山东大学

毕 业 论 文(设 计)

论文（设计）题目:

学术论文 PDF 全文信息挖掘

姓　　名　刘昭呈

学　　号　201100300169

学　　院　山东大学软件学院

专　　业　软件工程

年　　级　2011 级

指导教师　王慧

2015 年 5 月 8 日

山东大学毕业设计（论文）成绩评定表

学院：　　专业：　　年级：

<table>
<tr><td>学号</td><td></td><td>姓名</td><td></td><td>设计（论文）成绩</td><td></td></tr>
<tr><td colspan="2">设计（论文）题目</td><td colspan="4"></td></tr>
<tr><td rowspan="2">指导教师评语</td><td colspan="5"></td></tr>
<tr><td colspan="5">评定成绩：　　签名：　　年　月　日</td></tr>
<tr><td rowspan="2">评阅人评语</td><td colspan="5"></td></tr>
<tr><td colspan="5">评定成绩：　　签名：　　年　月　日</td></tr>
<tr><td rowspan="2">答辩小组评语</td><td colspan="5"></td></tr>
<tr><td colspan="5">答辩成绩：　　组长签名：　　年　月　日</td></tr>
</table>

注：设计（论文）成绩=指导教师评定成绩（30%）＋评阅人评定成绩（30%）＋答辩成绩（40%）

目录

学术论文 PDF 全文信息挖掘

中文摘要

随着互联网数据核爆炸式的膨胀爆发，对于海量数据的存储和分析，成了企业和工程师需要面对和解决的问题。海量数据中蕴含着事物发展的规律，我们需要挖掘数据中这种规律，让我们对于这件事物发展产生更加深刻地认识。

在这种背景下，一个又一个新的分布式处理数据的尝试和方法不断诞生。 目前来看，Hadoop 平台无疑是这其中最为吸引关注的一个。 Hadoop 基础分布式架构主要由 HDFS 分布式文件系统和 MapReduce 计算模型组成。HDFS 主要负责海量数据的存储，而 MapReduce 主要负责在海量数据上的计算。

本文首先介绍 HDFS 的设计原则和前提假设，接着介绍 MapReduce 计算模型下数据处理的流程，都会经历哪几步，每一步都会做什么事情。最后列举一个较为复杂的实际海量数据处理实例，pdf 文件从解析到匹配。这个应用都是笔者实际做过，且数据量都在亿级别，都是典型的适合借助 HDFS 存储，利用 MapReduce 计算模型来处理的应用实例。

关键字：hadoop;mapReduce;hdfs;海量数据

Abstract

With the expansion of the Internet data's explosive eruptions, the enterprises and engineers need to face and solve the storage and analysis of massive data. The massive data contained the laws of development of things. We need to min that laws which helps us to have a more profound understanding of the development of things.

In this context, many new distributed data processing methods and practice born. At present, the Hadoop platform is undoubtedly the one of the most attention. Hadoop based distributed architecture is mainly composed of HDFS distributed file system and MapReduce calculation model. HDFS is mainly responsible for the massive data storage, MapReduce is mainly responsible for the massive data computation.

This paper first introduces the design principles and assumptions of HDFS. Then introduce the MapReduce model data processing. We will talk about What steps will experience, every step will do what. Finally we will talk about a complicated actual data processing example. The massive PDF files from the parse to match. This is the application I actually did, and the amount of data in billion level, is typical for using HDFS storage, application to handle the calculation model using MapReduce.

Keywords: hadoop;mapReduce;hdfs;massive data

第 1 章 绪论

1.1 大数据背景

据第35次CNNIC报告［1］显示，截止到2014年12月底，中国的网页数量为1899亿个，我国网民规模达6.49亿，手机网民规模达5.57亿，中国网民的人均周上网时长达26.1小时。另外根据该次报告显示，互联网有效地降低了交易和沟通的成本，也营造了更加分享的网络环境。根据本次调查显示，大约有60.0%左右的网民对于分享信息到互联网上持积极的态度，其中非常愿意分享的占13.0%左右，比较愿意分享的占47.0%左右 。借助互联网，网民在资源和信息等方面分享，不仅降低了相关交易成本，也创造出来了新的价值。互联网给广大的网民提供了平等表达自己见解的“新公共领域”。大约有43.8%的网民喜欢在互联网上发表各种各样的评论，其中非常喜欢发表评论的占6.7%，比较喜欢发表评论的占37.1% 。互联网已经成为人们发表各种言论的重要场所。

所谓大数据是指一类数据集，这类数据集的特点是体量非常大，数据类别非常大，并且用传统数据库工具无法对该类数据集进行抓取、管理和处理。

我们可以从四个层面来看待大数据的特点：第一个层面，数据体量巨大。从TB级别，跃升到PB级别。第二个层面，数据的类型异常各种各样。前面文章中所提到的例如video、image、graphic information。第三个层面，价值密度低。现在如果以video作为例子，在昼夜不停的连续的监控获取的数据中可能只有1,2秒的数据是有用的。第四个层面，处理速度快。1秒定律。业界将其归纳为4个“V”——Volume，Variety，Value，Velocity。

毫无疑问，我们已经进入了互联的时代，在这个时代中，我们每分每秒都产生着各种各样的数据，由此，我们实际上进入了一个大数据［2］的时代。所以，对于海量数据的存储和分析，成了企业和工程师需要面对和解决的问题。海量数据中蕴含着事情发展的规律，我们需要挖掘数据中这种规律，让我们对于这件事物发展产生更加深刻地认识。

1.2 大数据挑战

对于海量数据的处理，我们面临三个重要挑战。

首先，从数量上来讲，对于海量数据的存储问题，海量数据从大小上来讲，一般通常可以到达 GB~PB 级的超大数据的规模，因此，这就要求支持扩展功能的海量数据存储系统。此外，我们需要简便的存储系统，为了实现这个要求，我们增加各种各样的模块或者数目更多的磁盘柜来达到目的。当前互联网中各种各样的非格式化数据每天在不断膨胀地增长。随着数据的不断膨胀增大，整个系统的开销在不断加大。

海量数据数量庞大，在存储的时候，需要占用大量的存储空间，而且，由于数据庞大，中间一旦哪个地方出现了错误，就会导致数据完整性背破坏。即如何在保证数据完整性的前提下，提高数据的吞吐率。

第二，由于海量数据的数据量和分布性的特点，如果像传统一样在单机运行程序来处理该数据，处理的时间将是不可接受的。也就是说单机的计算能力已经满足不了海量数据的处理要求了。

分布式并行处理技术面临了处理各种各样的海量数据的新的挑战，于是，在一系列研究工作中开始出现以 MapReduce 为代表的项目。MapReduce 是 2004 年由谷歌公司提出的一个用来进行并行处理和生成大数据集的模型。

1.3 分布式计算

在计算机科学学科中，分布式计算<Distributed computing>主要研究分布式系统（Distributed system）[3] 如何进行计算。分布式系统是指由数量庞大的计算机组成的计算机集群，通过计算机网络实现集群中计算机之间相互连接，通过传递和分发消息进行通信，以协调它们工作而形成的系统。将需要进行大量计算工作任务的海量数据分区成一个又一个小块，分发给多台计算机分别进行独立地计算，当每台计算机独立完成计算后，会上传自己的运算的结果，最后将集群内的各个计算机上传的结果统一合并后，即可得出海量数据的整体的处理结果。

1.4 本文的组织结构

本文共分为六个章节，具体内容结构安排如下：

第一章：绪论介绍。简单介绍海量数据的背景，海量数据带来的挑战以及应对海量数据产生的分布式计算；

第二章：介绍 HDFS 的设计原则和前提假设，阐明 HDFS 是为了什么样的数据处理应用而设计的。接着介绍 MapReduce 计算模型下数据处理的流程，都会经历哪几步，每一步都会做什么事情。进而提出什么样的应用适合用 MapReduce 计算模型来处理。

第三章：开始介绍应用实例，简述了实例的背景，实例所面临的关键问题，以及每个关键问题所对应的解决办法，最后还介绍了本次实例的结果。

第四章：通过对这次应用实例分析，总结了整体对于基于 mapReduce 应用的处理海量数据的心得，并且对于复杂的应用总结了设计心得。

第 2 章 HDFS 和 MapReduce 简介

2.1 HDFS 简介

Hadoop [4] Distributed File System(HDFS) [5] 是被设计用来运行在通用硬件上的分布式文件系统。HDFS 和很多的其他分布式文件系统很像。但是，HDFS 是高容错，高吞吐率，非常适合运行有非常大的数据集的应用程序。另外，HDFS 还适合部署在低成本的机器上。HDFS 实现了对文件系统中的数据以流的形式访问，这一特性非常重要。

2.4 MapReduce 程序的数据流程

一个典型的 mapReduce [6] 作业开始运行的时候，client 首先会向 JobTracker 请求一个 Job ID,来唯一标识这个作业。同时会将运行该作业所必须的各种各样的资源文件上传到 HDFS 上，包括 Map 和 Reduce 的各自的应用程序、各种各样的需要的配置文件和在客户端这边设置的各种各样的输入和输出的划分的信息。这些所有的各种各样的文件都会被存放在一个由 Job Tracker 为该作业特地创建的一个文件夹目录中，文件夹目录的名字是这项作业的 ID。每个应用程序文件默认会有多达 10 个副本用来备份，可以使用 mapred.submit.replication 这个属性来控制副本的个数。剩余的各种各样的配置的信息，比如输入的具体的划分信息向 Job Tracker 指明了到底应该为这项作业启动多少个 map 任务来处理数据等信息。

Job Tracker 在接收到该作业请求之后，会立马将该作业放入一个各种各样的作业的队列里面，同时开始等待作业调度器对该作业进行相关的调度，当任务调度器根据自己的调度算法的逻辑决定调度该作业的时候，会根据之前配置好的输入划分信息为每个划分的分片创建一个 map 任务，并将各个 map 任务分配给 Task Tracker 来执行。对于 map 和 reduce 任务，Task Tracker 会首先根据主机核的数量和内存的大小，同时参考整体集群的槽位占用情况，分配一定数量的

map 槽和 reduce 槽。map 任务是直接分配给那些含有该 map 处理的数据块的 Task Tracker 上面的，同时会将应用程序各种各样的包复制到该 Task Tracker 上来运行。另外，在分配 reduce 任务的时候并不会考虑数据本地化。

