暨南大学硕士学位论文

题名：暨南大学硕士学位论文LaTex模板v0.0.1
论文英文题目

作者姓名：作者姓名（若是同等学力人员请注明“同等学力申请”，若是港澳台侨及海外留学生请注明申请人生源地）

指导教师姓名及学位、职称：导师姓名 导师学位 导师职称

学科、专业名称：学科 专业

学位类型：(学术学位/专业学位)

论文提交时间：2015 年 6 月

论文答辩时间：2015 年 6 月

答辩委员会主席：委员会主席 职称

论文评阅人：评阅人名字 职称

学位授予单位日期：2015 年 6 月

独 创 性 声 明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 暨南大学 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：　　签字日期：　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解 暨南大学 有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权 暨南大学 可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本授权书）

学位论文作者签名：　　导师签名：

签字日期：　　年　　月　　日　　签字日期：　　年　　月　　日

学位论文作者毕业后去向：

工作单位：　　电话：

通讯地址：　　邮编：

摘 要

本模板由Yongtao Zhou制作，非官方模板，请选择性使用，最终解释权和所有权归原作者和数据存储与集群计算实验室所有。仅供个人学习、学术交流使用，可以随意修改，保留原作者信息即可∧_∧。

如果您在使用中有建议和发现任何的bug，可以随时联系我修改

$$y.t.zhou@foxmial.com$$

本项目地址：https://github.com/icrt/JNUMasterThesis

关键词：重复数据删除；相似文件；局部性；指纹预取；备份系统

ABSTRACT

With the advent of huge amounts of data and the age of big data, digital TV、digital camera、web、 video、 large shopping site, etc, dramatically accelerated the growth of the data. In recent years, the global total data is growing at a speed more than 60% every year, the data need to backup and archive has brought great pressure on backup system. Deduplication technology has effectively reduced the storage cost by removing redundant data and reducing the actual use of storage capacity.

Key Words: Key Word 1; Key Word 2; Key Word 3

目　录

摘　要 …… I
ABSTRACT …… II
第一章 绪论 …… 1
1.1 课题背景 …… 1
1.2 国内外研究现状 …… 1
1.3 简单的数学公式和定理 …… 2
1.4 参考文献 …… 2
1.5 图片 …… 2
1.6 表格 …… 2
第二章 重复数据删除相关技术综述 …… 3
2.1 伪代码实例 …… 3
2.2 更新日志 …… 3
结　论 …… 5
在学期间发表论文的清单 …… 6
参考文献 …… 7
致　谢 …… 8

第一章 绪论

1.1 课题背景

我们正处在海量数据和大数据的时代，数字电视、摄像机以及其他通讯技术的出现正快速的加剧着数据的增长。据IDC（国际数据中心International Data Center）统计，2007年数字内容总量第一次超过了全球存储总容量，并且每年数据总量以指数的速度不断增长[1]。数据的爆炸性的增长给大型企业的数据中心带来了较大的压力，以不断扩大甚至扩建数据中心的方式并不能有效的缓解需要存储的数据的增长速度。同时，随着科技的快速发展和信息化的全面普及，数据对于企业甚至国家越来越重要，对于银行和互联网等公司，数据是它们赖以生存的根本，决定着未来的命运。

但是，由于种种未知的原因，人们无法预知或者避免数据的丢失和损坏。例如恐怖事件、系统故障、人为操作、自然灾害、黑客攻击、计算机病毒等各种因素，时刻威胁着大量对企业和国家至关重要的数据。在1993 年，美国世贸中心由于恐怖袭击发生爆炸。在爆炸前，大约有350 家企业在该大楼中办公，然而一年后，世贸大楼的公司只剩下了150 家，其余的200 家公司由于无法获取原有的重要数据而被迫倒闭。根据Gartner Group 的数据表明，企业数据灾难导致很多公司停止运营，其中2/5的公司无法再重新恢复，剩下的也有1/3在两年内相继宣告破产。

