

INSIGHTING INTO DATA

洞见数据之密

八篇大家之言揭示数据关系挖掘深层奥义

实战案例剖析

+

技术方法探索

=

数据价值探索真经



卷首语

最初的数据挖掘技术雏形出现在 20 世纪 60 至 70 年代，那时候全球互联网泡沫才刚刚开始膨胀。随着各行各业数据量的激增以及产品生命周期变化的愈发复杂，数字已变成一种珍贵性资产。

从知识发现、商业智能、预测建模到预测分析，数据挖掘相关技术名称的不断变更，不仅体现了其内涵与定义上的延伸，还意味着它在业务流程和商业决策上产生的越来越重要影响。业务驱动的需求扩展促使数据挖掘从简单的数据检索和统计成长为更复杂的分析预测工具。如今，数据挖掘和分析型 CRM 已发展为主流，不同的数据挖掘软件也随着技术和算法进步而日趋成熟。

同时，数据挖掘是一个分析过程，旨在探索大量数据背后，模式或变量间较为一致的系统关系，并将分析出的模式应用于新的数据子集。也就是说，数据挖掘的最终目标是实现精准预测，特别是对个体行为、商业决策乃至社会层面的关系和发展预测。几十年来，社会科学将交互的强度和频率量化以衡量个人间的关系，因此便有了弱关系和强关系之说。弱关系（松散的熟人）可以帮助传播不

同群体间的思想或创新，有利于信息流动。这种潜在关系的联结也让数据挖掘技术有了真正的用武之地。在这样的运作机制下，关系数据挖掘也对数据特征的准确性提出了更高的要求。

明略数据作为国内领先的企业级大数据技术商业化公司，可以有效帮助企业将多源异构、非结构化的数据进行统一存储并挖掘出其中所隐藏的价值。除了利用传统的数据挖掘方法，如通过人工智能自动地按照一些基本特征对数据进行分类、聚类外，明略还会根据真实的数据样本通过机器学习进行样本训练，从而得出比“人”的主观意识更精确的规则集和模型。在提升数据特征精准性上，明略可以做到从原始数据中提取特征，紧密联系不同领域不同业务选择合适的特征，以实现数据间的区分度。在解决数据关联问题上，明略会把数据转化成类似知识图谱的形式去进行存储，使业务人员能更容易地理解这些数据。

为了在 DT 时代下为企业变革提供更好的支撑和服务，明略在 2015 年 10 月同时发布了数据系统革命性产品 SCOPA(数据关联分析)、数据安全与运维安全的大数据平台产品 MDP(Mininglamp DataPlatform)，以及分布式全量数据挖掘产品 DataInsight。多条业务线和不同类别产品的拓展，已使得明略能更好地在海量社会数据中洞察出真正的数据之密，在信息浪潮中挖掘出更多隐藏在数据中的“宝藏”。

这正如明略所主张的思想：“孤立无序、种类繁多的数据本身不具备价值，只有将数据统一、关联起来，才能发掘信息，发挥价值。”

目录

- 05** 安全有效地输出价值：大数据是这个游戏的名字
- 16** 一线专家谈谈：数据挖掘在实际领域中的那些事儿
- 27** 当 AlphaGo 火了以后，我们来聊聊深度学习
- 41** 数据量决定了特定领域自然语言处理最终效果
- 46** 特定领域自然语言处理最终效果
- 53** MDP 打造新一代高性能、高可用、高安全大数据平台
- 61** 数据关系挖掘算法、技术难点及应用场景分析
- 66** SCOPA 架构升级下的实践与优化

安全有效地输出价值： 大数据是这个游戏的名字

作者 江金陵

【编者按】Hadoop 于 2006 年 1 月 28 日诞生，至今已有 10 年，它改变了企业对数据的存储、处理和分析的过程，加速了大数据的发展，形成了自己的极其火爆的技术生态圈，并受到非常广泛的应用。在 2016 年 Hadoop 十岁生日之际，InfoQ 策划了一个 Hadoop 热点系列文章，为大家梳理 Hadoop 这十年的变化，技术圈的生态状况，回顾以前，激励以后。

要建立一个大数据系统，我们需要从数据流的源头跟踪到最后有价值的输出，并在现有的 Hadoop 和大数据生态圈内根据实际需求挑选并整合各部分合适的组件来构建一个能够支撑多种查询和分析功能的系统平台。这其中既包括了对数据存储的选择，也涵盖了数据线上和线下处理分离等方面的思考和权衡。此外，没有任何一个引入大数据解决方案的商业应用在生产环境上承担的起安全隐患。因此， 本文将从计算框架，NoSQL 数据库，大数据平台安全等三个方面详细阐述在将数据转换成价值的过程中可能产生的技术选型，对比分析不同的应对场景和未来的框架和技术发展方向。

计算框架篇

1. 大数据的价值

只有在能指导人们做出有价值的决定时，数据才能体现其自身的价值。因此，大数据技术要服务于实际的用途，才是有意义的。一般来说，大数据可以从以下三个方面指导人们做出有价值的决定：

- 报表生成（比如根据用户历史点击行为的跟踪和综合分析、应用程序活跃程度和用户粘性计算等）；
- 诊断分析（例如分析为何用户粘性下降、根据日志分析系统为何性能下降、垃圾邮件以及病毒的特征检测等）；
- 决策（例如个性化新闻阅读或歌曲推荐、预测增加哪些功能能增加用户粘性、帮助广告主进行广告精准投放、设定垃圾邮件和病毒拦截策略等）。

进一步来看，大数据技术从以下三个方面解决了传统技术难以达成的目标（如图 1）：



交互式查询: 激活快速决策
» 例如检测一个站点为何变慢并解决



流数据查询: 激活在实时数据上的决策
» 例如欺诈检测，DDoS攻击检测



精细化的数据处理和分析: 激活“更好”的决策
» 例如异常点检测和趋势分析

图 1 大数据的价值

- 在历史数据上的低延迟（交互式）查询，目标是加快决策过程和时间，例如分析一个站点为何变缓慢并尝试修复它；
- 在实时数据上的低延迟查询，目的是帮助用户和应用程序在实时数据上

做出决策， 例如实时检测并阻拦病毒蠕虫（一个病毒蠕虫可以在1.3秒内攻击1百万台主机）；

- 更加精细高级的数据处理算法，这可以帮助用户做出“更好”的决策，例如图数据处理、异常点检测、趋势分析及其他机器学习算法。

2. 蛋糕模式

从将数据转换成价值的角度来说，在 Hadoop 生态圈十年蓬勃成长的过程中，YARN 和 Spark 这二者可以算得上是里程碑事件。Yarn 的出现使得集群资源管理和数据处理流水线分离，大大革新并推动了大数据应用层面各种框架的发展（SQL on Hadoop 框架，流数据，图数据，机器学习）。它使得用户不再受到 MapReduce 开发模式的约束，而是可以创建种类更为丰富的分布式应用程序，并让各类应用程序运行在统一的架构上，消除了为其他框架维护独有资源的开销。就好比一个多层蛋糕，下面两层是 HDFS 和 Yarn，而 MapReduce 就只是蛋糕上层的一根蜡烛而已，在蛋糕上还能插各式各样的蜡烛。在这一架构体系中，总体数据处理分析作业分三块（图 2），在 HBase 上做交互式查询（Apache Phoenix，

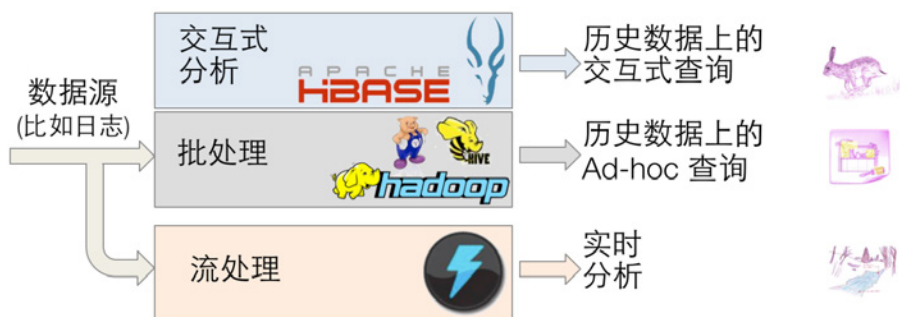


图 2 基于 Yarn 上的开发堆栈

Cloudera Impala 等），在历史数据集上编写 MapReduce 程序抑或利用 Hive 等做批处理业务，另外对于实时流数据分析 Apache Storm 则会是一种标准选择方案。虽然 Yarn 的出现极大地丰富了 Hadoop 生态圈的应用场景，但仍存有两个显而易见的挑战：一是在一个平台上需要维护三个开发堆栈；二是在不同框架内很难共享数据，比如很难在一个框架内对流数据做交互式查询。这也意味着我们

需要一个更为统一和支持更好抽象的计算框架的出现。

3. 一统江湖

Spark 的出现使得批处理任务，交互式查询，实时流数据处理被整合到一个统一的框架内（图 3），同时 Spark 和现有的开源生态系统也能够很好地兼容（Hadoop, HDFS, Yarn, Hive, Flume）。通过启用内存分布数据集，优化迭代工作负载，用户能够更简单地操作数据，并在此基础上开发更为精细的算法，如机器学习和图算法等。有三个最主要的原因促使 Spark 目前成为了时下最火的大数据开源社区（拥有超过来自 200 多个公司的 800 多个 contributors）：

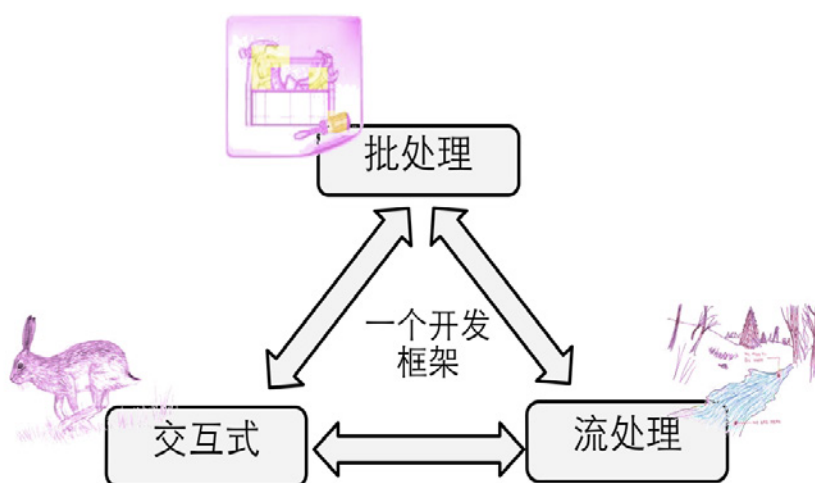


图 3 Spark 统一了开发框架

- Spark 可以扩展部署到超过 8000 节点并处理 PB 级别的数据，同时也提供了很多不错的工具供应用开发者进行管理和部署；
- Spark 提供了一个交互式 shell 供开发者可以用 Scala 或者 Python 即时性试验不同的功能；
- Spark 提供了很多内置函数使得开发者能够比较容易地写出低耦合的并且能够并发执行的代码，这样开发人员就更能集中精力地为用户提供更多的业务功能而不是花费时间在优化并行化代码之上。

当然 Spark 也和当年的 MapReduce 一样不是万灵药，比如对实时性要求很高的流数据处理上 Apache Storm 还是被作为主流选择，因为 Spark Streaming

实际上是 microbatch（将一个流数据按时间片切成 batch，每个 batch 提交一个 job）而不是事件触发实时系统，所以虽然支持者们认为 microbatch 在系统延时性上贡献并不多，但在生产环境中和 Apache Storm 相比还不是特别能满足对低延时要求很高的应用场景。比如在实践过程中，如果统计每条消息的平均处理时间，很容易达到毫秒级别，但一旦统计类似 service assurance（确保某条消息在毫秒基本能被处理完成）的指标，系统的瓶颈有时还是不能避免。但同时我们不能不注意到，在许多用例当中，与流数据的交互以及和静态数据集的结合是很有必要的，例如我们需要在静态数据集上进行分类器的模型计算，并在已有分类器模型的基础上，对实时进入系统的流数据进行交互计算来判定类别。由于 Spark 的系统设计对各类工作（批处理、流处理以及交互式工作）进行了一个共有抽象，并且生态圈内延伸出了许多丰富的库（MLlib 机器学习库、SQL 语言 API、GraphX），使得用户可以在每一批流数据上进行灵活的 Spark 相关操作，在开发上提供了许多便利。

Spark 的成熟使得 Hadoop 生态圈在短短一年之间发生了翻天覆地的变化，Cloudera 和 Hortonworks 纷纷加入了 Spark 阵营，而 Hadoop 项目群中除了 Yarn 之外已经没有项目是必须的了（虽然 Mesos 已在一些场合替代了 Yarn），因为就连 HDFS，Spark 都可以不依赖。但很多时候我们仍然需要像 Impala 这样的依赖分布式文件系统的 MPP 解决方案并利用 Hive 管理文件到表的映射，因此 Hadoop 传统生态圈依然有很强的生命力。

另外在这里简要对比一下交互式分析任务中各类 SQL on Hadoop 框架，因为这也是我们在实际项目实施中经常遇到的问题。我们主要将注意力集中在 Spark SQL，Impala 和 Hive on Tez 上，其中 Spark SQL 是三者之中历史最短的，论文发表在 15 年的 SIGMOD 会议上，原文对比了数据仓库上不同类型的查询在 Shark（Spark 最早对 SQL 接口提供的支持）、Spark SQL 和 Impala 上的性能比较。也就是说，虽然 Spark SQL 在 Shark 的基础上利用 Catalyst optimizer 在代码生成上做了很多优化，但总体性能还是比不上 Impala，尤其是当做 join 操作的时候，Impala 可以利用“predicate pushdown”更早对表进行选择操作从而提高性能。不过 Spark SQL 的 Catalyst optimizer 一直在持续优化中，相信未来

会有更多更好的进展。Cloudera 的 Benchmark 评测中 Impala 一直比其他 SQL on Hadoop 框架性能更加优越，但同时 Hortonworks 评测则指出虽然单个数据仓库查询 Impala 可以在很短的时间内完成，但是一旦并发多个查询 Hive on Tez 的优势就展示出来。另外 Hive on Tez 在 SQL 表达能力也要比 Impala 更强（主要是因为 Impala 的嵌套存储模型导致的），因此根据不同的场景选取不同的解决方案是很有必要的。

4. 各领风骚抑或代有才人出？

近一年比较吸引眼球的 Apache Flink（与 Spark 一样已有 5 年历史，前身已经是柏林理工大学一个研究性项目，被其拥趸推崇为继 MapReduce，Yarn，Spark 之后第四代大数据分析处理框架），与 Spark 相反，Flink 是一个真正的实时流数据处理系统，它将批处理看作是流数据的特例，同 Spark 一样它也在尝试建立一个统一的平台运行批量，流数据，交互式作业以及机器学习，图算法等应用。Flink 有一些设计思路是明显区别于 Spark 的，一个典型的例子是内存管理，Flink 从一开始就坚持自己精确的控制内存使用并且直接操作二进制数据，而 Spark 一直到 1.5 版本都还是试用 java 的内存管理来做数据缓存，这也导致了 Spark 很容易遭受 OOM 以及 JVM GC 带来的性能损失。但是从另外一个角度来说，Spark 中的 RDD 在运行时被存成 java objects 的设计模式也大大降低了用户编程设计门槛，同时随着 Tungsten 项目的引入，Spark 现在也逐渐转向自身的内存管理，具体表现为 Spark 生态圈内从传统的围绕 RDD（分布式 java 对象集合）为核心的开发逐渐转向以 DataFrame（分布式行对象集合）为核心。总的来说，这两个生态圈目前都在互相学习，Flink 的设计基因更为超前一些，但 Spark 社区活跃度大很多，发展到目前毫无疑问是更为成熟的选择，比如对数据源的支持（HBase，Cassandra，Parquet，JSON，ORC）更为丰富以及更为统一简洁的计算表示。另一方面，Apache Flink 作为一个由欧洲大陆发起的项目，目前已经拥有来自北美、欧洲以及亚洲的许多贡献者，这是否能够一改欧洲在开源世界中一贯的被动角色，我们将在未来拭目以待。

