大数据分析技术方案

李万鸿

一. 目标

现在已经进入大数据时代, 数据是无缝连接网络世界与物理世界的DNA。发现数据DNA、重组数据DNA是人类不断认识、探索、实践大数据的持续过程。大数据分析可以有效地促进营销,个性化医疗治病，帮助学生提高成绩,利于老师提高教学水平,利于掌握舆情了解产品问题加以改进，公司的许多产品可以用到大数据技术，公司需要加强大数据技术的研究并实际应用。这里对目前最先进和最实用的用户画像技术、文本挖掘、机器学习、人工智能等进行讲解，并分析大数据分析的常用算法。

http://xinsheng.huawei.com/cn/blog/detail\_79549.html

二. 用户画像

1. 定义

http://xinsheng.huawei.com/cn/index.php?app=space&mod=Blog&act=spaceIndex

用户画像指根据用户的特性和行为数据建立标签，准确描述用户的特点，支持对用户的精准营销等活动。用户画像，即用户信息标签化，就是企业通过收集与分析消费者社会属性、生活习惯、消费行为等主要信息的数据之后，完美地抽象出一个用户的商业全貌作是企业应用大数据技术的基本方式。用户画像为企业提供了足够的信息基础，能够帮助企业快速找到精准用户群体以及用户需求等更为广泛的反馈信息。

大数据用户画像其实就是对现实用户做的一个数学模型，在整个数学模型中，其核心是，怎么描述业务知识体系，而这个业务知识体系就是本体论，本体论很复杂，我们找到了一个特别朴素的实现，就是标签。建好模型以后，要在业务的实践中去检验，并且不断完善，不断丰富这个模型，来达到利用比特流对人越来越精确的理解。用户画像不是一个数学游戏，不是一个技术问题，实际上是一个业务问题。因为最核心的是你去如何理解用户，了解你的用户。它是技术与业务最佳的结合点，也是一个现实跟数据的最佳实践。

一个画像样例：基于他这个人可以知道他所在的城市是在北京，男性，公司在百分点，喜欢的品类是男鞋、运动鞋，喜欢的品牌有耐克、阿迪达斯等等。每一个标签都有一个权重值。可以看到，耐克的权重值比阿迪达斯更高一些。

2用户画像方法

用户画像的焦点工作就是为用户打“标签”，而一个标签通常是人为规定的高度精炼的特征标识，如年龄、性别、地域、用户偏好等，最后将用户的所有标签综合来看，就可以勾勒出该用户的立体“画像”了。具体来讲，当为用户画像时，需要以下四个阶段：

http://xinsheng.huawei.com/cn/index.PHP?app=space&mod=Blog&act=spaceIndex

1.战略解读：企业选择构建用户画像平台，可以实现不同的战略目的，如提升产品服务质量、精准营销等。根据战略目的的不同，用户画像的构建也有所区别。因此首先需要明确用户画像平台的战略意义、平台建设目标和效果预期，进而有针对性的开展实施工作。

2.建模体系：对用户画像进行数据建模，结合客户实际的需求，找出相关的数据实体，以数据实体为中心规约数据维度类型和关联关系，形成符合客户实际情况的建模体系。

3.维度分解：以用户、商品、渠道三类数据实体为中心，进行数据维度分解和列举。根据相关性原则，选取和战略目的相关的数据维度，避免产生过多无用数据干扰分析过程。

4.应用流程：针对不同角色人员的需求（如市场、销售、研发等），设计各角色人员在用户画像工具中的使用功能和应用/操作流程。

用户画像平台的战略意义：

1.完善产品运营，提升用户体验：改变以往闭门造车的生产模式，通过事先调研用户需求，设计制造更适合用户的产品，提升用户体验。

2.对外服务，提升盈利：根据产品特点，找到目标用户，在用户偏好的渠道上与其交互，促成购买，实现精准运营和营销。

如何搭建用户画像平台？

公司应搭建一个用户画像平台，将本身拥有大量用户数据的数据平台和可视化数据工具平台连接起来，根据不同的用户交互场景，应用挖掘数据平台的价值，让研发生产，用户研究，市场营销等人员能够根据需要，随时自主地分析不同产品用户特征，快速洞察用户需求。该平台需要回答的核心问题是：

用户是谁?

用户需求是什么?

用户在哪里？

用户画像建模体系

完善的用户画像平台需要考虑周全的模型体系。通常来讲，构建用户画像平台所需的数据分成用户、商品、渠道三类实体。

1.用户：数据维度包括自然特征、兴趣特征、社会特征、消费特征。从数据特点上看，又可分为基本属性和衍生标签，基本属性包括年龄、性别、地域、收入等客观事实数据，衍生标签属于基本属性为依据，通过模型规则生成的附加判断数据。

2.商品：数据维度包括商品定位和商品属性。商品属性即商品的功能、颜色、能耗、价格等事实数据，商品定位即商品的风格和定位人群，需要和用户标签进行匹配。

3.渠道：渠道分为信息渠道和购买渠道。用户在信息渠道上获得资讯，在购买渠道上进行商品采购。不同类型的用户对渠道有不同的偏好，精准的选择对应的渠道才能提高效率和收益。

用户画像数据维度

针对每一类数据实体，进一步分解可落地的数据维度，形成字段集。

1.用户数据：

用户画像使用场景

场景一，按需设计：改变原有的先设计、再销售的传统模式，在研发新产品前，先基于产品期望定位，在用户画像平台中分析该用户群体的偏好，有针对性的设计产品，从而改变原先新产品高失败率的窘境，增强销售表现。比如，某公司想研发一款智能手表，面向28-35岁的年轻男性，通过在平台中进行分析，发现材质=“金属”、风格=“硬朗”、颜色=“黑色”/"深灰色"、价格区间=“中等”的偏好比重最大，那么就给新产品的设计提供了非常客观有效的决策依据。

