计算机学院信息专业实习方案

指导教师 **崔禾磊** 学生人数 **24**

一、实习对象与时间

实习对象：

计算机学院计算机科学与技术专业本科生，2016 级。

实习时间：2019-7-1 至 2019-7-12，共 2 周时间。

二、实习目标

1. 经历和实践一个项目的完整开发过程；
2. 初步具备 1）加密算法（如 AES、SHA）使用和安全去重项目或 2）机器学习算法的使用和大数据分析预测项目的开发的能力；
3. 培养团队合作能力和协作精神。

三、实习方式与基本要求

本次实习设定四个项目题目，让学生自行选择项目类别和开发环境，最终设计并完成指定题目的功能。

本队一共有 24 人，按照 3-5 人一组的形式自由组合，共 5 组，设定项目组长，完成选题和开发工作。

首先，教师对题目要求进行详细讲解，说明本次开发所需完成的任务，相关知识点，开发工具和开发环境，实习总结报告的具体形式、考核方式。学生根据教师的要求进行题目的选择和分组，由学生选定项目组长，组长对组员进行任务的分配和进度的监督，以及向指导教师汇报过程中遇到的问题，教师和学生讨论问题解决方法，共同完成整个实习题目。

本次实习主要培养学生的团队开发能力和协作能力，锻炼动手实践能力，鼓励在项目中的创新。

考核以项目组为单位，采用学生演示、教师提问、学生回答的方式来进行。考核点主要包括数据集的合理性、加密算法使用的准确性、核心算法实现的正确性以及展示界面的美观和易用性。项目组成绩包括整体分数和成员分数。整体分数根据考核点确定，组员的分数根据个人分工有所调整。

四、实习拟订项目

项目一：加密数据安全去重客户端模块

1. 背景及基本功能：

在大数据时代，每时每刻都在不断地产生新数据。国际权威数据机构 IDC 在 2018 年 11 月发布了一份关于数据持续增长的报告，报告预测世界范围内的数据总和还在经历高速的扩张，将会从 2018 年的 33ZB（zettabytes，泽字节）增

长到 2025 年的 175ZB。而这其中往往会存在很多冗余数据，常见的包括文档、视频、软件安装包等，他们或是来自不同用户或是来自同一个用户的不同目录。已有研究指出，在传统的文件系统中来自不同用户的数据冗余能够达到 50%以上，而在文件备份这一应用场景中荣誉数据甚至能超过 90%以上。因此，通过使用数据去重（data deduplication）技术能够缓解数据增长对存储资源的浪费。

为了更好地应对数据的爆炸性增长，利用相对经济、易于管理的公共云平台来进行海量数据存储、高效数据分发、复杂数据分析等业务越来越成为一种主流趋势。IDC 在上述报告中还预测到 2025 年，全球 49%的数据将会存放在公共云平台之中。但是，当把用户、企业或是政府组织的数据外包给云平台的同时，他们在某种意义上相当于是放弃了对数据的完全控制。出于利益的驱使，云端存储的服务器无时不刻都在面对着各种可能存在的外部或内部攻击，随之而来的数据泄露、篡改、丢失等现实中已经屡次出现的灾难性后果，使得广大用户往往陷入对数据安全、隐私的顾虑和本地存储能力不足的两难境地。因此，通过使用数据加密（data encryption）技术以及数据审计（data auditing）技术能够为外包数据提供有效的安全和隐私保护。

因此，为了推动云存储服务在新形势、新需求下朝着更加高效、安全的方向发展，数据冗余删除和数据隐私保护这两项关键问题应当同时得到重视和解决。为了保护数据的隐私，用户可以使用传统的加密手段（如 AES）将外包的数据通过使用他们各自的秘钥进行加密。但是，常见的加密手段通常是随机的算法（probabilistic encryption），会使得相同的数据在加密后变成不同的密文，从而阻止了服务器端通过求（密文）文件的哈希来判断冗余（如图 1 所示）。即便是使用确定性的加密算法（deterministic encryption），来自不同用户的冗余数据也会因为秘钥不同而得到完全不同的密文。虽然采用全局一样的秘钥并使用确定性加密算法可以实现这一目的，但是让所有用户使用相同且唯一的秘钥会使得整个系统容易遭受单点攻击而变得十分脆弱。