NoSQL数据库篇

NoSQL 数据库在主流选择上依旧集中在 MongoDB, HBase 和 Cassandra 这三者之间。在所有的 NoSQL 选择中, 用 C++ 编写的 MongoDB 几乎应该是开发者最快也最易部署的选择。MongoDB 是一个面向文档的数据库, 每个文档 / 记录 / 数据 (包括爬取的网页数据及其他大型对象如视频等) 是以一种 BSON (Binary JSON) 的二进制数据格式存储, 这使得 MongoDB 并不需要事先定义任何模式, 也就是模式自由 (可以把完全不同结构的记录放在同一个数据库里)。MongoDB 对于完全索引的支持在应用上是很方便的, 同时也具备一般 NoSQL 分布式数据库中可扩展, 支持复制和故障恢复等功能。MongoDB 一般应用于高度伸缩性的缓存及大尺寸的 JSON 数据存储业务中, 但不能执行 “JOIN” 操作, 而且数据占用空间也比较大, 最被用户诟病的就是由于 MongoDB 提供的是数据库级锁粒度导致在一些情况下建索引操作会引发整个数据库阻塞。一般来说, MongoDB 完全可以满足一些快速迭代的中小型项目的需求。

下面来主要谈谈 Cassandra 和 HBase 之间的比较选择。Cassandra 和 HBase 有着截然不同的基因血统。HBase 和其底层依赖的系统架构源自于著名的 Google FileSystem (发表于 2003 年) 和 Google BigTable 设计 (发表于 2006 年), 其克服了 HDFS 注重吞吐量却牺牲 I/O 的缺点, 提供了一个存储中间层使得用户或者应用程序可以随机读写数据。具体来说, HBase 的更新和删除操作实际上是先发生在内存 MemStore 中, 当 MemStore 满了以后会 Flush 到 StoreFile, 之后当 StoreFile 文件数量增长到一定阈值后会触发 Compact 合并操作, 因此 HBase 的更新操作其实是不断追加的操作, 而最终所有更新和删除数据的持久化操作都是在之后 Compact 过程中进行的, 这使得应用程序在向内存 MemStore 写入数据后, 所做的修改马上就能得到反映, 用户读到的数据绝不会是陈旧的数据, 保证了 I/O 高性能和数据完全一致性; 另一方面来说, HBase 基于 Hadoop 生态系统的基因就已经决定了他自身的高度可扩展性、容错性。

在数据模型上, Cassandra 和 HBase 类似实现了一个 key-value 提供面向列式存储服务, 其系统设计参考了 Amazon Dynamo (发表于 2007 年) 分布式哈希

(DHT) 的 P2P 结构 (实际上大部分 Cassandra 的初始工作都是由两位从 Amazon 的 Dynamo 组跳槽到 Facebook 的工程师完成), 同样具有很高的可扩展性和容错性等特点。除此之外, 相对 HBase 的主从结构, Cassandra 去中心化的 P2P 结构能够更简单地部署和维护, 比如增加一台机器只需告知 Cassandra 系统新节点在哪, 剩下的交给系统完成就行了。同时, Cassandra 对多数据中心的支持也更好, 如果需要在多个数据中心进行数据迁移 Cassandra 会是一个更优的选择。Eric Brewer 教授提出的经典 CAP 理论认为任何基于网络的数据共享系统, 最多只能满足数据一致性、可用性、分区容忍性三要素中的两个要素。实际分布式系统的设计过程往往都是在一致性与可用性上进行取舍, 相比于 HBase 数据完全一致性的系统设计, Cassandra 选择了在优先考虑数据可用性的基础上让用户自己根据应用程序需求决定系统一致性级别。比如: 用户可以配置 QUORUM 参数来决定系统需要几个节点返回数据才能向客户端做出响应, ONE 指只要有一个节点返回数据就可以对客户端做出响应, ALL 指等于数据复制份数的所有节点都返回结果才能向客户端做出响应, 对于数据一致性要求不是特别高的可以选择 ONE, 它是最快的一种方式。

从基因和发展历史上来说, HBase 更适合用做数据仓库和大规模数据处理与分析 (比如对网页数据建立索引), 而 Cassandra 则更适合用作实时事务和交互式查询服务。Cassandra 在国外市场占有率和发展要远比国内红火, 在不少权威测评网站上排名都已经超过了 HBase。目前 Apache Cassandra 的商业化版本主要由软件公司 DataStax 进行开发和销售推广。另外还有一些 NoSQL 分布式数据库如 Riak, CouchDB 也都在各自支持的厂商推动下取得了不错的发展。

虽然我们也考虑到了 HBase 在实际应用中的不便之处比如对二级索引的支持程度不够 (只支持通过单个行键访问, 通过行键的范围查询, 全表扫描), 不过在明略的大数据基础平台上, 目前整合的是依然是 HBase, 理由也很简单, HBase 出身就与 Hadoop 的生态系统紧密集成, 其能够很容易与其他 SQL on Hadoop 框架 (Cloudera Impala, Apache Phoenix, or Hive on Tez) 进行整合, 而不需要重新部署一套分布式数据库系统, 而且可以很方便地将同样的数据内容在同一个生态系统中根据不同框架需要来变换存储格式 (比如存储成 Hive 表或

者 Parquet 格式)。我们在很多项目中都有需要用到多种 SQL on Hadoop 框架应对不同应用场景的情况，也体会到了在同一生态系统下部署多种框架的简便性。但同时我们也遇到了一些问题，因为 HBase 项目本身与 HDFS 和 Zookeeper 系统分别是由不同开源团队进行维护的，所以在系统整合时我们需要先对 HBase 所依赖的其他模块进行设置再对 HBase 进行配置，在一定程度上降低了系统维护的友好性。目前我们也已经在考虑将 Cassandra 应用到一些新的客户项目中，因为很多企业级的应用都需要将线上线下数据库进行分离，HBase 更适合存储离线处理的结果和数据仓库，而更适合用作实时事务和并发交互性能更好的 Cassandra 作为线上服务数据库会是一种很好的选择。

大数据安全篇

随着越来越多各式各样的数据被存储在大数据系统中，任何对企业级数据的破坏都是灾难性的，从侵犯隐私到监管违规，甚至会造成公司品牌的破坏并最终影响到股东收益。给大数据系统提供全面且有效的安全解决方案的需求已经十分迫切：

- 大数据系统存储着许多重要且敏感的数据，这些数据是企业长久以来的财富；
- 与大数据系统互动的外部系统是动态变化的，这会给系统引入新的安全隐患；
- 在一个企业的内部，不同 Business Units 会用不同的方式与大数据系统进行交互，比如线上的系统会实时给集群推送数据、数据科学家团队则需要分析存储在数据仓库内的历史数据、运维团队则会需要对大数据系统拥有管理权限。

因此为了保护公司业务、客户、财务和名誉免于被侵害，大数据系统运维团队必须将系统安全高度提高到和其他遗留系统一样的级别。同时大数据系统并不意味着引入大的安全隐患，通过精细完整的设计，仍然能够把一些传统的系统安全解决方案对接到最新的大数据集群系统中。一般来说，一个完整的企业级安

全框架包括五个部分：

- Administration：大数据集群系统的集中式管理，设定全局一致的安全策略
- Authentication：对用户和系统的认证
- Authorization：授权个人用户和组对数据的访问权限
- Audit：维护数据访问的日志记录
- Data Protection：数据脱敏和加密以达到保护数据的目的

系统管理员要能够提供覆盖以上五个部分的企业级安全基础设施，否则任何一环的缺失都可能给整个系统引入安全性风险。在大数据系统安全集中式管理平台这块，由 Hortonworks 推出的开源项目 Apache Ranger 就可以十分全面地为用户提供 Hadoop 生态圈的集中安全策略的管理，并解决授权 (Authorization) 和审计 (Audit)。例如，运维管理员可以轻松地为个人用户和组对文件、数据等的访问策略，然后审计对数据源的访问。与 Ranger 提供相似功能的还有 Cloudera 推出的 Apache Sentry 项目，相比较而言 Ranger 的功能会更全面一些。而在认证 (Authentication) 方面，一种普遍采用的解决方案是将基于 Kerberos 的认证方案对接到企业内部的 LDAP 环境中，Kerberos 也是唯一为 Hadoop 全面实施的验证技术。另外值得一提的是 Apache Knox Gateway 项目，与 Ranger 提高集群内部组件以及用户互相访问的安全不同，Knox 提供的是 Hadoop 集群与外界的唯一交互接口，也就是说所有与集群交互的 REST API 都通过 Knox 处理。这样，Knox 就给大数据系统提供了一个很好的基于边缘的安全 (perimeter-based security)。

基于以上提到的五个安全指标和 Hadoop 生态圈安全相关的开源项目，已经足已证明基于 Hadoop 的大数据平台我们是能够构建一个集中、一致、全面且有效的安全解决方案。我们明略的 MDP 大数据平台就是这样一款兼顾数据安全与运维安全的产品。

结语

本文主要介绍了如何将 Hadoop 和大数据生态圈的各部分重要组件有机地联

系在一起去创建一个能够支撑批处理、交互式和实时分析工作的大数据平台系统。其中，我们重点尝试从计算框架、NoSQL 数据库以及大数据平台安全这三方面分析了在不同的应用场景中相应的技术选型以及需要考虑到的权衡点，希望让大家对如何建立一个完整可用的安全大数据平台能有一个直观的认识。

江金陵，明略数据数据科学家，中山大学本科，硕士毕业于沙特阿拉伯阿卜杜拉国王科技大学，博士就读于丹麦奥尔堡大学，攻读博士期间赴斯德歌尔摩参与创立一款个性化新闻阅读工具并提名瑞典最佳新媒体类移动应用，后加入欧洲前三大博彩公司 Unibet 负责实时个性化赛事推荐系统的大数据平台开发工作。他曾在 ICDE、ICDM 等数据库和数据挖掘顶级会议中发表过学术文章，对大数据环境下的搜索、推荐、自然语言处理等方面均有十分丰富的经验。目前供职于明略数据数据科学家团队，负责公安和金融领域的大数据建模与开发工作。

一线专家谈谈： 数据挖掘在实际领域中的那些事儿

作者 余伟

【编者按】本文是4月18日大数据杂谈群分享的内容。关注“大数据杂谈”公众号，点击“加群学习”，更多大牛一手技术分享等着你。讲师：余伟（明略数据技术合伙人兼研究院执行院长）。

企业中的数据挖掘

我们先来看看在企业中数据挖掘都是怎么做的，以及有着哪些问题。

图1中的左边是SPSS在1999年提出的《跨行业数据挖掘标准流程》，在图中定义了数据挖掘的6个步骤。虽然这个图已经提出有10几年了，但是在大数据环境下，这个流程依然适用。

1. 理解商业问题。这需要大数据科学家和行业专业，以及客户的业务专家一起来明确问题。这是整个大数据挖掘中最关键的一步。如果不理解业务就贸然开做，最后的项目一定是失败的。
2. 分析数据。当明确了业务问题之后，我们就需要去分析数据，看看到



图 1

底哪些数据能够支撑我们的业务，用哪些数据去解决问题。在这个阶段，我们可能发现数据不足，或者数据质量太差，这个时候就可能要寻求第三方数据的帮助，或者规划如何去采集更多的数据了。

数据挖掘。前两步都是在做数据挖掘前的准备，当业务明确，数据可用时，我们就正式开始数据挖掘了。

1. 提取特征

首先我们要对数据进行处理，从数据中提取特征。这是数据挖掘非常关键的一步，特征的好坏直接影响最终模型的效果。在数据挖掘过程中，算法其实并不是最主要的因素，影响效果最直接的因素就是特征。

良好的特征需要有非常好的区分度，只有这些特征，才能很好的去解决问题。举个例子，我们要辨别一个西瓜是好是坏，可能颜色是一个特征，条纹，重量，瓜蒂也是特征。但是，大家都知道西瓜一般都是绿色的，所以用绿色去作为判别西瓜好坏是没有区分度的。而条纹，重量，瓜蒂是判别一个西瓜是好是坏非常重要的因素，因此他们是好特征。

我们在解决不同问题时，所用的特征是不一样的。可能在解决某个问题有用的特征在解决另外一个问题时就不具备区分度。因此，我们必须紧密的联系业务，去选择合适的特征。

在提取特征时，因为我们是大数据挖掘，所以要使用大数据技术去从原始数据中提取特征。这需要大数据科学家有着非常丰富的大数据处理技能。

2. 建立模型

当特征提取完毕后，我们就需要去应用算法建立模型了。在实际的建模过程中，由于数据量过于庞大，算法训练过程往往十分缓慢，如何加速算法计算速度，是一个非常突出的问题。

此外，由于传统的数据挖掘算法都是针对小数据集的，当数据规模到了一台服务器无法处理的程度，传统的数据挖掘算法就不再使用。此时，我们需要有新的数据挖掘技术来支持大数据上的数据挖掘。

当模型建立完成之后，我们需要对模型进行评估，来确定模型效果。此时最重要的是建立模型的评价指标。这个评价指标必须是要结合业务来建立的。当模型效果不佳时，我们要回到特征提取，建模过程来不断的迭代，甚至可能要重新分析业务和数据。

3. 后期工作

当一个效果非常好的模型建立完毕了，我们的数据挖掘就结束了吗？传统的数据挖掘软件往往只做到模型建立这一步，但是在模型建立完成之后还有很多工作要做。我们如何将模型在生产系统中使用起来，如何去管理、运行、维护、扩展模型。

我们先来看看 DataInsight 对业务的支持（见图 2）。

用户需要针对不同的业务去建立不同的模型，这个建模过程可以由用户自己完成，也可以由明略的大数据科学家去完成。建立好的模型以插件的形式插入到 DataInsight 中去，方便模型的管理和扩展。

用户的业务系统会通过 API 和 DataInsight 进行通信，来运行或者更新 DataInsight 中插入的模型。

一个典型的 DataInsight 模型运行过程如下：用户通过 API 调用 DataInsight，在请求中指定模型，模型的输入和模型的输出。DataInsight 会



图 2

将数据从数据源中取出，送入模型，并且将模型分成多个步骤，并行化的在分布式执行引擎中运行。当模型运行完毕后，结果将送入用户指定的目的数据库中。这样，用户的应用系统就可以直接从目的数据库中获得模型运行的最新结果了。

DataInsight 中将解决客户业务问题的模型成为业务模型，或者应用。DataInsight 对业务模型也进行了一定层次的抽象。每个业务模型都是由若干步骤组成的。每个步骤被称作一个算子。

图 3 是一个文本分类的业务模型，其解决的问题是将若干文本进行分类。例



图 3

如我们有很多文章，我们要对每篇文章的情感进行分类，就可以使用这个模型。

我们将文本分类模型抽象为很多算子的组合。每个算子都是对数据进行了某种转换，将一组输入转化为一组输出。这个转化过程可能是对数据进行的预处理，也可能是某种机器学习算法。

每个算子都有输入和输出，且算子的输出可以作为另外一个算子的输入。这样，整个业务模型就抽象成了一个有向无环图(DAG)。DataInsight 在执行模型时，会去调度模型中的每个算子，将适合分布式计算的算子送入不同的执行容器中去运行，加速了整个模型的计算速度（见图 4）。

DataInsight 总体的体系架构见图 5。下面我们介绍一下明略在各个领域中的一些案例。由于时间关系，我这里只举两个案例。



图 4、图 5

精准营销

明略是从秒针系统拆分出来的，秒针系统是一家以互联网精准营销为主要业务的公司，因此明略在精准营销方面有着接近 10 年的积累。

首先，明略的大数据平台 MDP 会将企业各种自由数据，包括 CRM 数据、交易行为数据以及官网数据等，和第三方数据一起收集起来，并对这些数据进行关联和打通，一起存储到大数据平台 MDP 中去。

我们针对企业不同的业务，建立多个模型，例如智能推荐模型，用户画像模型，消费预测模型，商圈聚类模型等等，这些模型作为插件插入到我们的大数据挖掘平台 DataInsight 中去。

我们可以将原始数据从 MDP 中取出，经过 DataInsight 中模型的计算之后，生成最终的结果数据，结果数据将送入用户画像系统和推荐系统的离线部分。

用户的推荐系统分为在线和离线两个部分，离线推荐的结果就是 DataInsight 中计算出来的结果。在线推荐系统将会接收一个在线的推荐请求，通过客户画像系统和离线推荐结果，并结合当时的一些场景，共同计算出最终向用户推荐的物品。