Task Tracker 每当过了一段时间就会给 Job Tracker 发送一个 heartbeat 信号，利用这个信号来告诉 Job Tracker 它的运行状态，比如当前各个 map 任务完成的进度等。

2.4.1 Map 和 Shuffle

首先读取数据，数据流入 Map 阶段，每个 map 处理一个分片，HDFS 的默认分片大小为 64M，当然也可以通过命令来设置分片的大小。Map 的输出结果会暂时放到一个环形缓冲区内，这个环形缓冲区的大小默认为 100M，当超出大小限制之后，map 会把缓冲区中的数据先溢出到一个溢出文件中。然后根据之前设置的 reduce 任务的数目将整个数据划分为几个具有相同数目的分区，最后，形成一个 reduce 任务只对应处理一个分区的数据。这样做可以避免出现有些 reduce 任务分配不均衡，比如，某些任务分到过多数据，而有些任务却仅仅被分配到寥寥的数据，甚至有些任务完全没有分到数据的情况。其实对整体数据进行分区的过程就是对每行数据进行 hash 的过程。

分区之后，在环形缓冲区中，会对数据进行预排序，每当缓冲区满了，会写到磁盘上的溢出文件上。所以，最后排序就相当于多路归并。

2.4.2 Reduce

每个 reduce 任务只对应处理一个分区的数据，数据经过应用程序的逻辑处理后，产生相应的输出。

第 3 章 学术论文 PDF 全文信息挖掘

3.1 背景

百度学术搜索，致力于方便用户获取论文全文信息。互联网上存在着数以千万级别的免费的论文的全文 PDF 文件。为了更加方便用户获取到这部分全文 PDF 文件，需要对已经收录的数亿 PDF 文件进行一一解析，解析出标题，作者，参考文献［7］等有用信息，然后找到该 PDF 文件和学术搜索收录的题录信息的对应关系，以实现检索结果中提示用户，该篇题录信息在互联网上存在免费论文全文数据，从而更加方便用户获取全文数据。

下面的图 3-1 表示的是一个实际的检索的过程，其中检索结果中有些论文会有“免费下载”，这个就是题录信息和已经收录的 PDF 文件对应上之后，在整个产品上的结果。

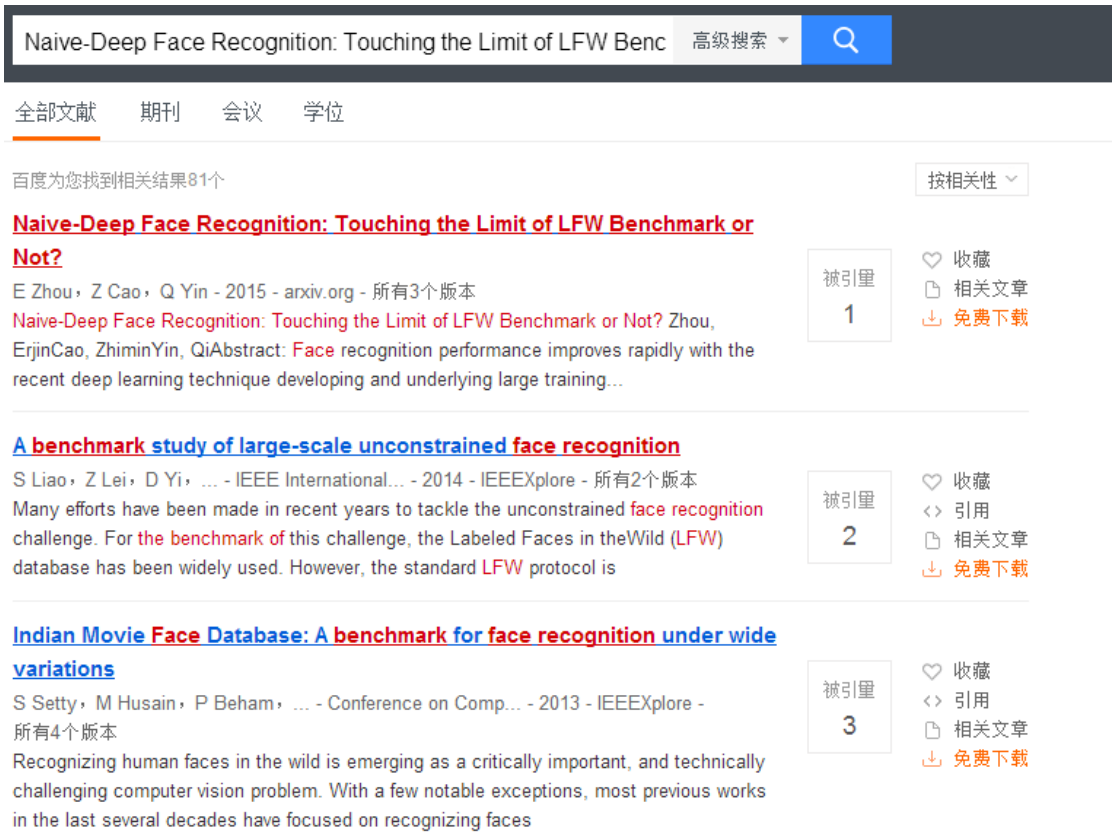
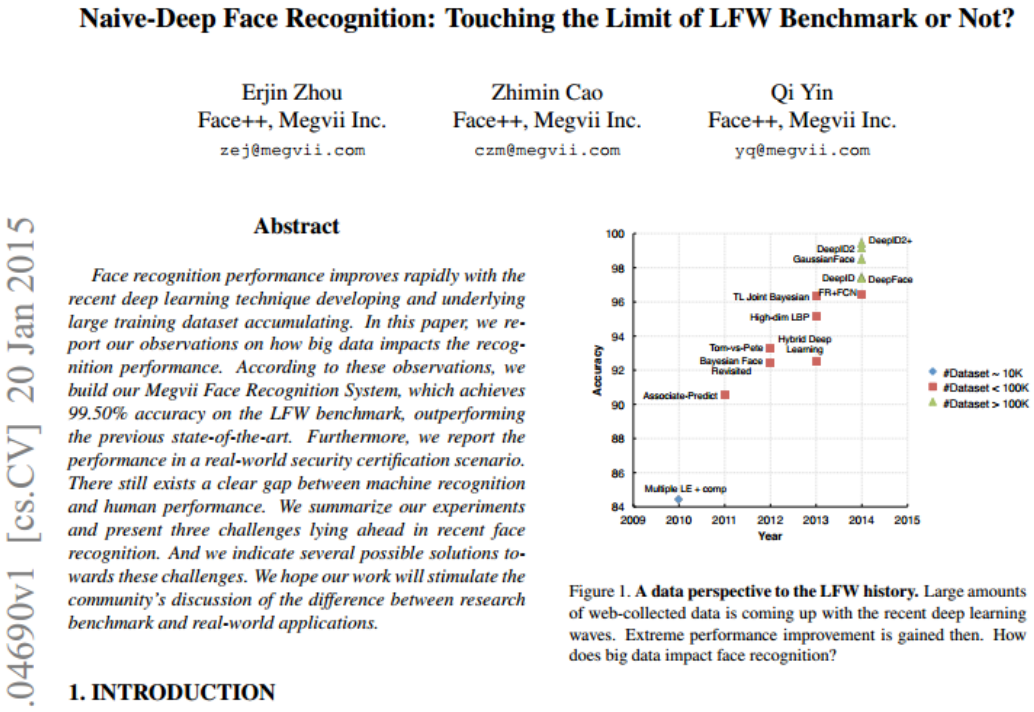


图 3-1 全文信息挖掘成功在产品上的体现

3.2 关键问题

3.2.1 PDF 解析困难

首先，PDF 格式不是结构化数据，无法直接获取到标题，作者等信息。学术论文的 PDF 文件，排版差异很大，不同的 PDF 文件中，标题和作者等信息的字体，字号，出现的位置可能会有很大的差异。下面的图 3-2，图 3-3 和图 3-4 显示了这一差异。



Naive-Deep Face Recognition: Touching the Limit of LFW Benchmark or Not?

Erjin Zhou
Face++, Megvii Inc.
zej@megvii.com

Zhimin Cao
Face++, Megvii Inc.
czm@megvii.com

Qi Yin
Face++, Megvii Inc.
yq@megvii.com

Abstract

Face recognition performance improves rapidly with the recent deep learning technique developing and underlying large training dataset accumulating. In this paper, we report our observations on how big data impacts the recognition performance. According to these observations, we build our Megvii Face Recognition System, which achieves 99.50% accuracy on the LFW benchmark, outperforming the previous state-of-the-art. Furthermore, we report the performance in a real-world security certification scenario. There still exists a clear gap between machine recognition and human performance. We summarize our experiments and present three challenges lying ahead in recent face recognition. And we indicate several possible solutions towards these challenges. We hope our work will stimulate the community's discussion of the difference between research benchmark and real-world applications.

1. INTRODUCTION

Figure 1. **A data perspective to the LFW history.** Large amounts of web-collected data is coming up with the recent deep learning waves. Extreme performance improvement is gained then. How does big data impact face recognition?