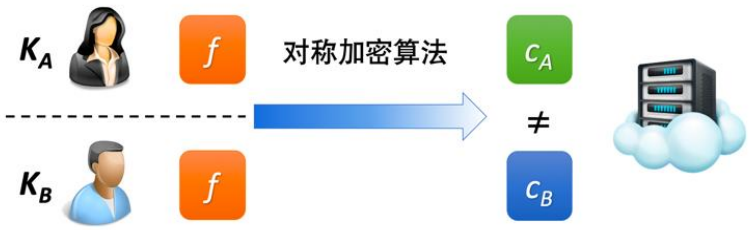

图 1 相同文件经不同用户加密后产生不同的密文

为了解决这一技术上的矛盾，安全数据去重（secure data deduplication）技术应运而生，其基本思想是通过文件本身生成加密数据用的秘钥，从而使得拥有相同数据的用户能够产生相同的秘钥，而无需事先统一秘钥。本项目将首先学习加密去重的经典算法，并在选定的具体加密去重算法基础上实现一个客户端工具，能够在本地给定文件目录下进行加密数据去重操作。

本模块具体任务：

- 相关密码学基础知识：熟悉本项目用到的密码学算法，包括 AES 和 SHA；
- 安全数据去重算法：学习 MLE 算法。
- 加密数据安全去重客户端模块：（要求有用户界面）
  a）实现文件选取、本地判定给定文件是否冗余；
  b）中间过程显示，包括相关文件 hash 的展示等；
  c）查重结果显示。
- 数据测试：使用冗余文件对算法进行完成度和正确性测试。

2. 开发语言和环境：

a）开发语言：Java、C++、Python 等；
b）开发环境：Eclipse、VS、Pycharm 等。

3. 开发进度：

a）7.1-7.2：分组、背景学习；
b）7.2-7.3：基本算法、工具学习及使用；
c）7.4-7.10：开发测试；
d）7.11-7.12：撰写总结报告。

4. 阶段成果和最终成果：

a）阶段成果：基本设计报告初稿；
b）最终成果：设计报告终稿，测试数据集及源代码，实习总结。

项目二：加密数据安全去重服务端模块

1. 背景及基本功能：

相关背景介绍同上，在此不再复述。

本项目将首先学习加密去重的经典算法，并在选定的具体加密去重算法基础上实现一个通过服务器端进行安全去重判定的协议实现，能够判定客户端请求的文件是否已经在服务器端存在。

本模块具体任务：

- 相关密码学基础知识：熟悉本项目用到的密码学算法，包括 AES 和 SHA；
- 安全数据去重算法：学习 MLE 算法。
- 加密数据安全去重服务端模块：（不要求有 UI，使用命令行展示亦可）
  a）实现客户端文件加密、去重标签的计算、发送请求到服务器端；
  b）实现服务器端解析去重请求，并进行去重检测；
  c）中间过程显示，包括请求过程、相关文件 hash 的展示等；
  d）查重结果显示。
- 数据测试：使用冗余文件对算法进行完成度和正确性测试。

2. 开发语言和环境：

a）开发语言：Java、C++、Python 等；
b）开发环境：Eclipse、VS、Pycharm 等。

3. **开发进度：**
   a）7.1-7.2：分组、背景学习；
   b）7.2-7.3：基本算法、工具学习及使用；
   c）7.4-7.10：开发测试；
   d）7.11-7.12：撰写总结报告。
4. **阶段成果和最终成果：**
   a）阶段成果：基本设计报告初稿；
   b）最终成果：设计报告终稿，测试数据集及源代码，实习总结。