明略的精准营销系统已经应用到了个性化推荐、精准营销、用户洞察、广告投放等多个领域，并取得了良好的效果。

智能推荐算法

和传统的协同过滤算法不一样，这个算法是采用了分类的思想，通过分类的方法来实现推荐的（见图 6）。

首先，在进行推荐之前，我们必须明确推荐的目标。那就是向用户推荐用户感兴趣的物品。这里的物品可以是商品，也可以是广告，甚至是文章、电影、音乐等等。

然后我们需要去寻找解决这个问题所需的数据。我们有物品内容数据库，用户 CRM 数据库，以及用户行为数据。

解决了目标和数据之后，我们就需要采集一批有标注的样本。因为是采用的



图 6

分类算法，这是有监督的算法，所以标注样本是建模的第一步工作。标注就是通过人工来判定用户是否对某个物品感兴趣。

标注问题解决后，我们就需要从数据中提取特征。我们的特征分为 3 类：物品自身属性，比如我们推荐的是手机，手机型号，手机价格，手机颜色都是物品的自身属性。其次，我们要提取人的属性，比如人的性别、年龄、收入、教育程度一类。最后，我们还需要知道人和物品的交互关系，他是浏览过商品还是加入过购物车，还是点击过商品，甚至购买过该商品。除了和推荐的商品之间的关系之外，我们还可以将用户和其他商品之间的关系也作为特征。

这样，我们就可以通过分类算法去建立模型了。常用的分类算法我们都可以尝试，诸如 GBDT，逻辑回归，SVM 等等。

当模型建立完毕之后，我们就可以得到分类结果了。分类结果是某用户对某商品是否感兴趣，以及感兴趣的程度。感兴趣的程度我们可以通过概率来表示。

有了分类结果还不是我们最终的推荐结果。我们根据分类概率对结果进行排序，最后选出 TopK 个结果作为最终结果返回。

设备诊断

我们的第二个案例是有关设备诊断方案的。我们知道，工业 4.0 是目前比较火热的一个话题。而设备诊断正是工业 4.0 中非常重要的一个应用。

设备诊断又分为故障诊断和故障预测两大类。故障诊断是当一个设备出现故障，我们需要辨别该故障的类型。故障预测是我们要预测出某个设备在未来会不会出现故障。这是两个截然不同的问题，但是处理的方法是类似的。故障诊断和故障预测已经在多个行业中得到应用，并且已经取得了非常突出的效果。

明略的故障诊断方案如图 7。



图 7

首先，各种设备的数据通过 ETL 汇聚进大数据平台中去。这些数据包括传感器实时数据，设备历史数据，时间历史数据等等。

然后，在 DataInsight 中建立故障诊断和故障预测模型，来对原始的数据进行分析，并得到诊断和预测结果。

DataInsight 中的模型会部署到生产系统中去，通过 API 和故障诊断和故障预测应用进行交互，提供最终的分析结果给到应用，在应用中根据分析结果进行各种统计和可视化的展现。

进行故障诊断和故障预测建模有两种方式，其一是传统的方式，其二是通过深度学习的方式。

这个过程中首先我们要对故障进行标注。对于故障诊断，我们要标注的是何种类型的故障，对于故障预测，我们要标注的是有没有发生故障。标注的工作是专业性极强的工作，一般需要用户的专家来进行标注。

对于传统方法而言，最复杂的部分是特征选取。上文我们也讲到，只有那些有

强区分度的特征才能有效的支持最终的模型。所以，需要由业务专家来指导如何从原始数据中提取特征。这就需要将业务专家的经验程序化，将人的知识变为机器能够处理的方法。这是非常困难的。

当特征提取完了之后，我们会采用分类算法来训练模型，最终得到故障诊断和故障预测的结果。

深度学习

在传统方法之外，我们还可以通过深度学习的方法来进行故障的诊断和预测，深度学习方法示意图（见图 8）。

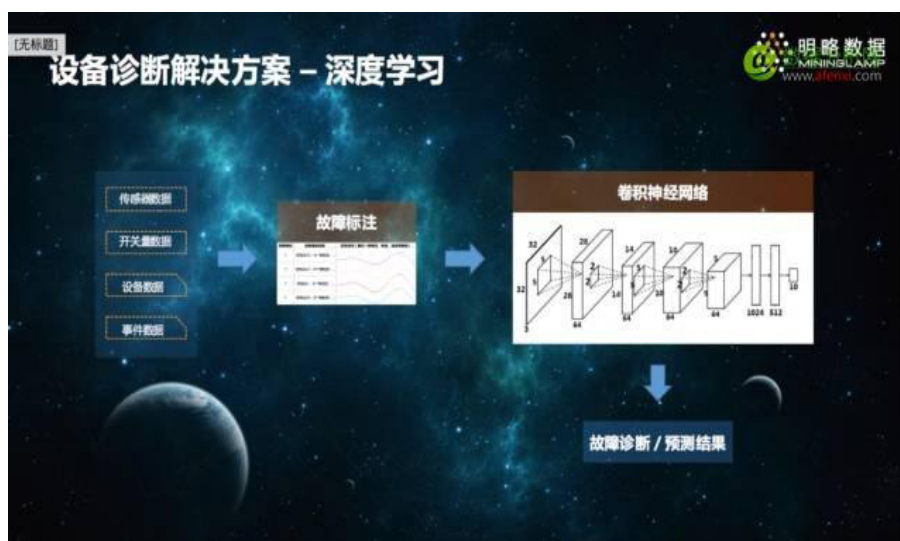


图 8

比起传统的方法，故障标注这一步是省不掉的，因为我们用的还是一个有监督的方法，这个方法必须要有一批标注好的样本。

和传统方法不一样的是，我们直接将样本送入深度学习算法，常用的如卷积神经网络去进行训练，来得到最终的故障诊断和预测的结果。

相比传统方法，深度学习方法省却了特征提取的过程。我们通过深度学习算法直接从原始数据中学习，省却了专家指导的过程。深度学习方法甚至能够学习到专家所不知道，或者在专家潜意识内但无法表达出来的特征。通过深度学习算法出来的模型，其效果往往好于传统方法的模型。

但是，深度学习算法对数据量的要求非常大。只有有大量训练样本才能使用深度学习。这在现实的工作中可能是一个问题。

Q&A:

Q1: 数据互联行业主要是指哪些业务？

A1: 数据互联是明略将第三方数据引入到企业中和企业自有数据结合起来去做数据挖掘的业务，这个业务需要对第三方数据如何与企业数据融合，去进行数据挖掘有比较深的理解。

简单的说，就是帮助用户分析需要什么样的数据，以及从何处去获得这些数据，外部数据和内部数据如何打通，如何去数据挖掘

Q2: 请问领域知识和数据专业知识哪个在实际工作中起的作用更大？

A2: 领域知识和数据专业知识应用的场景不一样。在进行数据挖掘之前，我们首先需要有领域知识。必须明白要解决的问题是什么。只有有了领域知识，并且有数据知识，才能把业务转化为数据挖掘的问题，在进行数据挖掘过程中，数据挖掘知识可能是更关键的，因为你要知道如何去解决这个问题。但是，进行数据挖掘时，还必须根据业务对模型进行调整。

刚才我也说了，模型调优必须建立合理的评价指标。这个评价指标根据不同的业务可能是不一样的。所以必须有业务知识才能知道如何去调优，才能知道什么样的模型是符合业务需要的，所以，在实际的数据挖掘过程中，领域知识和数据挖掘专业知识都是非常重要的，如果缺乏了任何一种，可能都很难取得比较好的效果。另外大数据挖掘中大数据处理能力也很重要，如果不会处理大数据，或者没有良好的编程能力，也是很难做好的

Q3: 请问一下明略大数据在特征工程上有哪些比较好的经验呢？

A3: 其实特征工程是一个非常 dirty 的活，需要大量的尝试性工作，明略的经验就是，在做特征工程时，了解业务是第一位，然后需要深入的去调查客户的每一张表，搞明白每一张表的每一个字段，以及字段间的关联关系，我们在实际工作中，经常要调研几百张表去找到我们需要的数据，此外，作为一个合格的数据挖掘人员，或者数据科学家，敏锐力非常重要，能够结合业务知道可以从数据中提取哪些特征。特征提取出来之后，是否是一个好的特征其实是不知道的。

我们可以大胆的尝试，多选取一些特征过来。然后在通过特征选择去进行筛选。特征工程是实际建模中最耗人力的过程。我们建模大概 70-80% 的时间都耗费在这个上面。

Q4: 请问在进行数据挖掘之前的怎么解决数据质量问题？

A4: 坦白的说，数据质量也是困扰我们的问题，目前我们遇到的客户，坦白的说数据很多都是碎片化的。可能是因为之前他们忽略了某些数据的收集，或者他们的数据只是总体样本的一小部分，对于第一种客户，我们会帮助他们制定如何去收集更多的数据，只有数据有了积累，数据质量问题才会解决。对于第二种客户，我们会帮助引入第三方数据，用第三方数据来补充客户现有的数据，大数据的数据质量差是有目共睹的，但是，正是由于数据量大，数据类型多，我们才能从大数据的沙子中挖到金子。如果传统数据是富矿石，大数据就是贫矿石，大数据数据只能以量去取代质。

Q5: 目前未回答问题中排名最高的是这个：二分类分类算法中，负面情况占比很小，训练集数据负面数据如何按比例分？训练集需要调高负面数据的比例吗？对算法有什么影响？

A5: 分类问题中对正负样本的平衡是必须的，这个也是影响最后分类结果的一个很重要的因素，如果样本不平衡，能做的事情是样本增益和样本抽样。比如正样本远远小于负样本，可以对正样本进行复制，或者加上随机扰动来扩充正样本，或者直接对负样本进行抽样。达到一定的正负样本比，这样最终的效果会比较好。我们的经验是正负样本比 1: 5 左右比较适合，达到一定的正负样本比，这样最终的效果会比较好。我们的经验是正负样本比 1: 5 左右比较适合。

余伟，明略数据技术合伙人兼研究院执行院长，2006 年硕士毕业于清华大学计算机科学与技术系。2006-2011 年先后供职于 IBM 和 HP，从事软件开发与算法研究工作。2011 年加入秒针系统，带领算法团队开展大数据下的数据挖掘技术研究。2013 年底作为技术合伙人加入明略数据，负责明略数据数据挖掘相关工作。目前作为明略数据研究院执行院长负责明略数据挖掘产品和项目方面的工作。

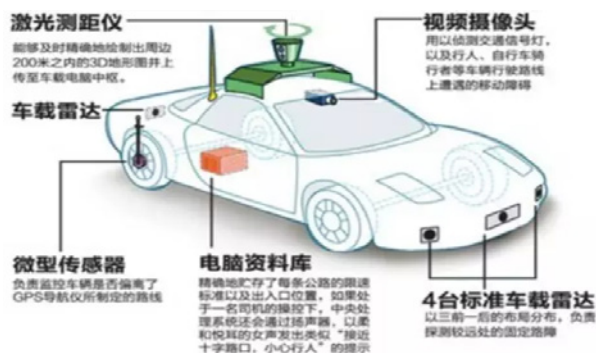
当 AlphaGo 火了以后， 我们来聊聊深度学习

作者 黄文坚

深度学习是我们明略重要的研究方向，是未来实现很多令人惊叹的功能的工具，也可以说是通向人工智能的必经之路。

1. 深度学习的丰富应用

Google 研究的无人驾驶，其组件由两个部分组成，一个是眼睛，一个是大脑，眼睛是激光测距仪和视频摄像头，汽车收集到这些视频信号之后，并不能很好的识别，为了让汽车能理解我们需要一个大脑，这个大脑就是深度学习，通过深度学习我们可以告诉我们的车载的计算机，现在前面有什么样的物体，并且结构化的抽取出来。

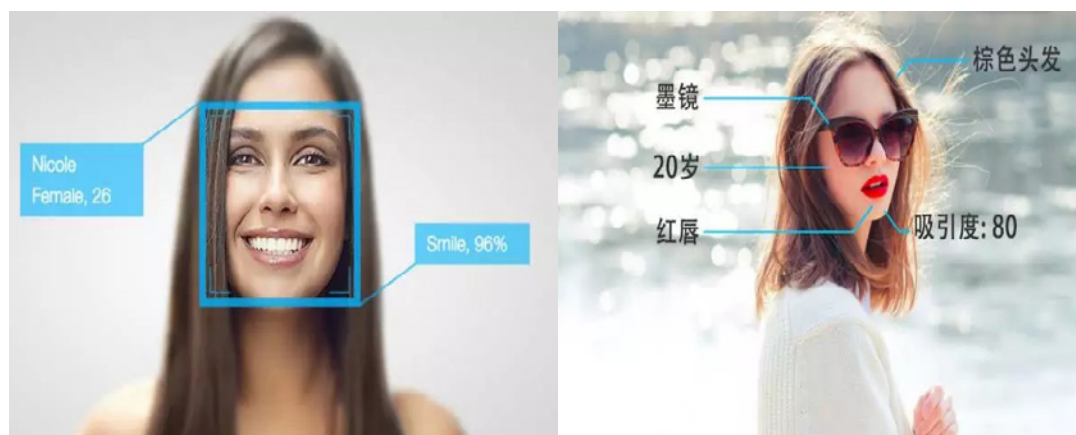


比如说这个是通过挡风玻璃看到的画面，让机器理解，必须要判断视野内的物体是移动还是静止，如果是静止的话，可以当作是安全的物体，只需避让即可，如果是移动的物体，那么还需要我们判断他的速度和行驶方向进行相应的路线规划。



人脸识别，我们有很多技术做人脸识别，人脸识别可以做什么其他的東西呢？

深度学习不止告诉我们人脸在图片中哪个位置，甚至告诉我这个人是谁的脸，是男性、女性，多大岁数都可以学习出来，包括人脸部的重要结点位置可以猜出来这个人是什么样的表情，甚至通过分析他嘴唇的动作，可以说这个人在说什么话，包括头发的颜色，戴什么样的墨镜，嘴唇涂什么样的唇膏都可以识别出来。



格林深瞳的例子，比如说我们在重要机构里面可以有安防监控，深度学习训练的卷积神经网络 CNN，可以识别被监控的人员是否有异常的举动，还有就是对

车辆的追捕，这辆车是否有逃逸的可能性，超速行驶，逆行变道的风险。

AlphaGo 2016 年 3 月，Google DeepMind 研发的 AlphaGo 4:1 战胜了世界冠军李世石。标志了一个时代的终结和一个时代的开始，人类在完全信息博弈的竞技中败北，人工智能发展的元年开始。

围棋很难被攻破的原因就是复杂度太高了。

每一步棋都有 300 多种可能，一盘棋平均有 200 多步，总的状态数量超过了整个宇宙中所有原子的数量，不可能被搜索完整的状态，我们只能通过估算和直觉进行围棋的计算和思考。

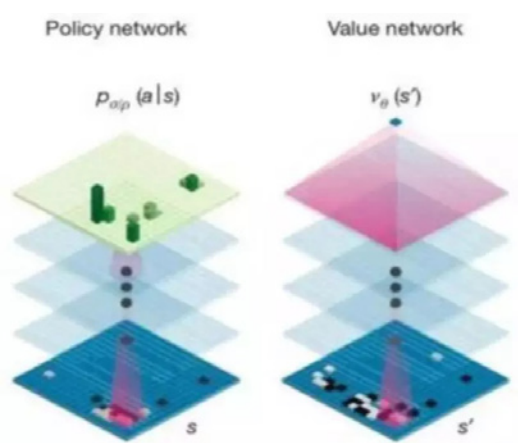
象棋很早就被攻破了，围棋可以坚持这么久。**深度学习可以让机器有人类的直觉**，预测人下一步要走什么，同时分析及其应该走哪一步。所以说 DeepMind 研发的 AlphaGo，基本上做到了知己知彼。

我们再来看看这两个 DeepMind 深度学习的网络，左边是策略网络，我走到一步的时候，分析棋盘上每个位置有多大价值，给每个位置打一个分数。右边这个估值网络是估算黑白双方的胜率的神经网络。

通过这两个网络的结合，再加上一些之前通用搜索的方法，比如蒙特卡洛搜索树，可以让计算机拥有一个非常强的对战能力。

事实上是通过复盘的结果，AlphaGo 和李世石对战的时候，AlphaGo 从一开始就认为自己的胜率有 60% 以上，到最后基本达到了 90%，他对整个棋盘的控制超过了人类的理解了，情况并不是很多评论员所认为的可能双方还是均势，李世石还有机会等等。大局完全都在 AlphaGo 的掌握当中。Deep Q Net，深度强化学习可以教会机器人如何灵活使用机械臂完成任务。