图 3-2 论文 PDF 样例 1

人脸识别技术综述

张翠平　　苏光大

(清华大学电子工程系 "智能技术与系统"国家重点实验室图形图象分室，北京　100084)

摘 要　首先对计算机人脸自动识别技术的研究背景及发展历程做了简单回顾，然后对人脸正面像的识别方法，按照识别特征的不同进行了分类综述，主要介绍了特征脸(Eigenface)方法、基于小波特征的弹性匹配(Elastic Matching)的方法、形状和灰度模型分离的可变形模型(Flexible Model)以及传统的部件建模等分析方法，通过对各种识别方法的分析与比较，总结了影响人脸识别技术实用化的几个因素，并提出了研究和开发成功的人脸识别技术所需要考虑的几个重要方面，进而展望了人脸识别技术今后的发展方向。

关键词　人脸识别　特征脸　小波特征　形　　状无关模型

中图法分类号：TP391.41　文献标识码　：A　文章编号　：1006-8961(2000)11-0885-10

Human Face Recognition : A

图 3-3 论文 PDF 样例 2

ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, No.3, March 2006, pp.525−534 http://www.jos.org.cn
DOI: 10.1360/jos170525

E-mail: jos@iscas.ac.cn
Tel/Fax: +86-10-62562563

基于 3D 人脸重建的光照、姿态不变人脸识别*

柴秀娟[1+]，山世光[2]，卿来云[2]，陈熙霖[2]，高　文[1,2]

[1](哈尔滨工业大学 计算机学院,黑龙江 哈尔滨 150001)

[2](中国科学院 计算技术研究所 ICT-ISVISION 面像识别联合实验室,北京 100080)

Pose and Illumination Invariant Face Recognition Based on 3D Face Reconstruction

CHAI Xiu-Juan[1+], SHAN Shi-Guang[2], QING Lai-Yun[2], CHEN Xi-Lin[2], GAO Wen[1,2]

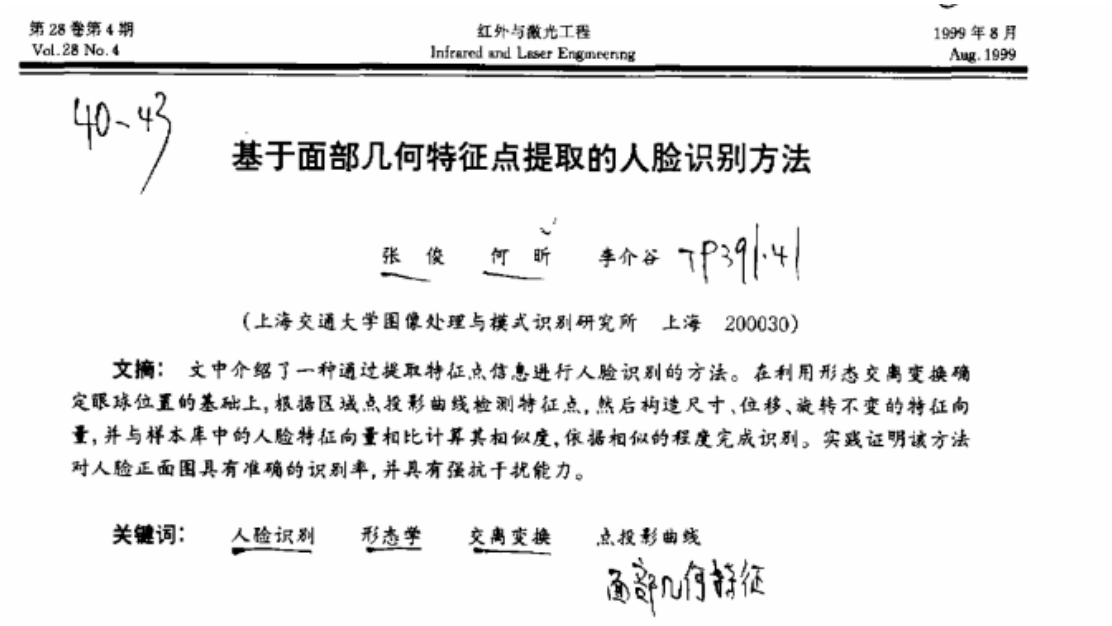
[1](Department of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

[2](ICT-ISVISION Joint R&D Laboratory for Face Recognition, Institute of Computer Technology, The Chinese Academy of Sciences, Beijing 100080, China)

+ Corresponding author: Phn: +86-10-58858300 ext 314, Fax: +86-10-58858301, E-mail: xjchai@jdl.ac.cn, http://www.jdl.ac.cn/

图 3-4 论文 PDF 样例 3

另外，有些 PDF 是扫描件，本身的质量就比较差，会导致解析出来的结果杂质比较多。图 3-5 反应了一个扫描件的 PDF，这个 PDF 解析出来可能会是标题多出 40-43



第 28 卷第 4 期
Vol.28 No.4
红外与激光工程
Infrared and Laser Engineering
1999 年 8 月
Aug. 1999

40-43

基于面部几何特征点提取的人脸识别方法

张 俊 何 昕 李介谷 TP391.41

(上海交通大学图像处理与模式识别研究所 上海 200030)

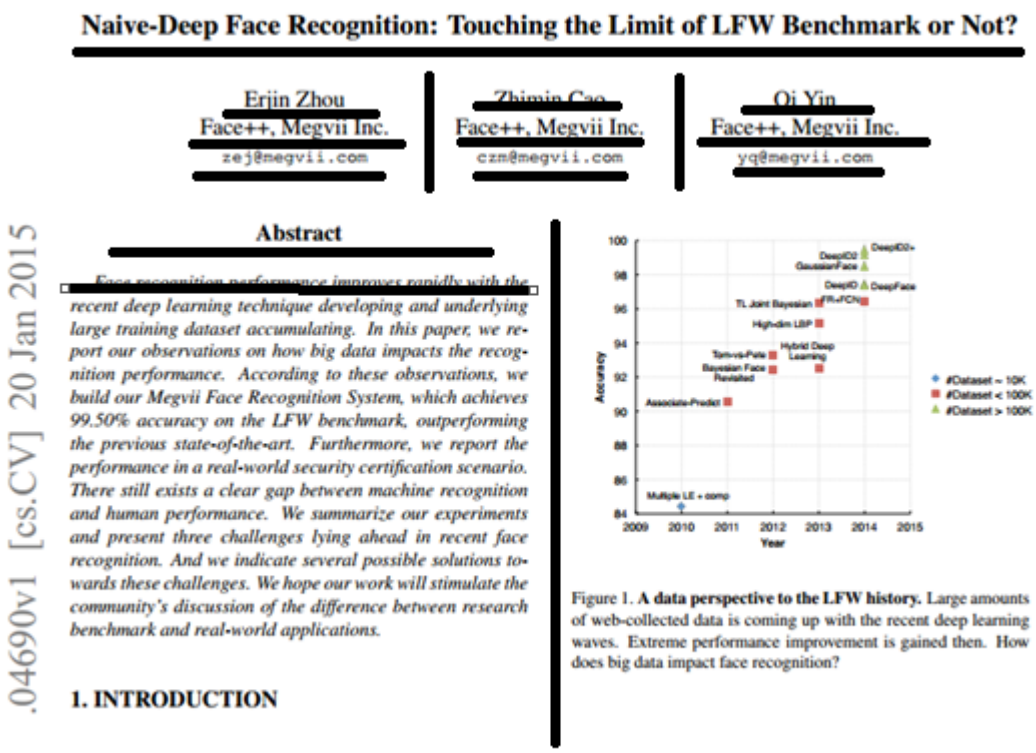
文摘： 文中介绍了一种通过提取特征点信息进行人脸识别的方法。在利用形态交离变换确定眼球位置的基础上，根据区域点投影曲线检测特征点，然后构造尺寸、位移、旋转不变的特征向量，并与样本库中的人脸特征向量相比计算其相似度，依据相似的程度完成识别。实践证明该方法对人脸正面图具有准确的识别率，并具有强抗干扰能力。

关键词： 人脸识别 形态学 交离变换 点投影曲线
面部几何特征

图 3-5 PDF 扫描件样例

PDF 格式偏重于渲染展示，支持自定义字体等，导致 PDF 文件解析存在不小的困难。在这个实例中，我采取了 XPDF [8] 来解析。在 XPDF 中，整个 PDF 文件好比在一张画布上画画一般，并不能直接提取出标题，作者，参考文献等信息。对于分行也只能使用 XPDF 内部分行算法，比较粗糙。

对应上面的图 3-2、图 3-3 和图 3-4，使用 XPDF 分行的结果可能如下图：



Naive-Deep Face Recognition: Touching the Limit of LFW Benchmark or Not?

Erjin Zhou
Face++, Megvii Inc.
zej@megvii.com

Zhimin Cao
Face++, Megvii Inc.
czm@megvii.com

Qi Yin
Face++, Megvii Inc.
yq@megvii.com

Abstract

Face recognition performance improves rapidly with the recent deep learning technique developing and underlying large training dataset accumulating. In this paper, we report our observations on how big data impacts the recognition performance. According to these observations, we build our Megvii Face Recognition System, which achieves 99.50% accuracy on the LFW benchmark, outperforming the previous state-of-the-art. Furthermore, we report the performance in a real-world security certification scenario. There still exists a clear gap between machine recognition and human performance. We summarize our experiments and present three challenges lying ahead in recent face recognition. And we indicate several possible solutions towards these challenges. We hope our work will stimulate the community's discussion of the difference between research benchmark and real-world applications.

1. INTRODUCTION

Figure 1. **A data perspective to the LFW history.** Large amounts of web-collected data is coming up with the recent deep learning waves. Extreme performance improvement is gained then. How does big data impact face recognition?