项目三：基于当前职业的个人履历推测

1. **背景及基本功能：**

通过机器学习方法，利用大数据分类，进行个人信息预测是机器学习和大数据相结合的重要应用方向之一。其中，学历信息是个人信息中较为关键的部分。但是在实际应用中，因为调查困难、涉及到用户隐私等诸多因素，获得学历信息是个人信息获取任务中相对较难的部分。鉴于学历信息对于职业的选择以及职业生涯规划有着至关重要的作用，因此针对这一问题，本项目旨在利用目标用户当前已知的职业信息，利用经典的机器学习方法（如 KNN，SVM 等），对目标用户的学历信息进行较高准确率的分析、推测。

本模块具体任务：

- 个人履历信息处理：
  a）熟悉数据类型，准备实验环境；
  b）设计数据分类方案；
  c）数据分类；
  d）数据后处理，如：标注，提取等。
- 个人履历推测算法：如 SVM 等。
- 程序接口：提供统一规范的程序接口。包括参数设定，检测结果显示，算法性能评估等。
- 数据测试：对算法进行完成度和准确率的测试。

2. **开发语言和环境：**
   a）开发语言：Python、C / C++、Matlab 等；
   b）开发环境：Pycharm、Matlab2016、VS2015 等。
3. **开发进度：**
   a）7.1-7.2：分组、背景学习；
   b）7.2-7.3：基本算法、工具学习及使用；
   c）7.4-7.10：数据处理及测试；
   d）7.11-7.12：撰写总结报告。
4. **阶段成果和最终成果：**
   a）阶段成果：基本设计报告初稿；

b）最终成果：设计报告终稿，测试数据集及源代码，实习总结。

项目四：基于机器学习/自然语言处理的群智任务分析

1. 基本功能：

通过机器学习方法，结合大数据分类与预测，对群智感知/众包等相关平台上的任务进行任务分析。群智是低成本低能耗解决社会问题的重要方法。本项目旨在于利用爬取到群智任务数据，对任务进行细粒度的分解并提取感兴趣的特征，提高任务自动识别和分配的准确率。

本模块具体任务：

- 信息处理：

a）熟悉群智平台数据格式，准备实验环境

熟悉主流的爬虫框架（spider 等），根据提供的数据标注的模板在相应的网站上爬取海量数据。

b）数据预处理

将爬取到的数据进行预处理，预处理主要包括：数据清洗、自然语言处理、数据标注。

I、数据清洗

数据清洗，就是把不感兴趣的、视为噪音的内容清洗删除。在对群智平台 API 和 HTML5 格式的群智任务数据有了了解之后，对于爬取的网页内容，去除广告、标签、HTML、JS 等代码和注释等。

II、自然语言处理

①文本分词：将爬取到的一批短文本或者长文本，比如：句子，文章摘要，段落或者整篇文章组成的一个集合。一般句子、段落之间的字、词语是连续的，有一定含义。而进行文本挖掘分析时，我们希望文本处理的最小单位粒度是词或者词语，所以这个时候就需要分词来将文本全部进行分词。

②词性标注：就是给每个词或者词语打词类标签，如形容词、动词、名词等。这样做可以让文本在后面的处理中融入更多有用的语言信息。词性标注是一个经典的序列标注问题。常见的词性标注方法可以分为基于规则和基于统计的方法。其中基于统计的方法，如基于最大熵的词性标注、基于统计最大概率输出词性和基于 HMM 的词性标注。