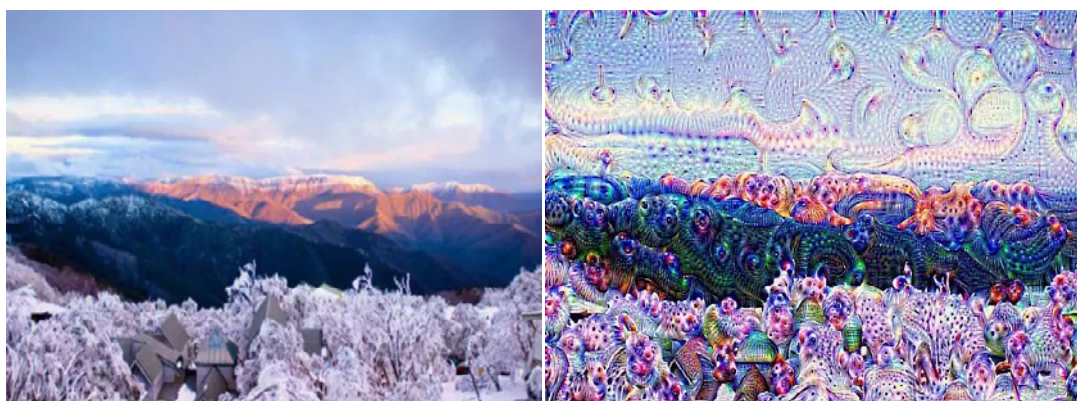
如果之前让一个机器人编程，让他去夹一个物体，不能有太多的干扰，否则就无法实现准确的抓取。现在我随便放一盒子东西，深度强化网络可以自动训练这个机器人拿什么样的物体，同时训练它怎么去夹，第一次没有夹到那就再学习，



再尝试，直到学会。可以说深度学习让机器人拥有几岁小孩拾起物体的能力。

Google DeepDream 实现梦幻般的图片生产，仿若梦魇一般。

大家看这个图，下边这个图是不是有点抽象，这个画是用深度学习网络自动生成出来的，基本原理就是人观察一张图片的时候，记不住所有的细节，在我们脑子里重构的时候会用之前的经验和概念在脑中塑造一个新的图片，而深度学习也是这个意思，在大数据量的需要上，积累了很多过往的经验和数据，我们给他一幅图片重构的时候，就制造出一个仿佛做梦或者脑海中胡思乱想的时候对这个图片产生的理解。所以我们可以说，它已经具备了人类对事物抽象和重构的能力。



使用深度学习实现的 EasyStyle，可以将任意图片内容与另一种图片风格融合。

这个图可能大家很熟悉，最上边这位是美国总统竞选人 Trump，中间这幅画是著名画家的画作，通过深度神经网络结合我们可以合成下面的图，没有进行任何算法的调优，它获取上边这个图内容的信息，再获取中间这个图风格的信息，完美的结合就成了中间这张图。

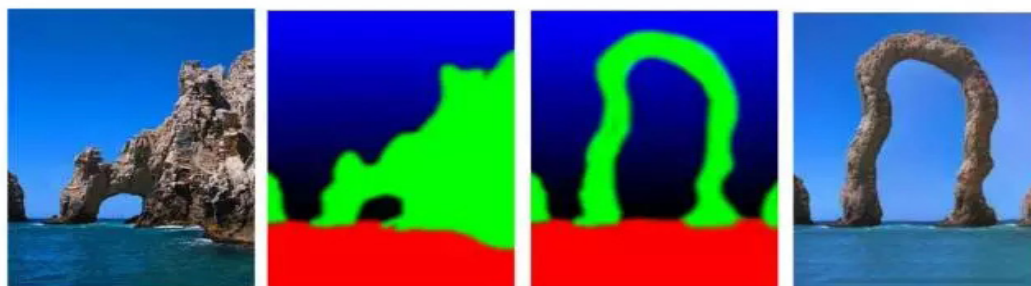


Neural Doodle：将涂鸦变成绘画

比如说随手涂鸦一幅画，可以得到一幅像模像样的一幅山水画。

我们还可以先解析一幅图的主要组成部分，然后调整其中的形状，再把原来的图重构出来，我们可以得出现实生活中不存在的图，这个是类似于人脑对物体的解析和重构的能力。

Image Analogies：使用深度学习变形图片



Deep Q Net：深度强化网络实现AI自动玩游戏

GoogleDeepMind 除了做围棋软件还有实现自动玩游戏的 AI，人类学习并不是一个监督和非监督的过程，是一个奖惩的机制，你做对的时候会有好的刺激，比如说我哭了，我妈妈过来把饭拿过来了，我吃了，很高兴，我下次可能饿了还要再哭。这套系统也是这样的，随即



采取一些策略获得比较高分的时候，他会记住这个策略。这幅图是太空大战游戏，使用程序玩游戏已经超过了世界上玩这个游戏选手的最高水平了。

动态记忆网络实现的图片问答系统

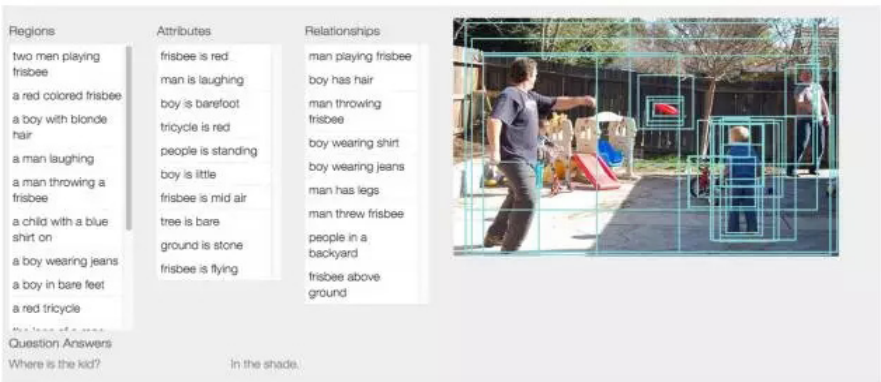
我们可以看看下面这副图，左边是使用一个基于 LSTM 长短期记忆网络的动态来对一段语言进行理解，并回答问题。而右边则是直接对图片进行提问并让计算机回答，使用的技术是动态记忆网络。

目前我们可以做到这种程度，问左上角大巴的颜色是什么，最后转换成语言回答，虽然回答只是简单的单词，但是事实上深度神经网络已经理解你的问题，

Visual Genome，不止对图片分类，还要看出有什么关联。

比如说这里有一个女人，戴着帽子，和帽子是什么关系，是佩戴的关系。她拿着吉他是拿和演奏的关系，我们要找不同物体之间存在的关系。

下图是一个简单的例子， 深度神经网络可以解析出这个图片上有两个成人和小孩，小孩扔飞盘，大人在看，我们要把关联的关系、动态的关系都挖掘出来。

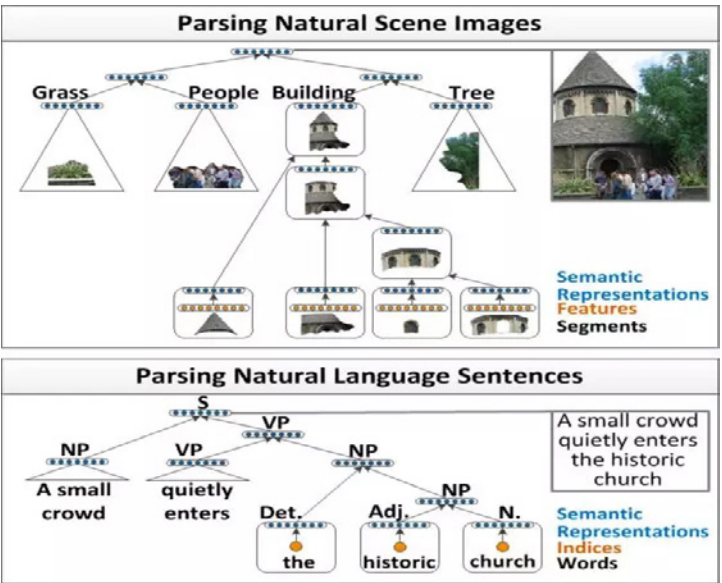


图像识别与 NLP，使用 Deep Learning 解析图像中的结构化信息，并生成描述性语言。

我们看看在下图上能做什么，我们可以让深度神经网络先尝试理解这幅画的结构，然后再用语言把这幅画描述出来，比如生成这样一段话：这张图右边有一棵树，左边有一个塔，塔有一个塔尖和一个塔身组成，塔身上有三个窗户，有一个门。塔前有许多人在站着。

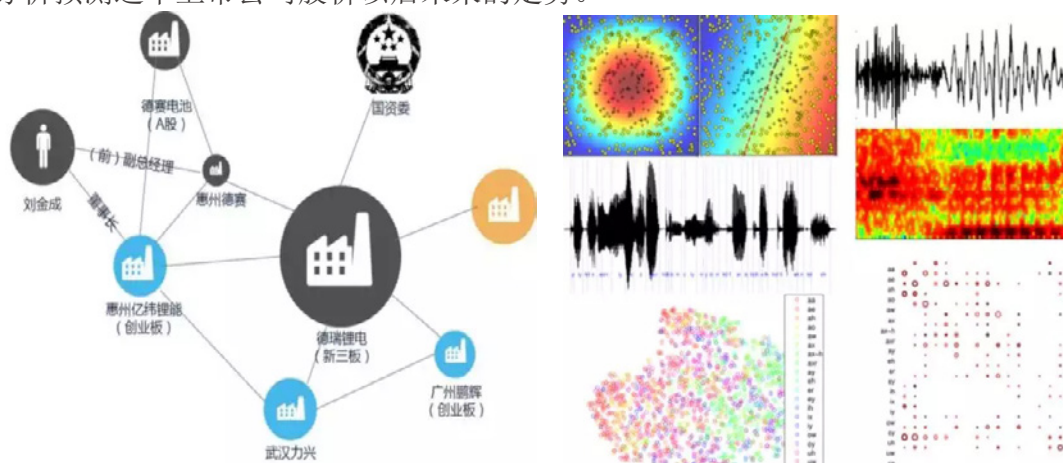
Word Embedding, 或者 Distributed Representation, 中文叫词向量，是使用深度学习学习出来的单词的向量化表示，有如下特性: $\text{King} - \text{Man} \approx \text{Queen} - \text{Woman}$

同时意思相似的词，在空间位置上距离相近。



词向量把我们常用的词汇转化为空间中的某一个点，点有什么特性呢：如果词汇意思相近的话，在空间中位置应该也是相近的，左上角可以找到许多点都是城市，虽然深度神经网络不知道北京、伦敦在什么地方，里面有什么建筑也都不知道，但是通过大量的学习出来城市的概念，并把他们放在空间中很相近的位置。我们并没有任何的语言和数据教它，是他通过大量的学习自己发现的。

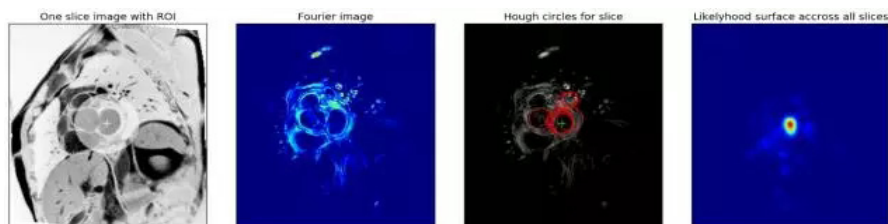
我们学习构建一个**深度的知识图谱**可以用在企业关系的挖掘，有一家上市公司，比如说是做锂电池的，可以找到他投融资的企业，并将上下游竞争关系全部联起来，这些企业之间会有信息的传递，如果网络构建足够大可以有一个模型，分析预测这个上市公司股价以后未来的走势。



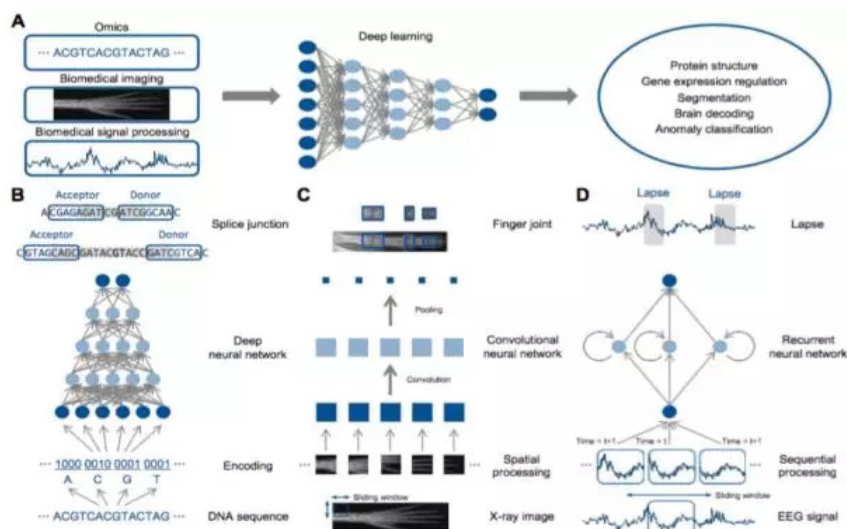
语音识别，百度最近有一个语音识别，叫双向循环神经网络 BDRNN，可以把每个音节都识别出来，发音中略微有一些口音和错误，也能够把大致的意思正确的识别出来。这个是语音识别的可视化的图像，我们把语音信号降维成一个平面的图，你会发现同一个元音音节和辅音音节在平面当中很相近的，都被抽象成有相邻关系的点，说明它真正理解了语音这个声频信号代表的含义。作为医疗诊断的重要根据，这是去年特别有名的心脏疾病诊断的比赛，当时参加比赛的最最后获得冠军的队伍，他们做到了准确度甚至超过了专家的水平。将几万张图片给深度学习的网络学习规律，其中正确答案是五位专家商讨得到的，但是计算机的水平超过了单个专家的诊断水平。

Deep Learning in Bioinformatics: 深度学习在生物医学领域，比如医学

图像处理、医学信号处理等有很好的应用基础



对 DNA 的解析，有很多遗传病是基因突变引起的，可能不是某一两个节点，可能是同时有几千个节点发生了问题，让人判断究竟怎么组合才会出问题是不可能了。这个时候深度学习可以来告诉我们，我们拿到一个人 DNA 之后，可以自动分析出来你在未来得某种疾病的几率有多少，可以提早的预防治疗。



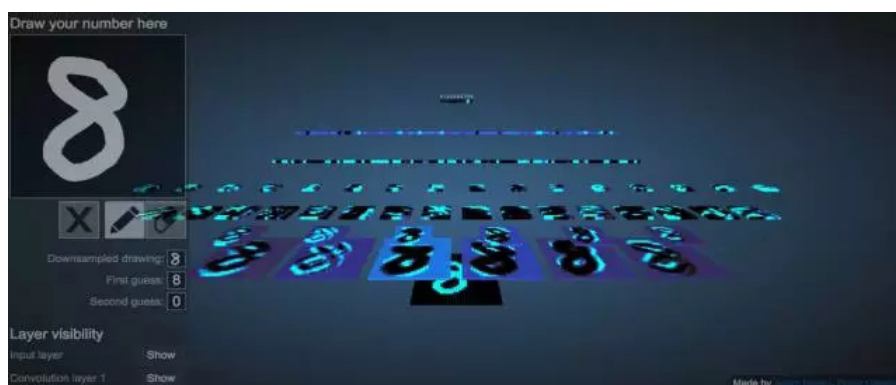
2.深度学习到底为什么这么厉害

深度学习是一个对特征不断抽象的过程，我们给他一个图片，深度神经网络首先提取出点和边，然后组合成人局部的器官，比如说一个眼睛和鼻子，局部的器官之后可以把拼接成一个个人脸，人脸外貌上有差异，我们用模版再匹配出最相似的就可以看看有没有人脸。

深度学习非常像人的学习过程，你必须一层一层的抽象才能理解更深的概念，之所以叫深度是有多层的学习网络，每一层是把特征抽象更高阶的概念，理解非常复杂的事物。

这是深度学习网络可视化的结果，我们给一个识别数字的神经网络一张数字

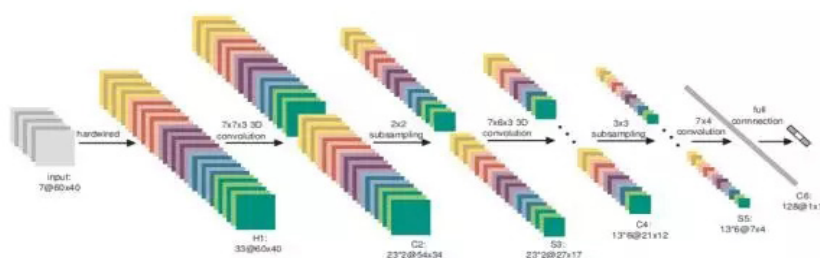
‘8’的图，可以清楚的看到每一层神经网络对原图进行了哪些特征变换。



这是一个深度学习常见的卷积结构，细节不讲了，大家可以感受一下，其中主要了 Convolution Layer, Max-Pooling Layer, 以及 ReLu Activation.