图 3-6 论文 PDF 样例 1 使用 XPDF 分行结果

人脸识别技术综述

张翠平　苏光大

(清华大学电子工程系"智能技术与系统"国家重点实验室图形图象分室，北京　100084)

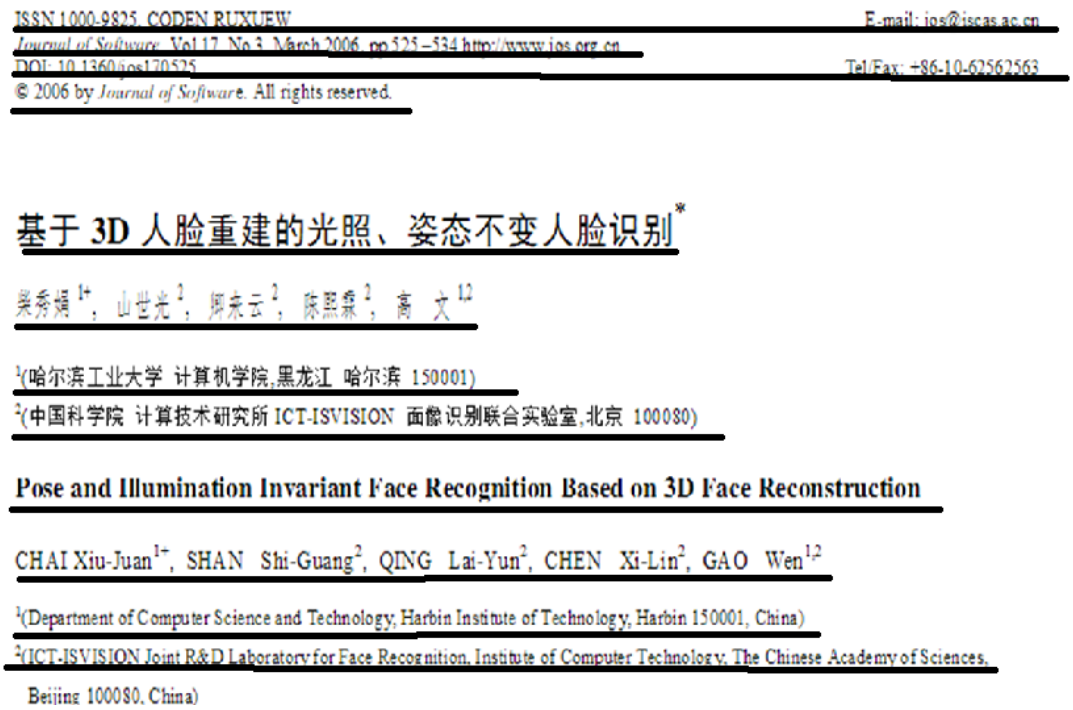
摘 要　首先对计算机人脸自动识别技术的研究背景及发展历程做了简单回顾，然后对人脸正面像的识别方法，按照识别特征的不同进行了分类综述，主要介绍了特征脸(Eigenface)方法、基于小波特征的弹性匹配(Elastic Matching)的方法、形状和灰度模型分离的可变形模型(Flexible Model)以及传统的部件建模等分析方法，通过对各种识别方法的分析与比较，总结了影响人脸识别技术实用化的几个因素，并提出了研究和开发成功的人脸识别技术所需要考虑的几个重要方面，进而展望了人脸识别技术今后的发展方向。

关键词　人脸识别　特征脸　小波特征　形　状无关模型

中图法分类号：TP391.41　文献标识码　：A　文章编号　：1006-8961(2000)11-0885-10

Human Face Recognition : A

图 3-7 论文 PDF 样例 2 使用 XPDF 分行结果


ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, No.3, March 2006, pp.525−534 http://www.jos.org.cn
DOI: 10.1360/jos170525
E-mail: jos@iscas.ac.cn
Tel/Fax: +86-10-62562563

基于 3D 人脸重建的光照、姿态不变人脸识别*

柴秀娟[1+], 山世光[2], 卿来云[2], 陈熙霖[2], 高 文[1,2]

[1](哈尔滨工业大学 计算机学院,黑龙江 哈尔滨 150001)

[2](中国科学院 计算技术研究所 ICT-ISVISION 面像识别联合实验室,北京 100080)

Pose and Illumination Invariant Face Recognition Based on 3D Face Reconstruction

CHAI Xiu-Juan[1+], SHAN Shi-Guang[2], QING Lai-Yun[2], CHEN Xi-Lin[2], GAO Wen[1,2]

[1](Department of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

[2](ICT-ISVISION Joint R&D Laboratory for Face Recognition, Institute of Computer Technology, The Chinese Academy of Sciences, Beijing 100080, China)

图 3-8 论文 PDF 样例 3 使用 XPDF 分行结果

如图 3-8，我们知道由于 PDF 文件内部并不会标识哪些是标题，哪些是作者，哪些是参考文献，而且，标题，作者和参考文献出现的位置会随着 PDF 文件排版的不同而不同，因此也不能简单地根据位置来提取。另外，还有一定比例的 PDF 文件本身为扫描件，扫描件<如图 3-5>上空白区域容易出现黑线等，非常容易导致分行出错。

3.2.2 目前无法解析的情况

一定比例的全图的 PDF，即为图片，这类 PDF 称为全图问题。xpdf 不支持解析这类 PDF 文件，无法分行，这类 PDF 需要使用 OCR 等方法来解析，目前在开发中。

另外，由于 PDF 文件本身为了支持自定义字体，所以，内部会有一套自己的字符编码，成为 CID，这个 CID 并不一定就是 unicode 码 [9]，因为比如想创造一个 unicode 码尚未编码的字符。一般 PDF 内部都会有一个 CID 到 unicode 码的对应表，但是，有一定比例的 PDF 文件会缺失该映射表，从而导致解析出来的结果乱码。这类问题称为 cmap 问题。

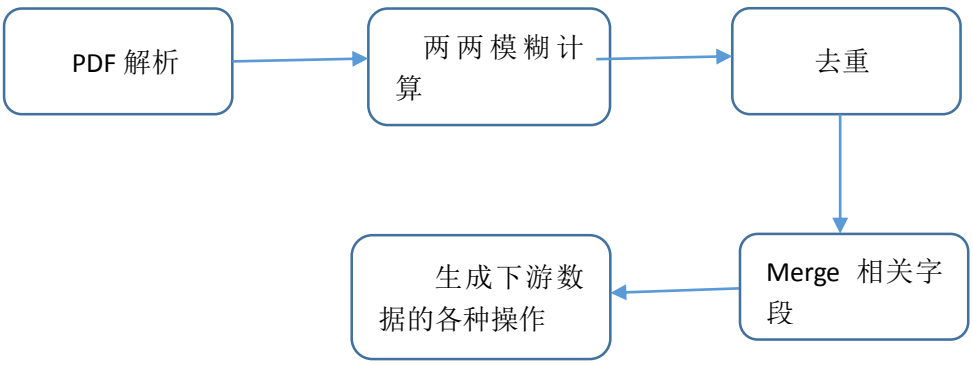
3.2.3 海量数据模糊匹配

一般认为，如果题录信息中的标题和作者和 PDF 文件中的标题和作者一致，则认为该题录与 PDF 文件相对应［10］。但是由于 PDF 解析结果准确率有限，需要对标题允许一定的模糊度［11］的情况下，比对作者是否一致。但是由于数据量庞大，题录信息和 PDF 解析结果都在亿级别，导致直接采取两两计算，就算使用庞大的集群也难以在可接受的时间内完成计算任务。

而且，传统的字符串编辑距离的算法的时间复杂度［12］为：O(n*m)，其中 n 和 m 分别为两个字符串的长度，这样的话，如果直接采取用动态规划［13］实现的编辑距离［14］来度量模糊度的直接两两计算，计算量为亿*亿*O(m*n)。在这种情况下，如果采取 map 切分的方式，即比如将题录信息当做一个词典，然后切分后，分发给每个 map 建成内存词典，然后，每个 map 流式处理所有的 PDF 文件的解析结果。即将题录信息切分为 a1，a2，a3，….，第 i 个 map 中使得 ai 与 PDF 文件解析结果两两计算，从而实现两两计算。但是就算我们开 10 万个 map 来切分题录信息，我们的时间复杂度也只是从亿*亿*O(m*n)变成了亿*亿*O(m*n)/10 万，还是无法在可以接受的时间内完成计算。所以，针对这个问题，我们要充分考虑 mapReduce 程序的特性，设计一个新的方法，减少不必要的两两计算，但是又要保证需要两两计算的题录和 PDF 文件解析结果进行了两两计算。

3.3 基于 MapReduce 的处理策略

3.3.1 主体流程



3.3.2 解析

由于 PDF 数据量庞大，无法用单机或者几台机器完成逐一解析，所以解析这一步需要在 hadoop 集群上完成。由于这整体的解析过程中，我们不需要利用按照一定的 key 排好序的特性来实现。所以，这一步，我们并不需要 Reduce 任务，只用 Map 任务即可。只是为了利用集群的庞大数目的机器来帮助我们完成 PDF 文件的逐一解析任务。

所以，相当于每个 map 处理的 PDF 文件数为——整体的 PDF 文件的数目/map 数目。每个 map 的逻辑就是利用 XPDF 逐一渲染 PDF 文件，通过 XPDF 内部的分行算法，获取 PDF 文件的每一行。然后根据字体大小，和上一行，下一行的距离远近，特征词(比如，英文论文的作者附近容易出现 by)，人名词典等等特征，对每一行进行相应的打分，从而提取出根据这些特征来看，最像标题，作者或者参考文献的相应内容。这里如果划分 20000 个 map，则每个 map 仅仅需要处理几万个 PDF 文件，由于 XPDF 是由 C++实现的，效率方面没有问题。

3.3.3 两两计算新思路

正如前面关键问题中的描述，传统的依靠切分词典的方式实现的两两计算，并不能满足目前的需求。我们需要设计一种新的两两计算的方法，既保证需要进行两两计算的题录数据和 PDF 文件解析数据进行相应的计算，又想要砍掉冗余的不需要进行的计算。

通过分析第一步的 PDF 文件的解析结果发现了，使用相应特征来打分的解析方式最后解析出来的解析结果，可以用下面几条原则概括：

1、 部分 PDF 文件解析出来的标题会少一个字符。

2、 部分 PDF 文件解析出来的标题会少几个词。

3、 PDF 文件解析出来的作者字段会含有杂质。

根据上面的几条原则，我们可以充分利用 mapReduce 计算模型下，reduce 的输入是按照一个 key 来排好序的特征。由于是按照 key 排好序，所以，相同的 key 下的数据总是在一起。

针对上面所述的第 1 种情况，采取将归一后(即删除一些特殊符号，转换一些编码等操作)的结果，如果长度为偶数，则直接将标题对半切分，切出来的两

个部分都当做这一条的 key 发往 reduce。如果长度为奇数，则将切 0 ~ int(length/2) - 1、int(length/2) ~ length 和 0 ~ int(length/2) + 1 、int(length/2) + 1 ~ length 分别当做这一条的 key 发往 reduce。对于题录数据，也要做相同的处理。

类似上面，针对第 2 种情况，我们知道题录数据比 PDF 解析的结果药长。所以，我们在 map 中可以以词为粒度，对题录数据进行切分，比如切掉第一个词，将剩余的标题作为 key 发往 reduce 等。