重复数据删除技术是一种新型的高级的数据压缩方式，它通过识别出重复的数据部分，删除冗余的部分，是一种更有效的节省磁盘空间的方法。研究发现，在数据备份系统中所存储的数据中有高达60%的数据是冗余的，而且这一比例会逐渐增加。

因此，基于重复数据删除技术的备份系统，对于节省磁盘空间、减少网络带宽、缩短恢复窗口期等有很重要的理论意义和实际意义。

1.2 国内外研究现状

重复数据删除技术作为一种有效的数据压缩方法被存储界列为十大存储热门技术之一，最早是由美国的 Data Domain 公司提出来的。在数据备份和归档领域得到了广泛的运用，由于基于重复数据删除技术的备份系统能够取得较好的压缩率以及

图 1.1: 实验室logo

表 1.1: 一个表的实例

r1	r2	r3	r4	r5
$4KB$	128929s	106511s	128201s	128201s

性能，以及相应的带来的节省带宽、降低成本等效益，重复数据删除技术越来越受到学术界和工业界的关注，成为存储界的一大新兴技术热点，并在高校和企业广泛研究和应用，并且取得了一系列的成果。

以上排版与《重复数据删除技术的实现与优化》一致

1.3 简单的数学公式和定理

定理 1.1 (存在性定理) $\Gamma\Theta\Lambda\Xi\Pi\alpha\beta\gamma\delta$

定理 1.2 xxxxx

$$\alpha = \sqrt[n]{\Re}. \tag{1.1}$$

定理1.2，公式1.1

1.4 参考文献

参考文献[1][2]

1.5 图片

图1.1是数据存储与集群计算实验室的logo。

1.6 表格

表1.1是一个实例。

第二章 重复数据删除相关技术综述

2.1 伪代码实例

算法 1: IntervalRestriction

输入: $G = (X, U)$ such that G^{tc} is an order.

输出: $G. = (X, V)$ with $V \subseteq U$ such that $G.^{tc}$ is an interval order.

```
1  begin
2      V ⟵ U
3      S ⟵ ∅
5      while S ≠ ∅ do
REM        remove x from the list of T of maximal index
8          while |S ∩ ImSucc(x)| ≠ |S| do
9              for y ∈ S − ImSucc(x) do
10                 { remove from V all the arcs zy : }
11                 for z ∈ ImPred(y) ∩ Min do
12                     remove the arc zy from V
13                     NbSuccInS(z) ⟵ NbSuccInS(z) − 1
14                     move z in T to the list preceding its present list
15                     {i.e. If z ∈ T[k], move z from T[k] to T[k − 1]}
16                 NbPredInMin(y) ⟵ 0
17                 NbPredNotInMin(y) ⟵ 0
18                 S ⟵ S − {y}
19                 AppendToMin(y)
20         RemoveFromMin(x)
```

2.2 更新日志

- 2015年4月9日v0.0.1

- 2015年4月15日更新第一页内容，增加学位类型。参考链接：
  http://gs.jnu.edu.cn/detail.html?1000109/W_13542_148575

结　论

结论，

在学期间发表论文的清单

1. paper1
2. paper2

参 考 文 献

[1] Andrei Z Broder. On the resemblance and containment of documents. In *Compression and Complexity of Sequences 1997. Proceedings*, pages 21–29. IEEE, 1997.

[2] Andrei Z Broder, Steven C Glassman, Mark S Manasse, and Geoffrey Zweig. Syntactic clustering of the web. *Computer Networks and ISDN Systems*, 29(8):1157–1166, 1997.

致　　谢

首先，感谢邓玉辉教授给我们提供了宽松的科研环境，邓玉辉教授严谨的治学态度和随和的为人，教会了我们非常多的东西，在此再次感谢邓玉辉教授。同时也要感谢谢俊杰师兄教会我使用LaTeX 2$_\varepsilon$，另外感谢刘瑞锴、杨儒、刘冰星、甘顺仪等小伙伴的帮助，最后感谢实验室及其班里其他同学的帮助和支持。