Auto-Encoder(Layer-wise Training), RBM, DBN

- PReLU, RReLU
- Dropout
- RNN, LSTM
- Max-Out
- Highway (Residual Net)
- Batch normalization
- Weight normalization



随着研究的不断深入，深度学习还有着各种各样的变种和组件，上面是一些最新的关于深度学习的研究成果。

3. 深度学习在实际项目中

我们讲讲深度学习在我们明略项目中的应用。我们有个很大的制造业客户，

他们有个故障预测的项目。我们能做什么呢？

深度学习除了建模的能力比普通的强一点，还可以学习时间序列的结构，设备传感器的数据是一个时间序列，每一秒钟或者多少毫秒产品信号，我们用传统的方法很难处理这么高纬度，这么大数据量的模型法国，深度学习可以理解在时间上的关系，大大提高我们对故障分类的预测。

另外一个就是在银行对不良客户检测的模型中，我们有数百维的储蓄、消费、信贷特征如果我们请专家来做非常困难，因为很多时候，当你的特征太多了，很难想到那么多规则的组合，用深度学习可以进行自动特征组合。

比如说发现我的银行的储蓄额很高，但是可能在月底突然取出来了，可能就代表着我可能只是临时在里面，跟别人借的钱放在里面，并不是我有这么高的资金做抵押，这个时候发现的时候，就可以排除在外，这个可能就超过了很多行业专家的工作效率了。

深度学习对我们的 DataInsight 是前沿重要的方向，我们会推出软硬一体的解决方案，我们也会使用 TESLA GPU 做深度学习加速器。

Q&A

Q1: 深度学习看上去很复杂，对于新手应该怎么去学，能不能推荐一些深度学习方面的书籍和流行的框架？

A1: 深度学习的数学理论其实并不复杂，但是需要注意的细节太多了，我们需要大量的时间来消化理解各个概念，对于新手，我推荐先上手代码，再研究原理，目前的深度学习框架非常之多，包括 TensorFlow, theano, lasagne, keras, sknn, no learn, caffe, mxnet, leaf, torch7, convnetjs 等。tensorflow 目前是 github 上最火的框架，我本人也是 TensorFlow 前 20 的 contributor，但是目前并不推荐新手直接使用 Tensorflow，新手最简单易用的框架是 keras，没有之一，可以先看上面的 tutorial 上手代码，理论方面，可以先学习 MIT 最新出的 Deep Learnign 的书来学习，这一本作者包括深度学习三巨头 Bengio

Q2: DL 在明略的具体应用场景，遇到的困难？

A2: DL 在明略的应用确实不是完全一帆风顺，我们在部署搭建 GPU 集群来训练 DL 时就需要非常多的坑，现在我们的产品 DataInsight 可以自动的帮我们解决这个问题，在数据上，我们需要收集比较大量的数据，才可以让 DL 发挥出强大的威力，在数据上，我们需要收集比较大量的数据，才可以让 DL 发挥出强大的威力，量级起码需要在十万条数据以上，最好是百万以上，然后就是参数调试的问题，不同的网络结构，激活函数，dropout，learning rate，参数实在太多，可能对初学者来说，感觉像是噩梦，然而随着对 DL 理论的理解，我们可以清晰的指导自己调参的思路，选择一条最佳的道路去解决这个问题，什么时候用 CNN，什么时候用 RNN，什么时候用 dropout，都是很有意思，并且很有内在道理的。能推荐几本书？或者资源吗？关于这个问题，我推荐去 github 上有一个项目叫 awesome deeplearning，上面有非常多的关于 DL 的资源

Q3: 请问现在对深度学习有一定理解，也有过相关经历用过 keras 等，请问怎么样更深入学习？

A3: 如果对基础代码熟练了，可以尝试去学习 DL 各个变种以及组件的原理，并在比较底层的框架，如 theano，tensorflow 上自己实现

Q4: 我是深度学习的硕士，请问现在有哪些工作还可以去做？

A4: 目前在 NLP 方向的研究还没有像 CV 和语音识别爆发起来，还存在很多灌水的空间

Q5: 内存和数据库以及 CPU 是否是深度学习的瓶颈？如果是则使用什么数据库能达到相应性能？

A5: 深度学习一般使用 GPU 训练，显存可能会成为瓶颈，数据库，cpu 一般不会成为瓶颈

Q6: 请问对于图像处理，是否用不到 RNN？这个问题很有意思

A6: RNN 用于有时间上前后关系的序列，以及需要记忆的网络，很有优势，可以说的是，在视频中，肯定是可以利用到的，还有刚才的图片问答的整体系统中，也是需要用到 RNN 的。

Q7: 请问深度学习在能应用在教育大数据分析中吗

A7: DL 基本可以用于任何数据集和问题，前提是有足够多的数据量，这

个是最重要的瓶颈，然而随着对 DL 理论的理解，我们可以清晰的指导自己调参的思路，选择一条最佳的道路去解决这个问题，可否稍微详细介绍下最佳道路？最重要的是网络结构，先设计好是用 CNN 还是 RNN，用多深的网络，用不用 maxpooling，接下来就是 Activation 函数的选择了，目前比较好的是 relu 系列的，包括 PRelu, Leaky Relu, RRelue, 另一个选择就是 maxout. 如果需要处理复杂问题，可以考虑 highway 或 resnet, 这两个可以训练非常深的网络，如果过拟合严重，可以考虑 dropout 或者 regularization。最后，再考虑 hidden layer unit number, learning rate 等等

Q8: 请问 spark 中那些可以在深度学习中应用呢？

A8: 目前有很多应用在 Spark 中的 DL 框架，包括 Elephas (依赖 Keras), SparkNet, CaffeOnSpark (依赖 Caffe)

Q9: 请问在图像处理领域，DL 还有哪方面的事情可以去研究？

A9: 目前比较火的一个方向是 图像生成，DCGAN，还有利用 DL 进行去噪，甚至是超分辨率，去模糊等方面的研究

Q10: 在 relu 等方法之后，一定程度上减轻了梯度扩散的问题，请问网络层数该如何选择（显然不是越多越好）？或者说，现阶段限制网络层数的主要原因有哪些？

A10: 网络层数过深目前依然会带来一些问题，比如训练过难，局部最优，鞍点等问题，还有就是也有过拟合的可能，如果问题确实非常复杂，比如图片分类，我们可以使用 resnet 或者 highway 来训练超深的网络，如果是简单一些的问题，我们可以考虑在深层一些的网络中加入类似 dropout, weight regularization 等减轻过拟合。

Q11: 我是新手，想问下深度学习需要哪些技术和理论基础？

A11: DL，首先需要在线性代数方面的基础知识，其次还有概率论，凸优化，基础数据挖掘概念等知识。

Q12: DL 在政务领域有哪些应用？

A12: 可以参见 MIT 的 Deep learning 的书籍，里面有很基础的讲解，可以简单入门。

Q13: 与 VR 有什么交叉点吗？

A13: 与 VR 关联可能不大，但是与 AR 关系很大，因为 AR 需要先识别出图像中的实体，再去进行 augmentation 增强，其中的识别需要用到 DL

Q14: 分布式环境上的深度学习，老师最推崇哪个？

A14: 分布式的深度学习框架，最推荐 tensorflow，在 100 台节点的 gpu 服务器集群上，tensorflow 的总体性能是单节点的 56 倍，也就是可以达到 56% 的分布式效率，远超其他框架

Q15: 先请问黄老师一个不正经的问题：有哪些任务目前为止的实验结果表明不适合 DL？

A15: 稀疏特征的数据上，DL 并不适用，比如适用 bag of words 再接 DL 效果一就很差，还有一个就是小数据量上 DL 可能会严重欠拟合

Q16: 话说对于一个大陆的学校的学生。如果本校没有什么出色老师研究这方面，怎么入门？

A16: 建议先研究代码，先研读 keras 的 tutorial 中的 sample code，然后再读 MIT 的 Deep Learning 这本书，最后读论文，通过 tensorflow 或者 theano 复现论文的方法。

Q17: 如何保证程序不被坏的数据教坏，比如垃圾数据，或者恶意构造的数据，以影响其本身的学习精度，就像微软的小冰，最近看到国外版本好像都要下线了，就是被人教坏了。

A17: 这个问题很有趣，目前确实没有什么特别好的办法解决这个问题，可能需要在网络结构上使用减轻过拟合的组件和策略，并且避免使用过大的 learning rate 和过小的 batch size。

Q18: 请问黄老师 在金融风控征信方式 dl 应用场景 能举些典型例子么

A18: 在政务和金融领域的应用，比如使用 DL 实现自动的征信打分模型，政府或者银行拥有超大规模的数据，完全可以满足 DL 的需求，加上 DL 自动组合特征的能力，是完全可以 在准确度上超过其他模型的。

明略技术合伙人徐安华： 数据量决定了特定领域自然语言处理最终效果

作者 刘羽飞

自然语言处理及文本挖掘技术的应用正变得更加广泛，尤其是在一些公共服务以及企业级应用方面的作用更加突出，比如执法机构需要用到的犯罪嫌疑分析，或者是企业决策用到的商业智能分析，以及普通人日常都需要用到的智能搜索功能等等，这些看似简单的应用背后，实际所需要的技术是比较复杂而专业的，因此为了更加深入地了解关于自然语言处理及文本挖掘技术发展情况相关的话题，InfoQ 专门采访了明略数据技术合伙人、SCOPA 产品搜索及自然语言处理组技术经理徐安华。

InfoQ：在进行自然语言处理的过程当中，会用到哪些工具？而这些工具又各自有什么样的特点？

徐安华：自然语言处理工具的使用，一般是由具体的自然语言处理过程来决定的。常见的处理过程主要就是进行分词处理，也就是将语句切分成几个有意义的词，接下来还要进行实体识别以及关系挖掘，有时可能还包括关键词提取、情

感分析分类等工作。

在这个过程中需要用到的，基本上就是跟这个过程相对应的那些工具，而这些工具基本上都在试图去覆盖这里面涉及所有的内容。一般来说这些工具可分为两大类，一类是来源于开源社区，比如中文分词组件 jieba，而另外一类是来源于国内的一些大学的研究成果，像复旦、哈工大等等，都提供了比较好的工具，尤其是哈工大，它在自然语言处理领域久负盛名。国外的话，做得比较好的开源工具有 NLTK 等，而公认的比较突出的国外大学，比如有斯坦福等，它提供的工具叫做斯坦福 NLP。

这些工具，实际都是在特定领域，或者是针对学术圈里面的一些特定情况下的应用，来去描述其自然语言处理效果的，因此在实际使用时，还是要根据自己的需求去增加定制化的东西。

至于这些工具各自的特点，可以说开源社区工具在覆盖的广度和可用性上肯定是要比学术圈内的工具做得更好，比较典型的比如刚才提到 NLTK，它其实已经面向整个文本挖掘里面所涉及到的方方面面内容，都提供了相关的工具。

然而要想进一步提升自然语言处理的精准度或者是精确的解决企业的业务问题，还是需要在这些现有的工具基础上，加入一些自己定制开发的东西，最后形成自己的一套自然语言处理工具。

InfoQ：能否谈谈自然语言处理及文本挖掘在技术层面上的难点有哪些吗？而对于企业级用户来说，又该怎样做才能克服这些困难？

徐安华：自然语言处理与文本挖掘比较难的一点就是准确率问题。准确率直接影响着自然语言处理技术在产品化运用过程中的最终效果。现在很多在学术界准确率较高的自然语言处理技术其实在产品化之后，用户体验都不太好，依然会感觉到人可以精确识别的部分，机器识别起来却还是非常困难。

从技术上来说，解决方法可能就是需要非常大量的人工来训练相应的模型，当然这里指的就是叫做基于监督的模型，基于监督的模型在特定领域的效果非常好，但它的缺点在于可能换一类文本，准确率就会下降。因此现在很多人都试图去寻找完全无监督的，可以不用人工去进行标注的这种方式去训练模型，通过机器自己的学习，也能够正常进行这种分词处理，或者是进行命名实体识别。

具体问题还是需要做具体分析。现在来看，并没有一种放之四海而皆准的方法，即使是基于深度学习的技术，其实也是在特定领域之内才可能会有实际用处，它的实际效果，其实跟现在主流的这种基于监督的方法而得到的模型的效果是差不多的，甚至有些情况下还不如后者。

而对于企业级用户来说，要想克服这些困难，一方面还是需要找到一些专业的人才，具体的问题还是应该在具体的应用场景下进行分析，除了之前提到的工具之外，更重要的是只有这种具有专业知识的人才能够有经验和能力去更好的解决这些问题，而想要完全依靠工具去解决则是非常不现实的。

InfoQ 国内目前在自然语言处理以及文本挖掘技术方面的发展情况怎么样？目前是否面临着一些普遍存在但是又应该去解决的问题呢？

徐安华：国内自然语言处理以及文本挖掘的应用，我认为做得比较好的是百度、腾讯等，还有一些其他的拥有海量非结构化数据的企业，正是因为他们有大量的数据，因此才在建立模型、固化模式等方面比其他数据量相对比较少企业更有经验，他们的数据就是他们最大的优势。

至于说到算法，其实这些年来并没有出现突破性的进展，大部分情况下各家比拼的还是数据量，数据量的大小决定了很多算法最终达到的效果，包括百度的语音识别技术，即使是利用了深度学习技术，但要想实现这么高的准确率，必须是建立在拥有海量的语音数据的基础之上才可以。

普遍存在的问题依然是准确率。要想做到一个全行业都能非常精准的识别自然语言，或者是非结构化数据里面提到的一些信息，也都是非常困难的。我们认为只有在垂直的领域下，才能根据很多的语言习惯和语言模型，将自然语言及文本挖掘技术进行产品化。但对于全行业都通用的高准确率自然语言处理解决方案，至少目前来看的话，现有的这些手段都没有办法解决这一问题。

InfoQ：对于明略数据目前的业务来说，哪些领域或行业的用户在自然语言处理以及文本挖掘方面的需求比较大，他们的业务有什么样的特点？

徐安华：公安用户其实在自然语言处理还有文本挖掘上需求量比较大。原因在于这个特定领域积累了大量的历史数据，这些数据中记录了很多相关的事件或案件描述，这些描述中包含很多关键的要素，而这其中的问题就在于如何把新加

进来的描述，跟已经存在库里的一些事件信息去进行关联分析，而文本挖掘就是解决这一问题的关键技术之一。

另外一个问题在于搜索，这个需求是在同时进行很多的事件描述的录入时，就同步进行关联分析或查找，而普通的搜索手段就无法有效的解决这种海量的实时数据录入问题了，再加上在这个过程中还要跟已有的库中的信息做关联，这时依然需要利用自然语言处理及文本挖掘技术来解决问题。

InfoQ：那么比较成功的案例是否也是在公安部门的应用当中？是否可以介绍一个比较成功的自然语言处理技术应用案例？

徐安华：自然语言处理比较成功的典型案例依然是在公安部门中的应用，也就是明略对于公安部门内部的一些事件描述文本的处理。利用这个领域之内长期积累而建立的模型，我们可以把一段事件描述里提到的一些关键要素拆解出来，准确率也越来越高。这样的话就更便于我们进行后续的文本关联分析，以补充和完善现有的信息库。

另外一方面，则是数据挖掘的效果最后需要将它进行产品化，也就是需要跟搜索引擎相关的技术做结合，将文本关联分析功能做成一个应用，形成一个可进行实时处理的具体产品。

InfoQ：今年在自然语言处理以及文本挖掘技术方面可能会出现哪些新趋势？

徐安华：目前比较受关注的趋势基本上在于两方面。首先是知识图谱的应用正变得更加广泛。之前对知识图谱的应用更多的是停留在大企业中，用来解决一些搜索问题。但是现在因为有更多相关的开源项目被公布出来，再加上也有一些企业将知识图谱构建技术公布了出来，这就使得更多的企业级客户，以及公安的客户来说去使用这种技术，来提高它现有的这种文本挖掘，以及这个文本处理的准确率，以及这个更加人性化的去识别一些这个相关的这个要素，其实会是一个趋势。

第二个趋势是深度学习，深度学习在图片挖掘领域获得成功之后，迅速朝向文本挖掘领域拓展，谷歌、Facebook 等公司都已经公布出来一些自己专门用来进行自然语言处理的深度学习框架以及相关的一些产品，甚至有些深度学习的产

品，已经在朝 iOS 或者是安卓系统设备平移，这样一来可能后续的自然语言处理以及文本挖掘技术将不再停留在服务器端，在说移动端或者其他更小的便携设备上，也可以去进行自然语言处理，或者说能够受益于自然语言处理技术，甚至是享受到更宽泛意义上的人工智能的优势。