需要注意的是，发往 reduce 的 key 不应太短，否则，会造成相同 key 下面数据条数太多，容易出现计算时间过长且容易爆内存。当然，我们也可以在 reduce 进行限制，比如，一个 key 下面的条目超出一定范围，就直接丢弃，不进行计算等，可以有效避免这种情况。

采取上面这种计算方式，可以使得所有符合我们分 key 策略的条目进行我们想要的实际两两比较，而且又不用全量实际两两计算，大大削减了计算量，可以在可接受的时间内跑完。我们的分 Key 的前提假设是通过观察对整体数据进行随机抽样，然后进行人工观察，进行总结出来的。这些分 Key 的前提假设直接影响的是最后的结果的召回率。所以，如果最后的召回率不能满足需求，我们可以对整体数据重新随机抽样，然后人工观察，各种 case 的规律，获取得到各类 case 的分布率，根据样本中的出现频率实际选择一些作为我们分 Key 的前提假设，从而逐步提高召回率。按照这个模式不断迭代，最终是可以达到召回率要求的，所以，这个分 Key 策略对于召回率是完全可以满足的。

那么，对于准确率，我们怎么来控制呢？上面描述的分 Key 策略只是能保证召回率。对于准确率的控制，就需要我们回答一个问题——到底什么样的论文信息和题录信息算一篇论文？首先，根据我们 PDF 解析的结果，标题会出现上面那几种情况，所以，在 reduce 里面，相同的 key 下的题录信息和 PDF 解析结果，我们需要比较他们的作者字段和年份字段等信息。关于比较的时候的细节策略，这就略过了，总是为了保证准确率，我们仅靠标题是不行，尤其是标题，我们还允许一定程度的模糊，所以要比较其他字段，包含作者，年份，甚至期刊等信息，可以进一步提高准确率。

3.3.4 实例来看分 Key 策略

这一小节，我们通过实际看一个例子来讲解一下上一小节的分 Key 策略。

一篇论文的标题是

Naive-Deep Face Recognition: Touching the Limit of LFW Benchmark or Not?

该论文的作者是

Erjin Zhou, Zhimin Cao, Qi Yin

该论文的发表时间是

2015

由于题录信息一般是准确的，所以针对这篇论文，我们拿到的题录信息也是这样的。

那么我们拿到的 PDF 解析结果可能的情况：

该篇论文的标题是

1Naive-Deep Face ecognition: Touching the Limit of LFW Benchmark or Not?2

<前后多出来的数字可能由于 PDF 文件本身质量存在问题，杂质导致。漏掉字母可能是 cmap 不全导致。>

该论文的作者是

Erjin Zhou

Face++, Megvii Inc.

zej@megvii.com

Zhimin Cao

Face++, Megvii Inc.

czm@megvii.com

Qi Yin

Face++, Megvii Inc.

yq@megvii.com

<观察 PDF 文件内的作者，确实是一个作者一个单位，下一个作者下一个单位…>

该论文的发表时间是

2015

针对这种情况我们是怎么分 Key 的呢？

对于题录信息：

第一步归一化处理，得到的结果是

naivedeepfacerecognitiontouchingtheimitoffwbenchmarkornot

总长度为 57，根据上面描述的分 Key 策略，将切分出下面几个 Key

K1：naivedeepfacerecognitiontouc

K2：naivedeepfacerecognitiontouch

K3：hingtheimitoffwbenchmarkornot

K4：ingtheimitoffwbenchmarkornot

对于 PDF 解析结果：

首先归一化，得到的结果是：

naivedeepfaceecognitiontouchingtheimitoffwbenchmarkornot

<我自己定义的归一化函数包含去掉前后数字等一系列细节策略>

长度为：

56

根据上面描述的切分出来的 key 有两个：

K1：naivedeepfaceecognitiontouch

K2：ingtheimitoffwbenchmarkornot

由此可见，题录信息的 K4 和 PDF 解析结果切分出来的 K2 相同，会把这两份信息分到一起。

接下来，我们在 reduce 就可以在 Key 相同的数据中，对作者，年份等信息进行比对，具体比对方法比较细节，这里就不提了。通过对作者和年份等信息比对，就可以在保证一定准确度的前提下， 保证召回率。

3.3.5 去重与 AWK One-Liners

经过上面那一轮 MapReduce，我们会发现结果中产生的对应关系非常容易出

现重复的情况。这是由于 Map 中对相同的一个条目切分了最多 10 次，导致这个条目可能会在 Reduce 中被计算多次，从而产生了重复的结果。由于 HDFS 的一次写入，多次读取的特点，我们需要再跑一轮 MapReduce 来进行去重。

关于去重这个 MapReduce 的经典应用场景，有一个最为简洁的写法。即在 map 中直接将数据 cat 到 reduce，shuffle 的时候根据我们要求的列进行排序，这样保证了我们认为相同的数据会被分到一个 reduce。然后 reduce.sh 的逻辑可以使用如下语句：

```
cat - | awk '!a[$0]++'
```

即将数据读入，然后通过 a 数组这个字典来判断数据是否是第一次读入，是的话，输出，不是的话，不输出。则达到去重的目的

3.3.6 长尾现象与 merge 操作

在庞大的 hadoop 集群上运行 mapReduce 任务处理数据的时候，经常出现，个别 Reduce 任务运行异常缓慢，比如一个 mapReduce 任务，一共分成了 1000 个 reduce 片来运行，如果出现了 990 多个任务很快就完成了，但是剩下几个任务运行异常缓慢，那么这种现象就称为“长尾现象”。

遇到长尾现象，首先要通过 job tracker 来查找到相应的 task tracker 的 debug 信息，通过分析这个 debug 信息来判断那些运行缓慢的长尾 task 是否是由于本身负责运行这个 task 的机器的故障问题。

如果排除掉相应的机器故障问题。那么我们应该查看相应任务的 attempt 的错误输出，来确认是否是程序运行到最后几步然后出错，由于 attempt 数还没到预先设置好的 attempt 上限，从而导致长尾现象。

如果排除了上面这个原因，我们就需要实际分析相应的代码逻辑，首先排除代码中存在死循环，导致长尾现象产生。接着需要检查是否，代码逻辑是否会导致对于特定的数据计算量太大，从而导致长尾现象。

最后，如果这些都没有问题，那么还有一种可能就是由于 hadoop 是按行来读取数据，如果一行的数据量太大，比如一行大小就在上百 M，那么也会导致长尾现象产生。

经过上面对长尾现象的原因的分析，前面几种问题导致的长尾现象都容易通

过人为干预，解决问题。但是针对最后一种可能导致长尾的原因，我们需要怎么办呢？

一种可能的解决办法是，如果这一行分成很多个不同的列，我们这一轮 mapReduce 其实只需要其中的某一列，那么我们就没必要把所有的列都带上发往 reduce，我们可以在 map 中取出我们所用到的列，发往 reduce 即可。这样可以大大减小这一行的大小，从而避免长尾现象出现。

关于 merge 操作，众所周知，mapReduce 计算模型中存在 KV 对，典型的 merge 操作就是根据 K 来把相应的一些另一份数据的字段 merge 到新数据中。Merge 操作的实现分成 map 和 reduce 来看。Map 里需要明确好哪一个字段来作为 Key，可以规定好一些标记字段，用来标记该行来自于哪份数据，然后把相应的数据发往 reduce。

经过 shuffle 根据 Key 排序之后，在 reduce 中，所有具有相同的 Key 的数据都会被发往同一个 reduce，而且会排在一起，那么我们只要实现一个简单的从上往下读取的算法就能实现 merge 操作。这里面只需注意好首行的初始化问题、相同 Key 的数据的存储问题和从一个 Key 切换到另一个 Key 的数据时候的输出问题。

3.3.7 hive 与统计操作

在生成数据之后，需要进行一系列的统计工作，比如，需要统计整体生成的关系对个数，找到原文的题录占比或者找到题录的 PDF 文件占比等等。每一项统计都可以在 mapReduce 计算模型下实现，但是，如果每一项都写个 mapReduce 程序，会造成不仅代码量大而冗余，后面例行维护起来也不方便。所以，这种情况最适合使用 hive [15][16] 来解决问题。通过指明分隔符，数据地址等，将数据建表，然后，就可以使用类似 sql 的 hql 语句来完成相关统计操作了。其实，hive 原理就是把写好的 hql 映射成一轮或者几轮 mapReduce 程序来执行的。

3.3.8 准召率评估结果

准召率其实是两个统计指标——准确率和召回率

准确率非常明显就是表示最后产出的数据的准确程度，即如果产出数据量较

小，则可以进行人工全部抽查。如果产出数据量较大，可以先对数据进行随机抽样，看抽样出来的数据中正确无误的数据的占比，即可得到准确率。

召回率表示的意思是，最后所有正确产出的结果数据中包含所有应该正确产出结果的占比。评估方法可以是多样的。比如，可以人工从所有应该召回的数据中抽样出子数据集，然后通过人工分析这个子数据集中的结果有多少是在最后产出的结果中存在，即可得到召回率。

通过对最后生成的数据进行多次随机抽样，人工将抽样数据进行实际打开 pdf 文件和题录信息进行对比发现，准确率达到 90%以上。另外，通过多次抽取没有匹配到 PDF 文件的题录信息，然后人工查找 PDF，发现召回率在 90%以上。所以，达到了预期的工程目标。

第 4 章 结论

4.1 设计和实现的总结

首先， 通过这次设计和实现，我对于 HDFS 的设计目标和前提假设有了更为深入的理解。充分理解设计的前提假设，有利于在面对问题的时候，更加清楚地知道这个问题是否能用 HDFS 来存储，或者说是否适合用 HDFS 来存储。

通过这次设计和实现，我深入理解了 MapReduce 这种编程模型，用于大规模数据集的并行运算。彻底搞清楚了两个重要概念"Map（映射）"和"Reduce（归约）"，和它们的主要思想，它的设计极大地方便了编程人员在不会分布式并行编程的情况下，将自己的程序运行在分布式系统上。作为用户客户端来说，只需通过编程指定一个 Map（映射）函数，用来把一组键值对映射成一组新的键值对，再通过编程指定并发的 Reduce（归约）函数，用来保证所有映射的键值对中的每一个共享相同的键组。熟练掌握基于 MapReduce 计算模型的编程技巧和解决问题的思路，是我们面对问题时应该具备的基础。有了这个基础，我们需要耐心分析面对的问题，首先确定该问题是不是符合 HDFS 的前提假设，如果不符合，则就不适合使用基于 MapReduce 的计算模型来解决。如果符合，则需要分析该问题是不是 mapReduce 的典型应用场景，比如，merge 字段，对数据进行去重等。如果不是，则就需要仔细分析问题，结合 shuffle 按 Key 排序的特性，包括所有的相同的 Key 的数据会发往同一个 Reduce，经过排序后，相同的 Key 的数据会连在一起等等特性，来思考问题的解决办法。

致　谢

在这里需要感谢我的校内指导老师——王慧老师

在这里需要感谢我在百度实习时候的导师——杰艺

参考文献

[1] 中国互联网信息中心; 第 35 次中国互联网络发展状况统计报告 2015 ;

[2] 邬贺铨. 大数据时代的机遇与挑战[J]. 信息安全与通信保密, 2013 (3): 9-10.