③去停用词：停用词一般指对文本特征没有任何贡献作用的字词，比如标点符号、语气、人称等一些词。去停用词操作不是一成不变的，停用词词典是根据具体场景来决定的。

III、数据标注

数据标注种类繁多，如分类、拉框、注释、标记等等

①常见的几种数据标注类型：分类标注、标框标注、区域标注、描点标注等

②数据标注过程：标注标准的确定、标注形式的确定、标注工具的选择、标注产品的设计

- 分析：
  a）特征工程
  做完预处理之后，考虑如何把分词之后的字和词语表示成计算机能够计算的类型。主流的做法是数据向量化。有两种常用的表示模型分别是词袋模型和词向量。而这一部分的工作依赖于 Python 来实现。
  b）特征选择
  构造好特征向量之后，要选择合适的、表达能力强的特征。文本特征一般都是词语，具有语义信息，使用特征选择能够找出一个特征子集，其仍然可以保留语义信息；但通过特征提取找到的特征子空间，将会丢失部分语义信息。有一些现成的算法来进行特征的选择。目前，常见的特征选择方法主要有 DF、 MI、 IG、 CHI、WLLR、WFO 六种。
  c）模型训练
  在特征向量选择好之后，接下来要做训练模型，对于不同的应用需求，我们使用不同的模型，传统的有监督和无监督等机器学习模型， 如 KNN、SVM、Naive Bayes、决策树、GBDT、K-means 等模型；深度学习模型比如 CNN、RNN、LSTM、 Seq2Seq、FastText、TextCNN 等。
- 准确率预测
  a）评价指标
  训练好的模型，上线之前要对模型进行必要的评估，具体有以下这些指标可以参考：错误率、精度、准确率、精确度、召回率、F1 衡量。
  b）程序接口
  提供统一规范的程序接口，包括参数设定。检测结果显示，算法性能评估等。
  c）数据测试
  给出测试方案，对算法施行上述评价指标的具体测试，给出最终用户的学历信息推荐方案。

2. 开发语言和环境：

a）开发语言：Python、C / C++等；
b）开发环境：Pycharm、Chrome 浏览器、Spider 框架等。

3. 开发进度：

a）7.1-7.2：环境搭建、熟悉业务流程；
b）7.2-7.4：数据爬取与标注；
c）7.5-7.10：数据处理及分析测试；
d）7.11-7.12：撰写总结报告。

4. 阶段成果和最终成果：

a）阶段成果：基本设计报告初稿；
b）最终成果：设计报告终稿，测试数据集及源代码，实习总结。

五、技术准备与推荐参考资料

技术准备：

1. C++、Java、Python、JavaScript 等编程语言；
2. OpenSSL、Java Cryptography Architecture (JCA)等密码学工具；
3. SVM、Random Forest 等常用机器学习工具。

推荐资料：

1. 《INTRODUCTION TO MODERN CRYPTOGRAPHY》by Jonathan Katz and Yehuda Lindell
2. 《THE ELEMENTS OF STATISTICAL LEARNING》by Trevor Hastie, Robert Tibshirani, and Jerome Friedman
3. 《机器学习》周志华

六．实验条件及保证

本次生产实习依托的单位为西北工业大学计算机学院“智能感知与计算工信部重点实验室”以及“陕西省嵌入式系统技术重点实验室”，在教育部 985/211 学科建设项目支持下，已投资 700 万元，建立了普适计算与传感器网络研究平台，包括 IBM 服务器/工作站，用于搭建普适环境的膝上电脑（laptop）、个人数字助理（PDA）、智能手机（Smart Phone）、等离子体平板等，包括 100 余个感知节点组成的传感器网络实验环境，部分节点配备了 GPS 接收模块，可作为定位基准节点。网关设备选用 Startgate，该产品基于 400MHz 的 Intel Xscale 处理器（PXA255），配置多种接口，可通过 IEEE 802.11b 与通用计算机相连。自主设计开发了微型化传感器节点，并且可以与 Mote 兼容，带有红外探测等多种新型传感器件，同时基于 Xscale 270 开发了功能强大的簇头节点，具有 802.11b、GRPS 等多种通信接口。实验室部署 USB 摄像头、麦克风阵列、便携式话筒、RFID 等多种传感设备。同时具有完整的 OSGi 网关和软件平台。基于已有的软硬件平台，本实验室有能力承担本院本科生的生产实习工作。

本次生产实习是在於志文、郭斌教授的共同指导下，由课题组崔禾磊副教授等成员具体指导。

於志文，工学博士，西北工业大学教授，博士生导师，德国洪堡学者。现任计算机学院党委书记，智能感知与计算工业和信息化部重点实验室主任，陕西省嵌入式系统技术重点实验室主任，普适与智能计算研究所所长。分别于 2000 年 7 月、2003 年 3 月和 2005 年 12 月于西北工业大学计算机学院获学士、硕士和博士学位。2004 年 9 月至 2005 年 5 月在新加坡信息通讯研究院访问研究。2006 年 2 月至 2009 年 1 月先后任日本名古屋大学博士后研究员、京都大学特别研究员。2009 年 2 月以海外人才引进的方式，特聘为西北工业大学计算机学院教授。2009 年 11 月至 2010 年 10 月受德国洪堡基金会资助，赴德国曼海姆大学从事合作研究。2011 年 9 月至 2014 年 3 月任计算机学院副院长，2014 年 3 月至 2018 年 3 月任西北工业大学学科建设办公室主任。2009 年入选教育部新世纪优秀人