徐安华，现为明略数据技术合伙人、SCOPA 产品搜索及自然语言处理组技术经理。2004 级北大计算机本科，2008 级北大计算机系统结构方向硕士，曾经在 Intel、爱奇艺工作 4 年时间，Linux kernel 代码贡献者，显卡虚拟化项目 XenGT 早期主要开发人员，拥有多篇专利。2014 年底加入明略数据，致力于在 Hive、Impala、SparkSQL 上实现行列级别权限，目前专注于自然语言处理与文本挖掘。

高手支招： 特定领域的实体关系如何提取

作者 徐安华

1. 公安领域的工作背景

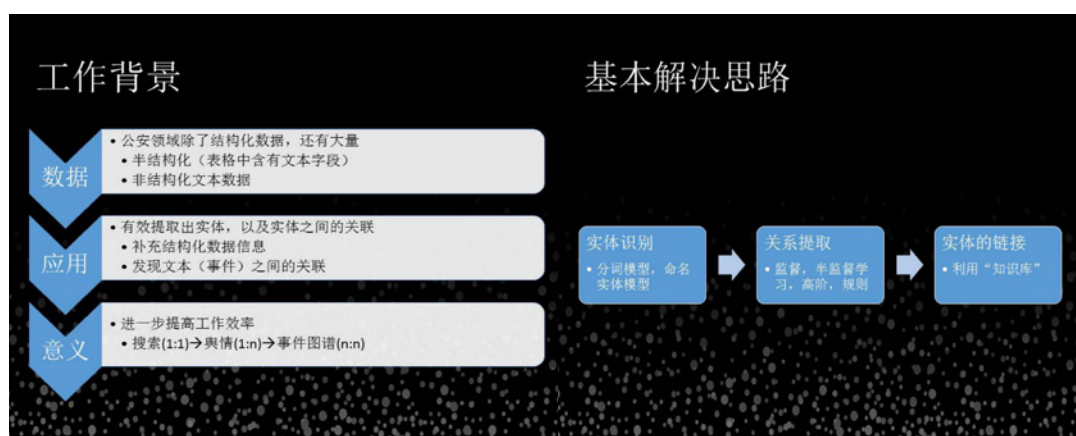
从特定类型文本中提取实体关系。这些领域除了有很多结构化数据之外，还有跟多的文本数据，通俗意义上都叫做非结构化数据（这里不包括语音、图片、视频等）。

在应用里面，结构化描述的数据是非常清楚的，对于文本来说，由于大家书写的形式各异，表达方式多样，这个里面提到的很多要素，如人名、车牌号、手机号、组织名等都是称之为实体。

实体和实体的关系，以及实体和事件的关系，是比较难发现之间的关联，说明了前两点之后再去看第三点，意义在哪里？

其实随着文本数据，文本数据的信息被大规模的被发挥出来，最早来自于互联网公司，比如说百度、Google 以及基于此创造很大的价值，随着互联网公司的兴起，大家对非结构化数字文化挖掘越来越重视。

这个搜索 1: 1 是什么意思，第一，就是一个人找一件事情，第二，1:n 是舆情，即使你做营销也好，你都需要知道统计的信息，舆情是一对多的情况，第三就是我们需要关注的所有的事情其实都是并发的，各种各样的发展的趋势都是同步进行的，你不可能用一个人来监测这些变化，我们也没有这么大的团队做这样的事



情，需要机器做事情，N: N就是用机器发现追踪非常多的线索，每条线索都涉及到很多的实体。

2. 有效地找到特定类型文本这种蕴含的实体关系

如何找到文本里面的实体关系。这是一个工业界也好，学术界都在持续做的事情，在这里列出来这些步骤，其实主要是想说我们做这个事情，从最开始到逐渐去完善，到最后去结合实际，把事情做到产品级别；

第一步，因为我们是处理中文，分词是必须的，也是最关键的第一步。第二步就是分词之后才会出现所谓的实体，这个实体就是一个字符串，什么样的字符串才是实体，这个我慢慢跟大家分享。

第二步就是关系的提取，这块虽然有各种深度学习的方法，但我们发现是统计的思路依然非常好用。但有一个问题，对于模型训练问题，我们需要巨大的人工。举例来说，你告诉机器 1+1 等于 2 一万次，这样下次它遇到的同等情况就会以较大概率猜到 1+1 等于 2。

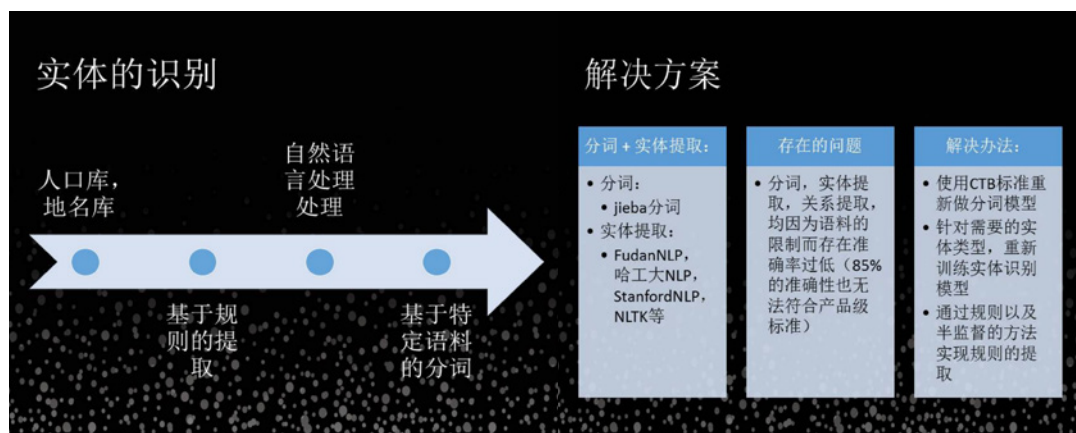
关系不只限于二元，你和你的亲戚、父母都是二元关系，还有多元关系，我和我的朋友一起去旅游，假如说我们一共有四个人，分别叫张三、李四、王五、刘七，那这个就是所谓的高阶关系。为了处理这些关系，我们除了基于统计，还可以利用一些规则的东西。

所有这些做完之后，我们得到了很多的事实，如何把文本本身提取到的实体进一步完善，确认这个实体，这个人名，我怎么确定这个人就是我想找的那个人，而不是重名的人，我们就需要利用知识库来做一些事情。

3. 把文本本身提取到的实体进一步完善

第一步就是实体的识别，在我们实际的处理过程中有一些主要尝试的思路，首先是最基本的字符串匹配，这件事做好的前提是我们原先知道足够多的人名、地名等实体名字，但是我们是没办法拿到非常全的各种各样的库的数据。那么我们就需要想办法，找规律。那么我们就需要想办法，找规律。

第二步就是用所谓的规则，可能大家都知道正则表达式。



上面两种思路都不好使的前提下，必须用自然语言处理的技术。市面上很多的工具，包括公布出来的斯坦福的工具，北京大学的工具，但这是这些工具中的语言模型很多都是从新闻媒体语料训练二来的。新闻媒体的提及的语言文字，范围广，涉猎多，所以处理的广度是有的，但是当应用到一个特定的领域之后，这个精度是远远不够的（尤其当这里的很多用语独居特色，和新闻语料有明显区分度）。作为一个业务人员使用这些已有的工具需要找到你所需要知识的时候是不够的。

这里提到的就是常见的市面上的工具，主要存在的问题，我在前面说得比较多了，解决办法就是我们希望用到更多更细的方法：为我们特定领域的文本重新训练的分词模型，重新训练识别模型。根据前两步已经做出来的模型，通过规则引擎以及半监督的方法来实现关系的提取。

4. 关系提取的方法

关系的提取有很多方法，并且学术界也一直在研究，但是当你遇到特定问题

的时候，我们不得不针对问题重新思考。

首先第一想到的就是基于监督的方法，不过你需要不少的人工的工作量在里面，人工告诉机器什么样的东西才是你所需要提取的关系，这个也是教机器很多次，希望它在后续的过程当中能够猜出来你希望的结果。它的优点就是针对领域，识别率比较高，我们在做的时候也用到了这种方法。缺点也很明显，如果更换了应用领域，原来的模型就不太适用了。

但是我们知道在这个过程中，人也是会累的，人是会犯错的，每个人的知识、背景都不一样，没有办法高精度确的教会机器。



下面介绍一种更好的方法，就是半监督的方法。

半监督的方法，思想非常好，比如说先说张三是张仁的儿子，在这句话中蕴含了一个模式，就是 xx 是 yy 的儿子，简而言之就是“父子关系”的模式，我可以把这个模式推广放到更广阔的范围，可以在互联网里面查找“xx 是 yy 的儿子”这样的形式，这个过程当中，除了初始化的种子里面提到这种模式之外，还可以通过新发现的关系来查找到更多的模式。再把更多模式里面学到的东西，不断地滚雪球，能够得到更多的关系。但是做这个事情最重要的前提问题域本身是有这么多的各种各样丰富的模式，不止是单一的模式，如果只有单一的模式，你希望从单一的模式扩展到所有的模式都是不可能的。

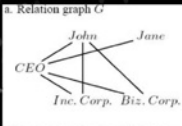
接下来是高阶关系，这个是我从别人论文取得里面的截图，举例说明高阶关系。高阶关系其实我们遇到的不是很多，高阶关系处理的还是比较复杂的，不能依赖于单一的简单的解决方法，论文中提到的方法并不实用，现在高阶也会规则

引擎去做识别。

这里举了一个使用规则提取关系的例子，这个规则看着非常复杂的东西，其实说明的是亲属关系，提到两个人，再加上两个人之间的关键词，我们这个规则可以识别出来这种模式。

关系提取：高阶关系

- 具体办法
 - 讲高阶关系分解为二元关系的集合
 - 构造实体关系图
 - 构建最大团



a. Relation graph G

b. Tuples from G

$(John, CEO, \perp)$
$(John, \perp, Inc. Corp.)$
$(John, \perp, Biz. Corp.)$
$(Jane, CEO, \perp)$
$(\perp, CEO, Inc. Corp.)$
$(\perp, CEO, Biz. Corp.)$
$(John, CEO, Inc. Corp.)$
$(John, CEO, Biz. Corp.)$

关系提取：基于规则

- $\$FAMILY = \text{"母亲|妈\{1,2\}|其母|父亲|其父|爸\{1,2\}|妻子|丈夫|哥\{1,2\}|弟\{1,2\}|嫂子|大嫂|姐\{1,2\}|妹\{1,2\}|儿子|女儿|儿媳|媳妇|女婿|孙子|孙女|祖父|爷爷|奶奶|祖母|外公|外婆|舅\{1,2\}|外甥|姨父?|侄子|侄女"};$
- $pattern: \{ (? \$prev [\{ ner:PERSON \}]) (? : [\{ 0, 10 \}]) (? \$relation \$FAMILY) (? : [\{ 0, 1 \}]) (? \$next [\{ ner:PERSON \}]) \}$

假设我们前面这些问题都已经解决了，我们已经利用文本中原有的信息解决了上述问题，但是把文本里面的信息挖掘的非常充分后，会发现涉及到很多人的重名。你需要确认两个事件人都是同一个人，这件事情需要上下文，还需要外部的知识库辅助完成这个事情。一方面需人口库，甚至地图库，你甚至需要确认说提到的地点在地理上确实是相近的。

5. 实体的链接

最好的知识库是什么？除了处理大量的非结构化数据之外，还有半结构化的数据，半结构化数据是天然的知识库，比如说你一行数据里面，有很多的名字、身份证号、手机号，后面还会加描述，事件的起因、结果等，文本里面提到的东西可以在自己所在的结构化字段行里找到更多的对应关系。

6. 一个文本处理工具

这个其实是我们自己研发的一款对于文本处理的工具，这个图说的是什么事情呢？大家可以看到最左边，左边是非常多的文字的描述，描述很多的事件，事件里面会有人、车，手机号，还会有实体关系的关键词的描述。

实体的链接

基本方法：文本上下文 + “知识库”

- “知识库”
- 结构化数据字段
- 人口库信息
- 地图库信息
- ...
- 考虑消歧

最好的“知识库”

- 半结构化数据
- 天然消歧

成果演示1



通过这些实体和标识实体关系的挖掘，最后可以把文本里面描述的事件，最后变成一种图的结构化形式的展示。可以看到右上角这个人是有电话号码，还拥有一辆车，还跟另外一个人有亲属关系，这个人又有一个电话号码，整个文字里面提到的故事情节都可以被结构化称实体之间的关联，这个可以非常方便的让我们的相关的业务人员能够看到这个文字里面到底描述的是什么。

一篇文本不算什么，但是一万篇文本里面的提到的人和事件做关联的话，很多事情就很好的分析了。这个产品其实是在整个的 SCOPA 里面的组成部分，在非结构化数据里面帮助我们进一步完善整个关系的挖掘，包括我们能够全方位，更多的把更加全面的关系挖掘出来。我今天的演讲就到这里，谢谢大家！

Q&A

Q1: 文本的实体分析目前我们明略识别的准确率是多少？

A1: 利用纯粹的自然语言处理技术，人名的识别率能够达到 95%，地名也在 90% 以上，但国内某些小城市的名字比较有特色，如“任城镇”这类，就不容易识别对，如果没有地理信息库的话。

Q2: 感觉明略的大数据专家团队女性很少很少，女性在从事这方面工作中有什么样的明显劣势？

A2: 女生确实偏少，但我相信后续会越来越多。我们的工作氛围和互联网公司非常像，扁平管理，所有人都能学到和了解到公司的方方面面。胆大心细的女生其实会有更大的机会。

Q3: 文本的实体关系分析有成熟的方法论吗？感觉还是在就事论事的阶段。

我们都是到 plantir 在 connect 方面做的很牛，有可借鉴的地方吗？

A3: 我们有自己的 NLP 开发 SDK，但是我们不会把所有问题都使用同样的 pipeline 来处理。我们非常在意准确率。Palantir 的方法其实大家都在猜测，但我们认为足够多的数据以及优秀的人才才是真正的核心。

Q4: 实体的识别是根据分词后的词性得到的吗？你提到分词模型要重新训练，所用的训练数据源是怎么得到的，是靠专家的标记么？

A4: 实体的识别可以把词性加到特征里面，还可以加入更多的特征。我们所使用的数据源来自于客户，但是并不需要专家来标记。

Q5: 老师，我想向大数据发展，怎么开始？

A5: 最好的建议是把现有的工作做好，提升内功，然后找机会进入一家大数据公司，比如明略数据。

明略技术合伙人杨威： MDP 打造新一代高性能、高可用、高安全大数据平台

作者 孟夕

InfoQ: 这次特意邀请了明略数据技术合伙人及 MDP 产品经理杨威先生，一起来聊一聊关于企业大数据处理平台的相关话题。杨威先生您好！首先想请您介绍一下自己，谈一谈之前的一些工作经历，以及现在所从事的主要工作。

杨威: 各位好，我是杨威，来自明略数据。在加入明略数据之前我一直在秒针系统工作，期间参与了秒针的广告监测平台、交易平台以及投放平台的建设，后来主导了秒针大数据平台的建设。从 2014 年加入明略数据之后，参与了一些大型金融企业的大数据平台建设，以及 Hadoop 数据库相关的建设，之后也参与了一些公安系统以及电信领域的一些大数据平台的建设工作，而现在则主要在明略负责大数据平台产品的研发工作。

InfoQ: 您职业生涯的大部分时间里都是在跟大数据平台打交道，那么能否分享一下在之前从事产品设计工作时，印象比较深刻的一件事？

杨威: 我参与比较多的确实还是跟大数据平台技术相关的工作，印象比较深

刻的应该是在秒针系统介入到大数据平台业务的这个阶段中所做的一些工作，因为这个阶段相当于大数据平台产品的研发雏形阶段。

建立秒针大数据平台的原因，其实是因为企业对于大数据平台的一种典型需求——首先要面对的就是海量数据，因为每天都有上百亿的曝光数据需要处理；其次还要面临海量的运算需求，因为每天都需要统计所有的广告曝光、PV、UV等数据，此外还有其他各种各样运算类型的需求，比如实时计算、反作弊广告监测等，这其中既有实时的计算系统，同时也有离线的批量计算系统以及数据挖掘系统，另外还有用户画像、用户模型的建立等等。