[3] Coulouris, George; Jean Dollimore; Tim Kindberg; Gordon Blair. Distributed Systems: Concepts and Design (5th Edition). Boston: Addison-Wesley. 2011.

[4] White T. Hadoop 权威指南[J]. 北京: 清华大学出版社, 2011, 201(1): 1-123.

[5] D Borthakur . The Hadoop Distributed File System:Architecture and Design. 2007

[6] J Dean,S Ghemawat. MapReduce_Simplified Data Processing on Large Clusters. 2004

[7] 王平. 参考文献引用原则的探讨[J]. 编辑学报, 2004, 16(1):35-36. DOI:10.3969/j.issn.1001- 4314.2004.01.016.

[8] Noonburg D. xpdf: A C++ library for accessing PDF[J]. 2009.

[9] Unicode Staff C. The Unicode standard: worldwide character encoding[M]. Addison-Wesley Longman Publishing Co., Inc., 1991.

[10] 余玄璇, 曾国荪, 丁春玲. 基于标题与正文匹配的科技论文可信质量评估方法[J]. 计算机应用, 2014,

[11] 陈鹤阳. 模糊匹配理论在学术论文检索中的应用[J]. 图书情报论坛, 2010, (2):9-11.

[12] 王晓东. 计算机算法分析与设计[J]. 2001.

[13] Bellman R E, Dreyfus S E. Applied dynamic programming[J]. 1962.

[14] Masek, Paterson W J ;, S. M. Faster Algorithm Computing String Edit Distances.[J]. Journal of Computer & System Sciences, 1980.

[15] Thusoo A, Sarma J S, Jain N, et al. Hive: a warehousing solution

over a map-reduce framework[J]. Proceedings of the VLDB Endowment, 2009, 2(2): 1626-1629.

[16] Capriolo E, Wampler D, Rutherglen J. Programming hive[M]. " O'Reilly Media, Inc.", 2012.

附录 1 英文原文

Defining Data Science

Beyond the study of the rules of the natural world as reflected by data

Yangyong Zhu and Yun Xiong School of Computer Science, Fudan University, Shanghai, China Shanghai Key Laboratory of Data Science, Fudan University, China. {yyzhu, yunx}@fudan.edu.cn

Data science has received widespread attention in academic and industrial circles. New data science research institutes and organizations have continued to emerge on the scene, such as the Columbia University Institute for Data Sciences and Engineering and New York University Center for Data Science. The University of California at Berkeley, Columbia University, Fudan University, and other universities have launched data science courses and degree programs. Cleveland and Smith proposed that data science should be considered an independent discipline2, 8. Facebook, Google, EMC, IBM, and other companies have established employment positions for data scientists. According to Harvard Business Review, the data scientist is “the sexiest job of the 21st century.” Currently, there are several viewpoints regarding the definition of data science (see page 2). However, there is no consensus definition. We believe that, as a new science, the research objectives of data science are different from those of other, more established branches of science. In addition, the scientific issues that data science addresses are not studied by natural or social sciences.

Our team has worked on data technology and research projects funded by Chinese government since 1998, and we have applied our work to life science, healthcare, finance, transportation, and other fields (Table 1). Over the years, we have noticed a number of common issues related to data in scientific research and industrial applications, most notably the similarity of data objects. We have come to realize that

there is a considerable need to conduct research specifically on the data itself, and we started to explore concept of data science in 20099. Since 2010, we have hosted the annual International Symposium on Data Science and Dataology (iwdds.fudan.edu.cn). The symposium provides us with forum for the discussion of data science issues with scientists involved in computer science, life sciences, astronomy, and other fields. Over the past 16 years, our understanding of data science has taken more solid shape. We believe

that data in cyberspace have formed what we call datanature9, 10. Data science is the scientific research of datanature.

There are several current viewpoints on data science.

VP1: Data science is the science of studying scientific data.

The Committee on Data for Science and Technology (CODATA) launched the Data Science Journal (codata.org/dsj/) in 2002. CODATA regards data science as the methods and technologies used to conduct scientific research through management and utilization of scientific data. As scientific data have become more accessible, data science has been used to better characterize the data-intensive nature of today's science and engineering. Many disciplines use data technology to deal with scientific data from their respective areas. From this, X-informatics emerged, including bioinformatics, neuroinformatics, and social informatics.

For example, researchers in NuMedii, Inc., a big-data company in Silicon Valley, predicted whether existing drugs could be used to treat ovarian cancer by examining gene expression data from over 2,500 ovarian tumor samples6.

As another example, mathematicians from Harvard University Aiden and Michel studied American history using Ngrams on Google1. They used Ngrams to search for the usage frequencies of two phrases: "United States are" and "United States is." The search results showed that before the American Civil War, the two phrases were used

at roughly equal frequency, but after the Civil War, the latter became far more common than the former. This is seen as indicative of the levels of acceptance by the public of the United States as a unified nation before and after the Civil War.

From this point of view, data mainly refer to data generated and used in scientific studies. This emphasizes that data science is the management, processing, and use of scientific data to support scientific research, i.e., the currently commonly known data-intensive scientific research or fourth paradigm of scientific research4.

VP2: Data science is the science of studying business data.

In 2010, Loukides discussed what data science is, arguing that data science should enable the creation of data products rather than working as a simple application with data5. In 2013, Provost et al. pointed out, “extracting knowledge from data to solve business problems” is one of the fundamental concepts of data science7.

Providing support for BI methodology research makes up a significant portion of the work

performed by many data scientists. To effect this, a large proportion of BI practitioners were transitioned into data scientists. Amazon, Google, LinkedIn, Facebook, and other internet companies opened job positions for data scientists and established data science teams. These data scientists study and analyze business data to provide services for management decision making. For example, Amazon uses collaborative filtering to generate high-quality product recommendations, and Facebook uses a “People you may know” feature to recommend friend connections.

From this point of view, the acquisition of knowledge from business data in order to make decisions is one aspect of data science. This is similar to what BI scientists work on. For this reason, many BI scientists are also called data scientists. However, compared to BI issues, data science focuses more on common issues in the analysis of various business data, i.e., the issues on BI methodology.

VP3: Data science is an integration of statistics, computing technology, and artificial intelligence (AI).

This viewpoint often comes up in discussions on what data scientists are. It is generally believed that data scientists should have skills in statistics, computing technology, AI, and related fields and that data scientists are not individual people specializing in one field so much as teams consisting of statisticians, computer scientists, AI experts, and domain experts.

For example, the data scientist teams at Google and Facebook are composed of statisticians, computer scientists, AI scientists, and experts in other relevant fields.

This viewpoint is simple: Because statistics, computing technology, and AI are all used to process and analyze data, they are all a natural part of data science. VP4: The purpose of data science is to solve scientific and business problems by extracting knowledge from data.

In 2013, Dhar discussed the implications of data science from a business and research standpoint3. He defined data science as “the study of the generalizable extraction of knowledge from data”3. He also pointed out that a data scientist needs to have comprehensive skills covering statistics, machine learning, AI, and database management and have a deep understanding of problem design. This viewpoint can be seen as an integration of the first three viewpoints.

What is data science?

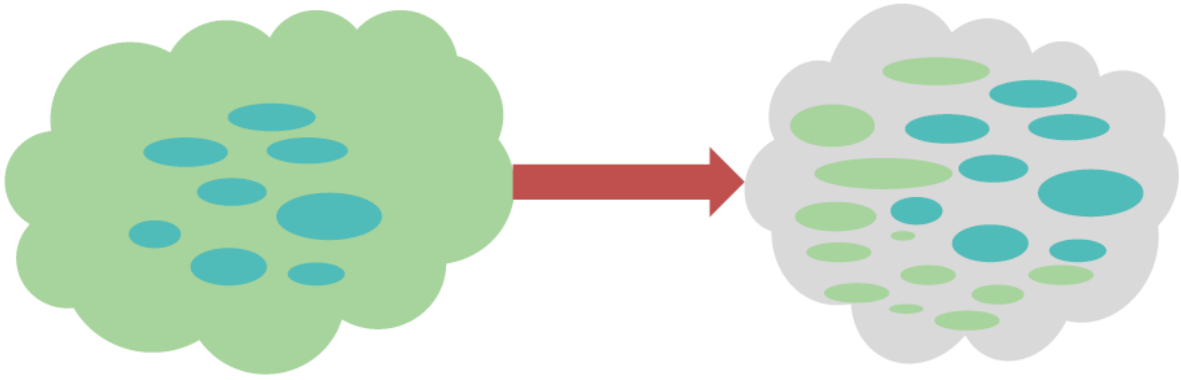
The basic ideas underlying the definitions given above are that data science is used to acquire

knowledge from data in some relevant fields and to provide support for existing scientific research and management decision-making schema. However, all work described above is still not enough to establish data science as a new, unique branch of science. This is because the objects of their study are things in the natural world,

and their research issues are also addressed in existing scientific fields.