才支持计划，2012 年获得首批国家优秀青年科学基金，2015 年入选科技部中青年科技创新领军人才计划，2017 年获得国家杰出青年科学基金、入选国家"万人计划"科技创新领军人才计划。主要从事普适计算、移动互联网、人机交互、感知大数据等领域的研究工作。已在国际顶级学术期刊和会议上发表论文 150 余篇，被 SCI 收录 80 余篇。论文发表的刊物和会议包括《ACM Computing Surveys》、《IEEE Pervasive Computing》、《IEEE TKDE》、《IEEE THMS》、UbiComp、PerCom、IUI 等。论文被来自美国、英国、德国、日本、法国、加拿大等近三十个国家和地区的学者他引 5000 余次，引用刊物包括 SCI 期刊《IEEE Intelligent Systems》、《IEEE Transactions on Knowledge and Data Engineering》、《IEEE Communications Letters》、《IEEE Multimedia》、《User Modeling and User-Adapted Interaction》、《ACM Multimedia Systems Journal》等，引用会议包括 UbiComp、KDD、WWW、SIGIR、ICDE 等。

郭斌，工学博士，西北工业大学教授，博士生导师，计算机系统与微电子系主任，智能感知与计算工信部重点实验室副主任。分别于 2003 年和 2006 年获得西安交通大学本科和硕士学位。2006 年 9 月起在“日本政府（文部科学省）奖学金”资助下在日本庆应义塾大学开展博士研究，2009 年 3 月提前完成博士学业，之后赴法国国立电信学院开展两年博士后研究。2011 年以“海外人才引进”的方式，特评为西北工业大学副教授，2014 年破格晋升教授。入选“教育部新世纪优秀人才计划”（2012），陕西省青年科技新星（2014），国家“万人计划”青年拔尖人才（2016）。主要从事普适计算与人机交互，群体感知与群智计算，移动大数据挖掘、大数据智能等领域的研究。目前已在国内外重要期刊和国际会议如 ACM Computing Surveys, IEEE TMC, IEEE THMS, ACM TIST, IEEE TITS, ACM TKDD, IEEE Comuter, IEEE Pervasive Computing, UbiComp, INFOCOM, PerCom, ICDM 等上面发表论文 100 余篇，出版专著或专著章节七部，其中 60 余篇被 SCI 索引。

崔禾磊，哲学博士，西北工业大学副教授。分别于 2010 年 7 月获西北工业大学软件工程工学学士（BEng），2013 年 11 月获香港中文大学信息工程理学硕士（MSc），2018 年 10 月获香港城市大学计算机科学哲学博士（PhD）。2018 年 9 月至 2019 年 4 月在香港城市大学开展博士后（Research Fellow）研究工作。博士阶段主要研究方向包括云计算安全与隐私保护、云存储数据安全、多媒体与移动互联网安全等领域，以主要参与人身份参与多项由其导师主持的基金项目，包括国家自然科学基金面上项目、香港优配研究基金（HK GRF）、香港创新及科技基金（HK ITF）等，项目所涉及的研究主题包括隐私保护的图像服务外包、隐私保护的云社交平台中内容感知服务、基于网络缓存提速的加密数据内容分发与共享、以及面向安全可信的网络中间件外包的研究等。相关论文成果发表在国际顶级学术期刊 IEEE TMC、TSC、TIFS 和国际重要学术会议 IEEE INFOCOM、ICDCS、ICPADS（获大会唯一最佳论文奖）等，合计共发表长文 9 篇，其中第一作者身份 7 篇，第二作者身份 2 篇。