除此之外，秒针大数据平台还处在多租户的环境下，比如秒针内部本身就有多个部门在使用该平台，大家要在这个平台上共享计算资源，并且要共同使用这个平台上的数据，而这就会产生很多权限和资源方面的管理需求，接着这些需求其实在后来明略的产品设计过程当中，都会被抽象成一个企业级大数据平台产品的需求，然后在整个秒针大数据平台的建设过程当中，我们自然而然的就积累很多这方面的经验，最后变成我们产品的一些功能的实现，然后能够应用在现在的很多大数据平台建设中。

InfoQ: 在明略数据工作的两年时间里，您感受到了一种什么样的技术团队的文化？这种文化氛围对您产生了什么影响？

杨威: 明略其实是一家非常新的技术公司，马上就要到它的两周年生日了。明略的技术氛围是一种既温和又热烈的技术氛围。温和其实是体现在这种人与人之间的关系上，包括上下级之间的关系。无论是CTO还是CEO，在平常很多工作生活当中都可以直接与他们接触和沟通，他们也会经常与我们沟通公司的一些发展情况、行业发展情况等等，大家并没有一种上下级的隔阂关系，反而是一种非常融洽的共同奋斗的关系。

而热烈则体现在大家对技术的追求上，明略虽然是很新的技术公司，但是在开源领域投入得非常多。作为初创型公司，这非常难得。明略在很多开源社区里都有代码提交者，除了在完成自身的工作之外，公司也将很多资源投入到了开源技术方面，这是在很多其他公司都很难看到的情况。

这使得明略的整个技术氛围都是非常先进、前卫的。比如针对前段时间的

AlphaGo 围棋赛，明略内部就曾安排了深度学习领域的大牛来给大家整体介绍 AlphaGo 是利用什么机制来实现的人工智能，以及战胜世界顶级围棋选手的原因是什么。在这样的氛围内，大家就可以很轻松地接触到当前大数据或人工智能领域最先进的技术或者是一些新的技术理念，这对大家的提升是非常有帮助的。

InfoQ: 去年 10 月份的时候明略数据公布了最新的大数据平台产品 MDP，那么能否请您介绍一下 MDP？

杨威：谈到大数据平台，可能需要先说一下对企业 IT 系统的理解。企业 IT 系统一般分为两种类型，一种是交易型的系统，与企业业务相关，比如说银行交易系统或者电商交易系统，一般是建立在关系型数据库上的业务系统，这一类系统的系统特点是不同的业务部门都有自己的一套系统，比如财务、ERP 等等都属于这样的系统；另外一类 IT 系统，比如以前有些企业会做自己的分析平台，或者是自己的数据仓库，这就是分析型的平台系统。

那么随着大数据平台相关技术的完善，企业其实完全可以构建统一的分析型平台，这个分析型平台能够介入企业现有的所有业务系统的数据。这种交易系统里面的数据以及外部的数据资源，可以在统一的平台上面进行数据分析、数据挖掘，产生新的数据价值。

这种平台的建设就可以称之为企业的大数据平台，那么 MDP 其实就是一个大数据平台的技术的实现，可以帮助企业建设自己的大数据平台，帮助企业接入业务系统里面的各种各样的数据，能够在上面支撑很多的不同的数据应用，另外这还是一个分布式的、可扩展的平台，不像上一代的数据仓库技术受限于技术框架限制，可扩展性非常有限，而现在的大数据平台的扩展性可以达到成百上千台的机器，计算资源其实是非常庞大的。在今后建设大数据平台过程中，一家企业只需要一个这样的平台就可以承担所有的分析性的计算应用。

InfoQ: 明略数据的上一代数据平台产品叫做 BDP，那么能否请您谈一谈 MDP 和 BDP 的区别？而 MDP 是否在数据存储、高性能计算框架、以及安全防护方面又有所加强呢？

杨威：MDP 这一代产品的主要增强都体现在了企业级需求上面，比如说高性能、高可靠、安全以及易用性，这些方面都有了很大提升，尤其在高性能方面涉

及的比较多。首先来说，运用了几种不同的计算手段，比如实时 SQL 分析的手段，引入了最新的 Impala 2.3 版本，计算性能非常好，MPP 计算架构可以并行处理整个平台上面的数据，进行 SQL 查询分析。而内存方面则引入了 Spark 1.6 的版本，这也是 Spark 最新一代的引擎，其内存迭代计算，以及新的 DataFrame 接口能够非常方便的进行数据处理，使得在 MDP 在内存计算方面上了一个新的台阶。

另一个企业级的非常重要的需求就是高可用，这方面明略也做了非常多的工作，新的 MDP 提供了一套从网络层面到系统设置再到整个产品内部的这种服务进程，以及包括产品内部的源数据，就是说在这种情况下任何一个网口、一台交换机，或者一台服务器，甚至一块硬盘、一个进程坏掉，都不会影响整个平台服务的正常运行，MDP 能够以 7×24 小时、5 个 9 的方式保证企业的服务正常运转。

还有一个就是安全方面，企业的数据安全也是明略数据非常关注的一个点，因此 MDP 从主机安全、网络安全、服务安全和数据安全，四个角度来构建了平台级的企业级数据安全保障的。

InfoQ: 在去年明略产品发布会上，另外一个提的较多的话题就是数据关系挖掘，是否可以请您谈一谈如何从大数据平台建设的角度，帮助企业级用户做好数据关系挖掘？

杨威: 数据关系挖掘其实就是接入企业各种各样不同的数据源，然后把这些数据在一起做关联挖掘他们之间的关系，这个过程其实涉及到几个方面，首先，有能力接入非常多的数据源，比如可以支持这种传统的关联数据库的接入，也能够支持一些非结构化的数据的接入，比如网页数据、音频、视频数据的接入等等，在接入这些数据之后，它要将这些数据进行关联分析，就要提供很多关联分析的存储和计算的手段，那么关系的存储可能我们会提供一些图数据库来支持关联分析应用，能够在这个过程当中存储它的实体之间的关系，分析型的话我们前面提到会提供像 Spark 这样的内存迭代式引擎来帮助关系的应用，利用这个引擎来引入一些关联分析，还有一点就是，本身来说这些数据也是关系到整个企业的非常多个业务的核心数据，在数据安全性的保障方面也是要做非常多的工作，我们要提供非常细力度的这种数据的权限控制手段来保证敏感数据不被一些无关的用户所看到。

最后还有一点就是关联分析它的整个平台资源消耗也是非常高的，明略需要提供的一个非常好的资源管控和隔离手段来保证关联分析应用的资源不被其他的应用所抢断，明略是通过 Cgroup 来控制来给关联分析应用的资源分配的。那么，主要是从这几方面，一个是数据的接入，第二是数据的关联数据的存储以及关联分析的手段的支持，最终从数据安全和条线隔离、资源隔离几方面对关联分析提供支撑。

InfoQ: 今年大数据领域又出现了一些新的热点，比如说数据湖、人工智能，还有像刚才说的阿尔发狗，那么您认为今年在企业级的数据安全技术方面又出现了哪些新的变化呢？

杨威: 今年企业技术数据安全有两方面变化，一个是从传统的数据安全保护的手段来看，它是从离线这种分析转向实时预警的方式，从大数据平台中来看就是以前可能大家都尽可能事前做一些规则的限制，限制某些用户对某些数据的访问权限，通过这种方式来限制。当然以前也有审计的手段，审计的时候记录用户的数据访问限制，但是现在会有实时的对这种访问限制的分析和预警，而且这个预警可能是一种不是基于规则，而是基于特征发现的基于数据挖掘手段来进行。

典型的就像比如现在有一个项目叫 Apache Eagle，这是 eBay 研发出来的一个项目，它能够分析整个大数据平台上面用户对数据的访问，能够实时的监控对敏感数据的访问，以及就是越过规则的一些访问操作，另外它还是能够通过对用户行为的分析建立这种用户行为的特征值，然后通过这种异常发生在异常空间里面异常访问行为空间里面的一些用户行为对它发出报警，这就是一些更快速然后也更智能的对数据安全的保护，另一方面其实是体现到受到这个数据安全现在比较火的这种数据交易。

企业在数据交易过程当中怎样保证数据的安全，这里面其实也有几个方面的事了，一个是在数据交易过程当中怎样保护用户的隐私不被泄露，那么这可能涉及到用户数据脱敏的一些工作，第二个事情就是在数据交易的过程当中怎样保证数据双方的交易双方的数据价值不被窃取或者丢失，因为数据本身是一个可复制的、可 copy 的一个东西，如果说买方买走了数据就随意的应用，那么对卖方是一个很大的打击，另外一方面就是说如果卖方随意把数据以很低的价格随便卖，

那对买方也是一个很不能接受的事情，因为这个数据还是有唯一性和稀缺性这种特点的，所以要保证这方面特点在平台方面也要做很多工作。

我们要从这些方面保证数据安全或者保证数据价值的话，其实都是只能从技术手段解决，技术手段我们其实是在交易双方都做了限制，其实双方都是用的我们这个 MDP 平台，无论是这个数据的输出和数据的接入都是由 MDP 来负责做技术上的处理的，从这种手段上面来保证双方的价值都不会被伤害。

InfoQ: 能否具体谈谈如何从网络、服务、运维这几个层面来把控企业数据安全？

杨威: 网络层面以前做的比较多，一般来说就是网络隔离、网口的限制等等，这个通常企业里面已经做的比较完善了。服务方面，我们主要说的是平台内部的服务，它其实是包含了几个层次，首先平台内部服务之间的数据交互是有一个传输的渠道的，这个渠道我们是做了加密限制的，第二是说服务平台，服务对外部提供数据访问接口的时候，这些接口都是提供加密的手段，第三个就是服务本身的使用权限限制，不同用户有不同的服务使用权限限制，以及每个服务内部可能有自己的不同的使用的服务，就是不同级别或者不同力度的这种控制，比如说有的人可以提交查询，有的人可以杀死别人的查询，有的人可以申请更多的资源，有的人可能就是只能用自己的资源，这些非常细力度的控制我们也提供一个接口来控制，还有的就是服务的上层，比如输出数据这个地方，我们也提供数据脱敏的手段来避免敏感数据泄露出去。

另外从大数据平台角度来看，整个平台运维也是很重要一点，区别于以往的这种像关系型数据库里面的运维决策，以前关系型数据库里面非常关键一个角色叫 DBA，它其实是一个系统的最高权限管理员，最高权限人员拥有所有的数据访问权限以及所有的相关的运维操作的权限，在明略数据平台里面其实是把这样一个角色分成两部分，一部分是运维的角色，另外一部分是权限管理角色，运维角色只有对平台的运维权限可以去服务、去操作配置文件、看日志等这些操作，但是其实是访问不了这里面数据的，而另一个权限管理员，可以给每个帐号分配你可以看到哪些数据的权限，两个角色拆分来看，保证了运维人员权限不会像以前在关系型数据那么高，没有数据权限，从而保证数据的安全。

InfoQ: 您刚才也提到数据脱敏，那么是否可以详细谈谈数据脱敏应用场景以及里面涉及到哪些技术关键点？

杨威：数据脱敏应用场景以前也存在，就是当你需要把一份数据交给第三方的时候，其中肯定会有很多敏感数据的，比如有时候可能会存在个人身份信息，这样的数据通常来说是企业不能把它转交给第三方的，但是又有很多时候你要转给其他的企业数据，你不得不面临你需要暴露出一些这种类似的数据，或者需要有一些特征，在这个时候就需要用到数据脱敏这种应用。

脱敏这种手段其实有很多种，最简单一种就是遮掩，我们把关键的链给它 mark 掉，用特殊符号代替或者怎样，另一种手段就是扰动，我们在里面做随机的扰动，使得数据变得不真实，这些都是一些比较简单的手段，这些手段确实能够避免隐私数据的丢失，但是它同时丢失了数据本身的统计特征，因为扰动是随机的或者遮掩是随机的，那么它原来应该带有一定的数据统计特征，比如北京地区的总收入应该在什么样的一个水平，上海地区的平均收入应该在什么样的水平，一旦随机做了扰动之后统计特征就丢失了，对于数据使用方来说数据价值就降低了。

那么在这些简单的方法之外还有很多更高级的方法，它能够使数据在数据脱敏过程当中不会丢失，这其实有一些像我们自己的系统里面已经实现这样一些类似的方法，能够保证数据在脱敏过程当中这些数据统计价值不会丢失，能保留数据的统计意义。

InfoQ: 那么从大数据平台应用这个角度来看，目前传统行业以及政府部门主要集中在哪些地方应用情况比较多？

杨威：无论是政府或者企业，有两类部门的应用对大数据平台建设需求比较多，一类部门就是本身就是自己的原来的核心的业务部门，比如说在一些焦点的业务部门，比如说在政府领域，像交通行业，交通部门就是非常典型的交警的行业，因为大家对交通问题非常关心，在交通行业应用大数据平台非常多，他们的大数据平台建设走在前面的行列，比如说在银行领域就是这种风控部门，风控部门是现代部门，因为现在业务面临很大压力，而且这些部门通常 IT 技术走的非常先进，所以他们也是一个使用大数据技术处在前列的一个位置。

另一方面就是企业的 IT 部门，就是说除了业务部门主动发起大数据平台建设之外，IT 部门也会考虑主动建设大数据平台，因为前面我提到就是在一个企业里面通常最终只会有一个大数据的分析型的平台，这种分析型平台管理和维护通常由 IT 部门统一进行，所以 IT 部门统一建个大数据平台也是非常常见的。

InfoQ: 您能否介绍一个比较成功的 MDP 平台的应用案例么？

杨威: 可以说一个电信行业的应用案例。现在大家都有手机，手机可以上网可以打电话，那么它其实在不上网不打电话的时候会发出去很多数据，其实就是说现在手机跟电信公司的通信都是会产生很多数据的，而明略建设的这个平台就是所有手机通信数据的一个存储和查询平台。

比如说一个省的数据，大概每天有数十 T 的数据进入这个平台，那么其需求就是要很快能够查到一个人所有的通信的数据，以及在任何一个时间段和时间点都能够分析出其通讯信号的好坏，或者详细的话费情况、流量情况等等。

在这样的背景下，另外还有一个要求就是高可用，因为是电信领域的在线业务系统，它是前端，是面对很多这种客服人员的查询需求的，那么它就不允许出现任何一分钟的业务中断。这就对高可用提出很高的要求，而明略则从网络层面到系统层面到明略自己内部的服务层面都提供了各种高可用的支持，以保证任何一个节点的故障都不会导致整个平台服务的不可用，而且最终这个平台也体现了非常好的性能，其在任意查询一个个人账户在半年之内任何时间段的通话记录可以在秒级的时间内返回结果。

InfoQ: 您认为今年的大数据平台建设方面出现了哪些新的趋势吗？

杨威: 其实也不是今年，从去年或者甚至前年就开始了，一个比较明显的趋势就是企业的 IT 数据越来越多的搬到云上面去，他们也会有非常强烈的需求把大数据平台建设在云上面，从国外大数据平台公司比如 Cloudera，以及 Hortonworks，他们都已经提前布局把大数据平台部署到云上面，我们其实也在和云厂商合作，然后把明略自己的 MDP 部署到这些不同的这种云的公有云上面去，以使用户能够非常快速的在云上面实现平台的部署。

明略技术合伙人任鑫琦： 数据关系挖掘算法、技术难点及应用场景分析

作者 刘羽飞

数据孤岛、零散数据等现象一直是企业大数据应用过程中所常见的问题，当数据以及数据来源增加过快时，不同数据之间的打通就成了最大的困难，有时这对于传统企业来说更是尤为困难。而数据关系挖掘作为解决数据孤岛等难题的手段之一，可以有效的帮助企业将多样化的数据进行统一存储并挖掘出其中隐藏的价值，目前在公安、电信、金融等传统行业中的应用也正变得愈加广泛。为了了解数据关系挖掘背后的算法应用、技术难点等问题，InfoQ 对明略数据技术合伙人及 SCOPA 产品负责人任鑫琦进行了独家专访。

SCOPA 是明略数据去年底刚刚推出的一款数据关系挖掘新产品，它构建在企业大数据平台之上，可结合明略数据在特定领域与行业中积累的业务知识，进行领域模型的转换，并且将转换后的领域模型对象数据进行关联，将所有数据转换成业务人员能轻松理解的数据形式，挖掘出这些数据之间的联系，把有关联的数据放在一起，最后交给上层的业务人员用以展示或分析。

InfoQ：提到数据挖掘和数据分析，就不得不谈算法的问题。前一段时间谷歌 AlphaGo 在围棋对战中战胜世界顶级围棋棋手李世石，这使得机器算法的话题引起了一阵热议。能否请您谈一谈明略的 SCOPA 在实际使用时都用到了哪些算法？怎么用的？这些算法各自又有什么不同的特点？