With the development of digital equipment, things in the natural world are increasingly being stored in cyberspace in the form of data. Data are entered, generated, and created in cyberspace in a variety of ways and have become more and more diverse, complex, and out of human control. More and more data are unknown to or poorly understood by humans. Data in the cyberspace already show features of an independent world, like the natural world, so all data in cyberspace are here referred to as datanature.

It should be noted that there are two types of data in the cyberspace. The first is the data that represent things in the natural world, here called real data. An example is personal information, which is data representative of personal characteristics. The second is data that do not represent things in the natural world, here called virtual data. Virtual data means that the instances of such data have no references in the natural world. An example is computer viruses, which are neither viruses in the natural world nor data representation of real viruses; instead, they only exist in cyberspace (Figure 1).



(a) Nature (b) Datanature

Figure 1. Real data and virtual data. The blue ellipses and the green ellipses in Figure 1(b) denote the real data that represent things in the natural word. Among these, the blue ellipses correspond to those things in the natural world which have been stored in cyberspace in the form of data (i.e., the blue ellipses in Figure 1(a)); the green ellipses correspond to those things in the natural world which would be stored in cyberspace gradually (i.e., the green part in Figure 1 (a)). The grey part in Figure 1(b) denotes the virtual data that do not represent things in the natural world, i.e., the instances of such data have no references in the natural world.

The formation of datanature has produced new objects of study and new scientific issues. These new objects of study are not things that exist in the natural world or in human society but rather in datanature, i.e., data. There are new scientific issues about datanature. What size is datanature? What is the growth rate of global data? How do the data flow in cyberspace? How should the authenticity of datanature be determined? None of these issues are addressed by the natural or social sciences. These new scientific issues need to be studied by a new science.

Data science is the science of studying datanature and the science of data itself. On a basic level, it involves extracting knowledge from data. Because some of the data in datanature do represent real things, the knowledge acquired from these data can be used for natural and social science. This type of work is considered as data science according to VP1-VP4. However, it is only one part of data science research.

Data science is here defined as follows:

Data science is the theory, method, and technology of studying datanature. It has two main components.

The first component is the study of the patterns and rules of data itself. Its goal is to explore datanature and scientific issues related to datanature. This does not take into account

the meaning of the data in the natural world.

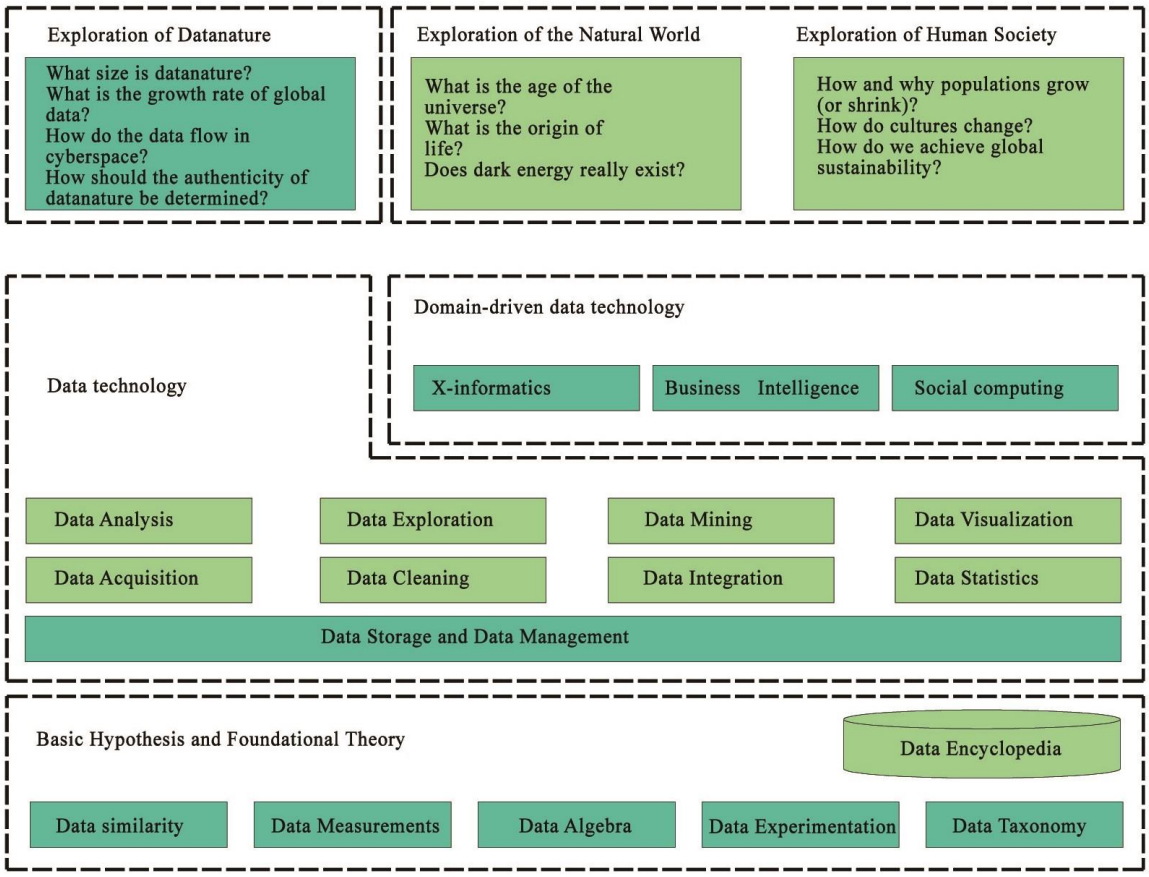
The second component is the study of the rules of the natural world as reflected by data, i.e., the study the natural world performed through the study of data. For example, the purpose of performing a study on data representing a person's behavior is to study that person's behavior.

As mentioned above, studies on data have been under way for some time, and data techniques such as data mining have been developed. However, the data science research community needs to establish fundamental theory and basic methods for scientific observation and measurement and to further develop data techniques. Figure 2 shows the main topics of data science.

Figure 2. Research Topics of Data Science

Conclusions

Data science is gaining more and more and widespread attention, but no consensus viewpoint on what data science is has emerged. As a new science, its objects of study and scientific issues



should not be covered by established sciences. In the present paper, data science is defined as the science of exploring datanature. We believe this is the most logical and accurate definition of data science, and it includes key parts of definitions VP1-VP4.

Reference.

1. Aiden, E., and Michel, J. B. Uncharted: Big data as a lens on human culture. Penguin Group (2013).

2. Cleveland, W. S. Data Science: an action plan for expanding the technical areas of the field of statistics. International Statistical Review 69, 1 (2001), 21-26.

3. Dhar, V. Data Science and Prediction. Commun. ACM 56, 12 (Dec. 2013), 64-73.

4. Hey, T. et al. The fourth paradigm: Data intensive scientific discovery. Microsoft Research. (Oct. 2009); http://research.microsoft.com/en-us/collaboration/fourthparadigm/.

5. Loukides, M. What is data science? An O'Reilly Radar Report 2010; http://radar.oreilly.com/2010/06/what-is-data-science.html.

6. May, M. Life science technologies: Big biological impacts from big data. Science (Science Products, Product Articles), (13 June 2014); http://www.sciencemag.org/site/products/ lst_20140613.xhtml.

7. Provost, F., and Fawcett, T. Data science and its relationship to big data and data-driven decision making. Big Data 1, 1 (2013), 51-59; http://online.liebertpub.com/doi/pdfplus/ 10.1089/big.2013.1508.

8. Smith, F. J. Data Science as an academic discipline. Data Science Journal 5 (2006), 163-164.

9. Zhu, Y. Y. et al. Data Explosion, Data Nature and Dataology. In Proceedings of International Conference on Brain Informatics. 2009.

10. Zhu, Y. Y., and Xiong, Y. Dataology and Data science: Up to Now. http://www.paper.edu.cn/ en_releasepaper/content/4432156, 2011.

附录 1 中文译文

数据科学定义

对数据所反映的自然世界规律的研究的超越

不论是在工业还是学术领域，数据科学都吸引了广泛的关注。新的数据科学研究机构和组织不断登场，例如哥伦比亚数据科学与工程中心和纽约大学数据科学中心。加州伯克利大学，哥伦比亚大学和复旦大学，还有很多大学都开始开设数据科学的课程。Cleveland和Smith提出数据科学应该被当做单独的学科。Facebook, Google, EMC, IBM和其他不少公司都开始专门招聘数据科学家。据哈佛商业评论报道，数据科学家是21世界最性感的工作。目前，关于数据科学的定义有几种不同的观点。但是，目前关于数据科学还没得到广泛接受的定义。我们相信作为一门新的学科，它的研究目标和已经建立的学科并不相同。此外，数据科学扎根的并不是自然或者社会科学研究的领域。

我们团队在中国政府的资助下从1998年就开始研究相关数据科学的技术，我们已经把技术应用到了医疗，金融，运输和其他不少领域。近年来，我们发现了不少数据科学在科学研究和工业领域的应用。在2009年，我们意识到是时候来研究数据本身，挖掘数据科学的概念了。自2010年开始，我们举办了国际数据科学专题研讨会——Dataology。 这个专题研讨会让我们有场所对设计计算机科学，生命科学等领域的数据科学进行讨论。16年来，我们对于数据科学的研究已经取得了不少坚实的成果。我们相信网络空间中的数据已经构成了数据生态。数据科学就是基于数据生态的研究。

关于数据科学目前有几种观点。

第一种观点：数据科学是研究科学数据的科学。

数据科学与技术委员会已经在2002年开办了数据周刊。该委员会认为数据科学就是基于科学数据的管理和利用的研究方法。科学数据现在越来越容易获取到，数据科学越来越被应用到数据密集型的应用上。很多学科都在自己的领域内应用数据科学来处理科学数据。由此，x-informatics应运而生，包括生物信息学，神经信息学，社会学。

