内容可以作为附录：

1) 为了整篇论文材料的完整，但编入正文又有损于编排的条理和逻辑性，这一材料包括比正文更为详尽的信息、研究方法和技术更深入的叙述，建议可以阅读的参考文献题录，对了解正文内容有用的补充信息等；
2) 由于篇幅过大或取材于复制品而不便于编入正文的材料；
3) 不便于编入正文的罕见的珍贵或需要特别保密的技术细节和详细方案（这中情况可单列成册）；
4) 对一般读者并非必要阅读，但对专业同行有参考价值的资料；
5) 某些重要的原始数据、过长的数学推导、计算程序、框图、结构图、注释、统计表、计算机打印输出文件等。

* 嗯，自由发挥吧 *

攻读硕士学位期间取得的学术成果

对于博士学位论文，本条目名称用“攻读博士学位期间取得的研究成果”，一般包括：

攻读博士学位期间取得的学术成果：攻读博士学位期间取得的学术成果：列出攻读博士期间发表（含录用）的与学位论文相关的学位论文、发表专利、著作、获奖项目等，书写格式与参考文献格式相同；

攻读博士期间参与的主要科研项目：列出攻读博士学位期间参与的与学位论文相关的主要科研项目，包括项目名称，项目来源，研制时间，本人承担的主要工作。

对于硕士学位论文，本条目名称用“攻读硕士学位期间取得的学术成果”，只列出攻读硕士学位期间发表（含录用）的与学位论文相关的学位论文、发表专利、著作、获奖项目等，书写格式与参考文献格式相同。

* 嗯，研究生不列科研项目 *

致　谢

致谢中主要感谢指导教师和在学术方面对论文的完成有直接贡献及重要帮助的团体和人士，以及感谢给予转载和引用权的资料、图片、文献、研究思想和设想的所有者。致谢中还可以感谢提供研究经费及实验装置的基金会或企业等单位和人士。致谢辞应谦虚诚恳，实事求是，切记浮夸与庸俗之词。

* 嗯，感谢完所有人之后，也请记得感谢一下自己 *

作者简介

博士学位论文应该提供作者简介，主要包括：姓名、性别、出生年月日、民族、出生的；简要学历、工作经历（职务）；以及攻读博士学位期间获得的其他奖项（除攻读学位期间取得的研究成果之外）。

* 嗯，“硕士学位论文无此项”，《手册》上是这么说的 *

This is BUAATHESIS, Happy TeXing! — from WeiQM.