任鑫琦：SCOPA 在做数据的关系构建或数据关联时，要用到的方法是多种多样的。因为在这一过程中所面对的数据形式、数据来源、数据种类同样也是多种多样的。基础的数据挖掘算法肯定是必要的，比如基础的分类算法和聚类算法，这也是明略数据在公安和金融领域通过实践而知的，不同于其他行业应用的一个重要方面。

传统行业的业务人员更多的是依靠自己的经验和习惯去总结一些类似于公式的东西，然后将抽样数据或者是能找到的结构化数据套用在这个公式上去计算，然后得到比如像重点人防控的数据模型或者是金融行业里的反欺诈数据模型等等。这些模型的问题大多在于它是源于“人”的经验，其数据特征都是由“人”的主管意识来决定的。

从传统的数据挖掘方法上来看，明略其实是利用相关技术，先将所有数据进行人工智能处理，比如先自动的按照一些基本特征去进行分类、聚类，虽然这中间产生的数据处理结果并一定能被人类完全理解，但是 SCOPA 会在这个基础之上再根据一些真实的数据样本，比如公安部门中的案件数据，或者金融领域里过去发生的欺诈行为的数据，来作为样本再进行训练。这样的话，之后得出来的规则集和模型，其实都是由真实的数据特征所决定的结果，相比“人”的主观意识来说会更精确。

另外在解决数据关联问题上，明略会把数据转化成类似知识图谱的形式去进行存储，帮助业务人员能够更容易地去理解这些数据。而在这之后，就可以结合很多在互联网领域中很成熟的图像数据挖掘与分析的方法，从中再继续提取数据特征，找到有用的信息。

比如一些离线的图挖掘算法，可以做一个省内一亿人口之间的数据记录关系网，然后就能从数据关系网当中挖掘出一些可疑的团伙或是一些正常的交集群，这些通过现成的数据挖掘算法就可以实现。甚至还可以做一些 link prediction

的预测工作，分析这张数据关系网里面哪部分处于活跃状态，哪部分未来可能会发生一些关联的事件。

同时这张数据关系网也可以做一些可视化的展示，或是可视化的分析。比如在一个群体内部，可以分析出哪些方面是权重点，而这就需要一些更具体的图挖掘或图分析的算法了。比如可以利用基于 Betweenness 或 Closeness 等方法去计算出一些核心点。

举一些简单例子，比如基于 Betweenness 计算的点，它相当于在一个犯罪团伙内所有通路和路径交汇最多的一个点，也相当于这个团伙组织架构的一个核心点，而这个点可能并不只一个，那么如果能够把这些点都一一破获的话，那么这个团伙或者组织就基本会落网了，这在公安部门打击一些非法传销或者非法金融链条的时候会有所应用。而基于 Closeness 的方法则是利用计算中心度的方式来寻找一些团伙内真正的核心人物，这个人关联到团伙内其他人的平均距离应该是最短的，这也是打击非法团伙的最快方式。

此外由于在大量的结构化数据之外还有很多非结构化的数据，尤其是像公安部门中的案情、笔录、出警描述这样的文本数据，里面往往都包含着非常重要信息，所以 SCOPA 所使用比较多的另外一类算法，就是自然语言处理 NLP，同时也会进行非常精准的命名实体识别，并计算实体之间的关系。比如可以通过一段文字描述锁定在某地区出现过的一群人，同时分析这些人之间的联系，其中哪些人跟某个案件有什么样的关联，受害人或被害人是谁，他们是否有一些共同的特征，某些地址、单位是否会跟他们产生关联，这些都是自然语言处理算法需要解决的问题。

当需要处理的案件描述非常多的情况下，比如 110 接警电话记录，或是警察调查走访的笔录等等，那么 SCOPA 就可以进行自动化的案件对比和分类工作，以便在大规模的案件描述里挖掘出一些数据特征，为一线调查人员的工作起到指引作用。

InfoQ：数据关系挖掘的作用毋庸置疑，理论上的方法也有很多，但是要想在实际的应用场景中做好落地，还需要考虑更多的细节问题。那么能否请您谈一谈进行数据关系挖掘时会面临哪些技术难点？

任鑫琦：关联数据挖掘或者更深入的说关系数据挖掘，研究的不仅仅是客观上的关联度，还会深入挖掘在物理世界中真实存在的某种准确的直接联系，同时还要确定是什么样的联系。那么在进行关系数据挖掘时的难点，主要就在于确定数据模型的特征时，必须要保证数据特征的准确性，否则可能做出来的模型也是不够精确的，而这种似是而非的数据模型在很多行业中其实是没有意义的，比如公安部门就必须使用非常严格的数据模型。

SCOPA 所使用的算法都是依赖于底层数据支持的，然而数据量越大并不一定就越好，而是数据的种类和来源越多越好。比如说公安部门需要确定犯罪嫌疑人之间的关系，那么如果能够拥有关联类数据、轨迹类数据、网络虚拟化数据、电信运营商数据等的话，就能确定嫌疑人经常出现的位置，这样可以依靠出现时间、空间、频次等几类模型来将这个人以及与之有关系的人或物给确定下来。

然而真正要完成这项工作，还需要克服两个挑战，第一个是如何尽可能多的收集和处理数据；第二个是如何在这么多复杂的数据之上挖掘关联性，这需要足够强的计算能力。

InfoQ：随着企业在大数据方面的需求不断扩大，数据的关联、关系挖掘在行业中的应用范围也正变得越来越广。您认为数据关系挖掘相关技术最近有哪些发展趋势呢？

任鑫琦：在没有大数据概念之前，很多时候是用数据库去做一些显性关联分析，而当有了大数据概念之后，更多地其实是想做隐性的关联分析与挖掘，也就是结合不同类型的数据，然后找到其中的联系。因此这其中的趋势，实际就是目前的数据关系挖掘更加偏向于跨领域数据或者跨类型数据的综合分析。

另外一个趋势就是数据分析中需要考虑的数据各种特征以及各种维度都越来越多，比如时间纬度、空间纬度、关系纬度、频次纬度等等，而这样一来数据关系挖掘的结果就会变得越来越准确。

目前还出现了一些类似于搜索引擎相关技术的数据分析技术，它可以通过一些文本及文字的匹配，进行一些类似关联度分析的数据挖掘。但是这种数据关系挖掘，可能今后发展的空间以及潜力会相对少一些。

InfoQ：目前看来，数据关系挖掘在保障公共安全以及维护治安方面的作用

是非常显著的，通过技术层面的手段，寻找数据之间的隐藏信息，这对于公安部门来说正是提升执法效率的途径之一，您能否简单地介绍一个相关的数据关系挖掘应用案例呢？

任鑫琦：明略曾经为一个市级公安局做了数据系统，之后当某个区域内经常出现电动车或者电动三轮车盗窃案后，直接通过数据关系挖掘在一分钟之内锁定了该盗窃团伙。

这其实是根据这个区域中的摄像头数据，先找出一些可疑车辆，接着分析在一定时间范围内这些车辆出现的位置，基于这些筛查工作的结果，再对比车主个人信息、违章记录以及与车主有关联的人，从而把范围缩小到一些小人群上，然后把这些人群的行为轨迹进行区域数据模型验证，确定他们在固定的时间段内，在固定范围内出现的概率，在进一步的筛查之后，计算出关联度最高的那群人，最后由调查人员再通过进一步的调查取证，锁定了该电动车盗窃团伙。

而过去一般遇到这样的案件，如果警方只用传统的数据检索和数据比对的方式的话，可能至少需要一个小团队工作三到四天才能破案。

任鑫琦，明略数据技术合伙人及 SCOPA 产品负责人，同时也是大数据架构、分布式计算、数据交互可视化领域的专家，主要从事大数据系统高效落地、优化架构以及便捷应用方面的工作。任鑫琦于 2009 年毕业于北京大学计算机科学与技术系，2009 年至 2012 年在 SLB 从事核心软件开发与架构设计工作；2012 年加入秒针系统，负责大数据集群运维和系统架构工作，在两年的时间内完成了公司计算架构的转变，集群规模达到 500 台，总数据量超过 3PB；2013 年加入明略，先后负责集群管理和日志分析两款产品的研发工作，曾落地实践多个金融、公安领域项目。

专访明略数据技术合伙人孟嘉： SCOPA 架构升级下的实践与优化

作者 孟嘉

随着“数据爆炸”时代的来临，数据挖掘成为一项重点工作，针对海量，混杂的大数据而非少量、随机化、样本化的精准数据，其关键是找到并建立不同数据间的相关性，并对其进行模式分析。

数据挖掘在明略数据构建的数据生态中处于极其重要的一环，为最大程度发掘数据中所隐含的关系、知识和规律，明略数据成立了科学家团队，不断打磨在人工智能、深度学习等领域的经验，同时它也组建了自己的 GPU 集群，以发挥其天生的计算优势，加快对深度学习平台的训练速度。在一些深度学习最新算法应



用上，明略数据也具有较为领先的创新。

为更好的了解明略技术背后的原理和机制，我们专访了明略数据技术合伙人孟嘉，跟他聊聊数据关联分析产品 SCOPA 的技术应用、架构演进及其对知识图谱和图数据库相关技术的一些思考。

SCOPA2.0架构演进背后的技术突破

今年，明略对 SCOPA 2.0 做了一次大规模的架构升级，目的是拓展 SCOPA 的平台化战略。具体内容包括三方面。



一是为向平台化方向发展，开放了 API 和插件体系。这样，SCOPA 项目团队成员以及合作伙伴都可以基于 SCOPA 快速开发和部署新的应用或功能。

二是在数据存储方面，把存储层抽取成一个独立的数据库，或者说，开发了一个面向知识图谱存储的独立 NoSQL 产品。SCOPA 底层存储的是一张巨大的知识网络（知识图谱），这样的独立数据库可为二次开发的人员提供独立使用并调试的可能，同时还可让开发过程直接以插件的形式接入到 SCOPA 的整体平台中。

三是在数据整合方面，SCOPA 2.0 已统一对非结构化数据和结构化数据的视图描述规范，提升了知识构建到存储的效率。之前明略数据的很多任务都是靠人工或者是半自动化的形式处理，这就好比是由几百甚至上千个离线任务组成的复杂的系统，互不相干的任务并不能并发执行。例如，在进行知识抽取的过程中只有先抽取实体，才能抽取到实体之间的关系。这样，面对任务多、串行时间太长等问题，SCOPA 2.0 加入基于 DAG 的任务调度系统，轻量级任务调度系统可以把

抽取实体关系执行条件组织成一张有向无环图（DAG），将离线和在线任务紧密结合在一起，同时还支持一部分任务重试、错误的监控，从而能极大提升 SCOPA 后端的整体效率。

知识图谱中的数据噪声清除及知识推理构建

知识图谱由节点和边构成，提供了从“关系”角度分析问题的能力。在 SCOPA 中，知识图谱是关系挖掘的载体，它将数据抽象成实体、关系和事件，利用包括属性图（Property Graph）在内的混合结构组织数据，点代表实体，边代表关系，再把各个实体通过带有属性的关系联系起来。

但同时，知识图谱中还有着对关系挖掘最终结果形成影响的数据噪声。其出现原因在于数据错误、数据缺失，或大量的数据冲突和数据冗余。处理错误数据最简单的办法是在数据治理的过程中做离线的规则过滤。例如，对于结构化数据的冲突，可以在治理过程中设置治理规则和增加数据优先级的概念，而对于非结构化数据，可采用自然语言处理中的实体消歧分析技术。另外，在 SCOPA 中，数据支持溯源与多版本，机器暂时无法处理的噪声会在用户分析的时候留给用户进行判断。

此外，在清除数据噪声之外，关系挖掘还有一项重要能力是发现已有结果中可能隐含的新知识，即利用算法完成对知识的推理。知识推理以知识图谱的构建为基础，SCOPA 的目标是通过数据挖掘，机器学习等方法，让机器学会人类的推理过程，使用户从海量而繁琐的业务数据中解脱出来，比如公安领域的团伙自动发现，重大事件预警都有知识推理的过程。另外，知识推理还可以用于发现实体间新的关系的预测与发现。

图数据库的优势及选型策略

大数据时代，包括文本、图片、视音频等都是未经处理的非结构化数据。为更好地挖掘非结构化数据中有价值的成分，SCOPA 对结构化和非结构化数据进行了统一的元数据描述，也称为视图，视图是链接其各个模块的桥梁。SCOPA 在结构化数据和非结构化数据上使用了不同的分析处理方法：在关系挖掘使用的技

术上，结构化数据主要运用了基于规则和机器学习的方法，非结构化数据运用了大量的自然语言处理方面的算法，而基于图的算法被同时应用到治理后的结构化与非结构化数据。

目前，基于图的数据模型比较常用的有两种，一种是 RDF（Resource Description Framework），一种是 Property Graph 属性图。RDF 是 W3C 标准，目前用作研究较多，市场表现比较平淡，优势在于数据交换。而属性图是从工程实践中总结出来的，得到了市场的认可，越来越多的企业和项目采用基于属性图的图数据库产品。

因为 SCOPA 是用图的形式来存储点边关系，因此会采用分布式图数据库。图数据库在多层关系挖掘分析方面比传统关系型数据库有着明显的优势。传统关系型数据库是按照表的结构来保存数据的，在查找数据间关系时往往需要对表之间做 join 操作。当需要 join 多个表，且每个表的数据量都很大时，这种关系查找就会变得很吃力。而图数据库更加符合现实世界对数据的描述。特别是处理关系数据，进行关系推演等的时候就有着天生的优势。例如，在查询命令执行上，一旦查询度数过多，关系型数据库可能就没法满足查询需求。而图数据库因为采取了邻接表的方式存储，随着查询深度的增加，它的代价是线性的，所以图数据库天生就适合这类的关系推演。

SCOPA 在进行图数据库选型时，主要考虑的是其与大数据平台的整合度、功能集与易用程度、性能等方面的因素。基于此，SCOPA 使用了分布式图数据库 Apache Titan，Apache Titan 在实现 Apache Tinkerpop 图协议的接口，使用 Gremlin 查询语言上具有很大优势，为存储和处理大规模图做了大量优化。

此外，源码级别的掌控能力也是图数据库选型时需重要考虑的因素。为使图数据库的代码和操作流程更接近于自然语言机制，SCOPA 研发团队又在 Titan 的基础上进行了多处优化，包括处理并发读取实体，超级节点的优化等。

人工智能 = 机器学习 + 大数据

从研究生阶段专注的应用服务器分布式集群方向，到后来工作过程中不断接触高并发、高吞吐的网络应用，再到近几年参与多个实际落地的大数据项目，孟

嘉经历过无数次的试错和复盘。在他看来，架构师既要见多识广，从整体提掌握、了解系统全局，又要深入到关键细节，思考如何突破系统的瓶颈。此外，他认为架构师也需要对业务具有深层次了解，尤其作为 To B 系统架构师，需要不断接触最终用户，理解需求。

为更好的把握数据挖掘技术接下来的突围点，孟嘉也提出了他对国内数据挖掘发展方向的预测——“人工智能 = 机器学习 + 大数据”。从目前的使用趋势看，人工智能已在更多领域得到广泛应用，在很多事情上也将会慢慢取代人力。因此，一方面，数据量不断增大为机器学习提供了更多的训练样本，另一方面数据的特征维度的不断变大也使越来越多的数据可以参与到训练学习中并发挥其价值。

孟嘉，2008年从北大计算机系硕士毕业，之后在某外企做研发工作，2014年底加入明略数据。目前在明略数据任系统架构师和技术经理。主要负责带领大数据关系挖掘分析平台SCOPA架构组对产品进行架构设计、带领研发组研发SCOPA的底层存储和在线计算部分。

版权声明

InfoQ 中文站出品

洞见数据之密

©2016 极客邦控股（北京）有限公司

本书版权为极客邦控股（北京）有限公司所有，未经出版者预先的书面许可，不得以任何方式复制或抄袭本书的任何部分，本书任何部分不得用于再印刷，存储于可重复使用的系统，或者以任何方式进行电子、机械、复印和录制等形式传播。

本书提到的公司产品或者使用到的商标为产品公司所有。

如果读者要了解具体的商标和注册信息，应该联系相应的公司。

出版：极客邦控股（北京）有限公司

北京市朝阳区洛娃大厦 C 座 1607

欢迎共同参与 InfoQ 中文站的内容建设工作，包括原创投稿和翻译，请联系
editors@cn.infoq.com。

网 址：www.infoq.com.cn

Geekbang>

极客邦科技

整合全球优质学习资源，帮助技术人 and 企业成长

InfoQ

技术媒体

EGO EXTRA GEEKS' ORGANIZATION
NETWORKS

职业社交

StuQ

职业教育