例如，numedii公司(硅谷的一家大数据公司)的研究人员，为了预测是否存在药物可用于治疗卵巢癌，就去研究了基因表达数据来自超过2500个卵巢肿瘤样品。

再比如，哈佛大学的数学家艾登和米歇尔使用ngrams GOOGLE研究了美国历史。他们用ngrams寻找两个短语的使用频率：“美国 are”和“美国 is”。最后的搜索结果显示，在美国内战之前，两个短语的使用频率大致相等，但在内战之后，后者远比前者更普遍了。这是表示美国公众接受美国作为一个统一的国家，南北战争前后的水平对比。

从这一点来看，数据主要是指数据的产生和应用的科学研究。这强调了数据的科学管理，处理，和恰当运用科学的数据来支持科学研究，例如支持目前常见的数据密集型的科学研究或科学研究。

第二种观点：数据科学是研究商业数据的科学。

2010年，loukides讨论了到底什么是数据科学，他认为数据科学是衍生出各种数据产品而并不是仅仅对于数据进行应用的科学。在2013年，Provost等人指出，“从数据中提取知识用来解决商业问题”是数据科学的基本概念之一。

为商业智能领域提供支持占了许多数据科学家日常的这个工作的一个重要部分。受此影响，很大比例的商业智能工作者过渡到数据科学家。亚马逊，谷歌，linkedin，facebook，和其他互联网公司为数据科学家和已经建立好的数据科学团队提供不少岗位。这些数据科学家研究和分析各种各样的商业数据，为管理层决策的时候提供辅助服务。例如，亚马逊使用协同过滤算法，获取到了非常高质量的产品推荐效果。再比如，facebook用“你可能认识的人”这个功能来推荐各种各样的朋友来供你联系。

从这一点来看，为了做出决策，通过商业数据来挖掘出知识，来辅助决策也是数据科学的一个方面。这和大多数商业数据科学家的工作很相像。正是出于这个原因，许多商业智能科学家也会被称为数据科学家。然而，与商业智能科学家相比，数据科学更加关注对于各种各样的商业数据的通用分析策略，例如商业智能中的通用策略。

第三种观点：数据科学是一个设计统计，计算以及人工智能（AI）的综合学科。

这种观点经常出现在数据科学家主要做什么的讨论中。人们普遍认为，数据科学家应该具备计算技术，人工智能，以及相关领域的知识和技术，数据科学家不是个人从事，团队多由统计人员，团队的计算机科学家，人工智能专家，与专家组成。

例如，在谷歌和facebook的数据科学家团队是由统计学家，计算机科学家，人工智能科学家，和其他相关领域的专家组成。

这个观点很简单：因为统计，计算技术，以及人工智能都是用来处理和分析数据，他们都是数据科学的自然的一部分。

第四种观点：数据科学的目的是通过从数据中抽取知识来解决科学和商业问题。

在2013年，Dhar从商业和研究的角度讨论了数据科学。他定义的数据科学是针对从数据中获取知识的一般化研究。他还指出，数据科学家一般需要有综合的技能包括统计，机器学习，人工智能，数据库管理和对问题的设计有深刻的理解。

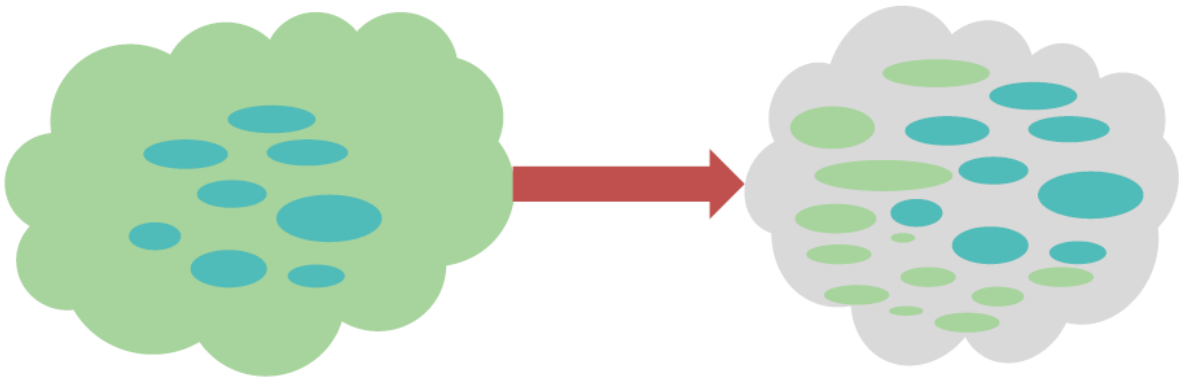
这种观点可以看作是前三个观点的整合。

数据科学到底是什么？

基本思想的基本定义是数据科学是从数据获取知识来支持相关领域的研究和相关领域的商业管理的决策。然而，上述所有工作还不足以建立数据科学作为一种新的，独特的学科分支。这是因为他们的研究对象是自然界中的事物，和他们的研究问题也是已有学科中的已经存在的问题。

随着数字设备的发展，现实中各种各样的事物慢慢开始在网际空间以数据的形式存储。数据在虚拟世界中的各种输入，生成和创造方法越来越多，并且更多样化的，复杂化，超出了人类的控制。越来越多的数据难以被人类理解。数据在网络世界的表现已经显示一个独立的操作系统，像所有的自然世界，在虚拟世界中的数据是指数据自然的。

需要指出的是，在网际空间中数据的类型有两种。第一是数据代表现实中的东西，这类数据称为真实的数据。一个例子是一个个人信息的数据，是代表个人的特点。第二个是不代表现实事物的数据，这类数据称为虚拟的数据。虚拟数据的例子，在自然世界中的没有参考。一个例子是计算机病毒，这是不正常的病毒在自然世界的真实数据表示的病毒；相反，他们只在网络空间的存在（图1）。



(a) Nature　　(b) Datanature

Figure 1. Real data and virtual data. The blue ellipses and the green ellipses in Figure 1(b) denote the real data that represent things in the natural word. Among these, the blue ellipses correspond to those things in the natural world which have been stored in cyberspace in the form of data (i.e., the blue ellipses in Figure 1(a)); the green ellipses correspond to those things in the natural world which would be stored in cyberspace gradually (i.e., the green part in Figure 1 (a)). The grey part in Figure 1(b) denotes the virtual data that do not represent things in the natural world, i.e., the instances of such data have no references in the natural world.

数据自然的形成产生了新的科学问题和研究新的对象。这些新的研究对象并不存在于自然界或人类社会中而是在数据自然，即数据的东西。存在着数据自然的新的科学问题。那么数据自然有多大呢？全球数据的增长率是多少？数据如何在网际空间中流动呢？数据自然的真实性是确定的吗？这些问题不都是由自然或社会科学解决。这些新的科学问题，需要用新的科学研究。

数据科学是研究数据自然和数据本身的学科。在一个基本水平，它涉及到从数据中挖掘知识。因为数据自然中一部分数据代表现实中的事物，从这些数据中获得的知识可以用于自然科学和社会科学。根据第一和第四种观点，这种类型的工作是数据科学。然而，它只是数据科学研究的一部分。

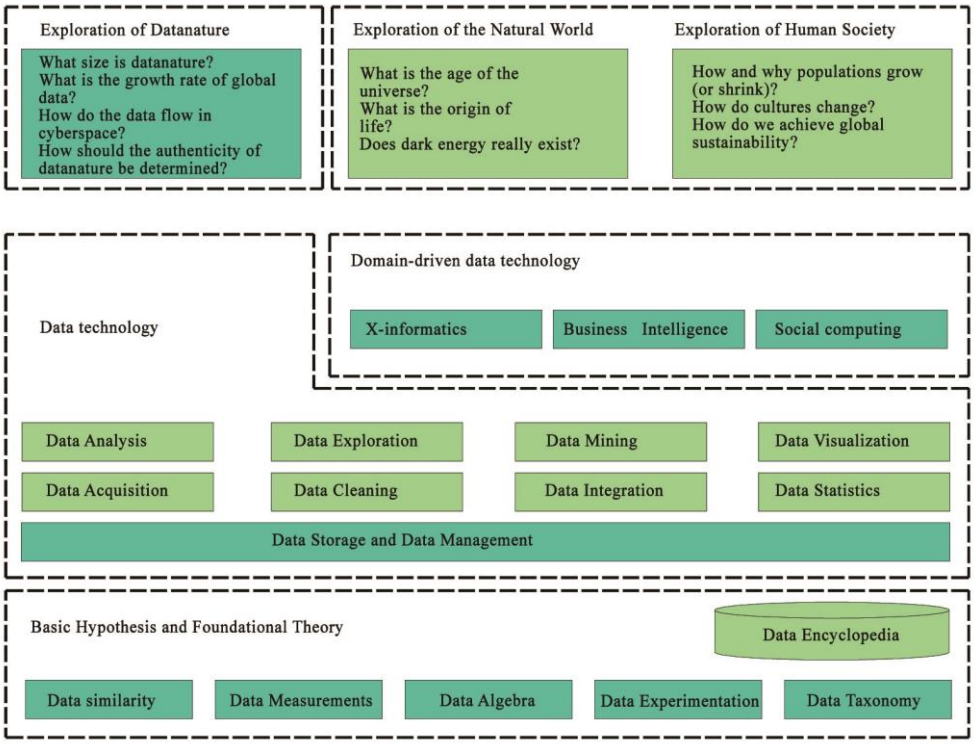
数据科学来定义如下：

数据科学是研究数据自然的理论，方法和技术。它主要有两部分组成。

第一部分是模式和数据本身的规律研究。其目的是探讨有关数据自然和数据自然的相关的科学问题。这没有考虑数据在现实中的意义。

第二部分是研究现实世界对应的数据的规律。

如上所述，研究数据已经进行了一段时间，相关的数据的技术如数据挖掘已经开发。然而，数据科学研究社区需要进一步发展的科学观察和测量的基本理论和基本方法。图2显示了数据科学的主要内容。



结论

数据科学受到越来越广泛的关注，但在数据科学是没有共识的观点出现。作为一种新的科学，它的研究对象和科学问题不应被局限在已经建立的科学领域。在本文中，数据科学的定义是探索数据自然的科学。我们相信这是最合乎逻辑的、准确的数据科学的定义，而且它包括定义第一种和第四种观点的关键部分。