场景二，精准营销：针对已有产品，寻找所偏好的精准人群分类，以及这些人群在信息渠道和购买渠道上的分布比例，来决定广告投放和活动开展的位置、内容等，实现精准营销。

用户画像平台技术方案

系统架构

从数据源到最终展现分成如下几层：

1.数据源：包括来自各个业务系统和媒介的分析数据源，其载体包括数据库、文件、大数据平台等。

2.数据建模：根据用户画像建模体系，配置数据模型。

3.数据集市：每个数据集市是基于一个主题做好轻量建模的细节数据，数据按照列存储的方式，被高效压缩，打好标签，存储在磁盘中。当需要计算时，采用内存计算来进行数据计算，并且每台机器节点会同时计算，最终会将结果送往可视化分析层做展现。

4.可视化分析：采用永洪敏捷可视化分析作为前端交互组件。无论是业务用户还是IT开发人员都可以通过主流浏览器来访问可视化分析系统，用户还可通过移动终端来访问系统。可视化分析系统提供系统监控，权限多级管理，多维数据分析，等等功能，还支持自服务式报表设计和数据分析。

可视化分析：敏捷可视化分析

通过对用户行为数据进行分析和挖掘，发掘用户的偏好，逐步勾勒出用户的画像。用户画像通常通过业务经验和建立模型相结合的方法来实现，但有主次之分，有些画像更偏重于业务经验的判断，有些画像更偏重于建立模型。

业务经验结合大数据分析为主勾画的人群，此类画像由于跟业务紧密相关，更多的是通过业务人员提供的经验来描述用户偏好。举个例子，比如：根据业务人员的经验，基于客户对金额、利润、信用等方面的贡献，建立多层综合指标体系，从而对用户的价值进行分级，生成用户价值的画像。一方面我们的产品经理可以根据用户价值的不同采取针对性的营销策略，另一方面通过分析我们的不同价值等级用户的占比，从而思考如何将低价值的用户发展成高价值的用户。

再比如，通过用户在下单前的浏览情况，业务人员可以区分用户的购物性格。有些用户总是在短时间内比较了少量的商品就下单，那么他的购物性格便是冲动型；有些用户总是在反复不停的比较少量同类商品最后才下单，那么他的购物性格便是理性型；有些用户总是长时间大量的浏览了很多商品最后才下单，那么他的购物性格便是犹豫型。对于不同购物性格的用户，我们可以推荐不同类型的商品，针对冲动型用户，我们直接推荐给他/她最畅销的同类商品，而理性型用户我们推荐给他/她口碑最好的商品。并且针对每一个用户，我们根据其购物性格定制了个性化的营销手段。

以建立模型为主勾画的人群，我们不能认为买过母婴类用品的用户家里就一定有小孩，因为这次购买很有可能是替别人代买或者送礼物。所以我们要判断这个用户所购买的母婴类商品是否是给自己买。根据用户下单前浏览情况、收货地址、对商品的评价等多种信息建立模型，最终判断出用户家庭是否有小孩。再根据购买的商品标签，比如奶粉的段数，童书适应年龄段等信息，建立孩子成长模型，在孩子所处不同的阶段进行精准营销。

3用户画像架构

用户画像的四阶段

用户画像的焦点工作就是为用户打“标签”，而一个标签通常是人为规定的高度精炼的特征标识，如年龄、性别、地域、用户偏好等，最后将用户的所有标签综合来看，就可以勾勒出该用户的立体“画像”了。

具体来讲，当为用户画像时，需要以下四个阶段：

战略解读：企业选择构建用户画像平台，可以实现不同的战略目的，如提升产品服务质量、精准营销等。根据战略目的的不同，用户画像的构建也有所区别。因此首先需要明确用户画像平台的战略意义、平台建设目标和效果预期，进而有针对性的开展实施工作。

建模体系：对用户画像进行数据建模，结合客户实际的需求，找出相关的数据实体，以数据实体为中心规约数据维度类型和关联关系，形成符合客户实际情况的建模体系。

维度分解：以用户、商品、渠道三类数据实体为中心，进行数据维度分解和列举。根据相关性原则，选取和战略目的相关的数据维度，避免产生过多无用数据干扰分析过程。

应用流程：针对不同角色人员的需求(如市场、销售、研发等)，设计各角色人员在用户画像工具中的使用功能和应用/操作流程。

用户画像的意义

完善产品运营，提升用户体验：改变以往闭门造车的生产模式，通过事先调研用户需求，设计制造更适合用户的产品，提升用户体验。

对外服务，提升盈利：根据产品特点，找到目标用户，在用户偏好的渠道上与其交互，促成购买，实现精准运营和营销。

用户画像的构建是有难度的

主要表现为以下四个方面：

为了精准地描述用户特征，可以参考下面的思路，从用户微观画像的建立→用户画像的标签建模→用户画像的数据架构，我们由微观到宏观，逐层分析。

首先我们从微观来看，如何给用户的微观画像进行分级呢?如下图所示

总原则：基于一级分类上述分类逐级进行细分。

第二分类…

第三分类…

…….

完成了对客户微观画像分析后，就可以考虑为用户画像的标签建模了。

从原始数据进行统计分析，得到事实标签，再进行建模分析，得到模型标签，再进行模型预测，得到预测标签。

最后从宏观层面总结，就是得到用户画像的数据架构。

百分点的画像标签体系包括：人口属性、上网特征、营销特征、内容偏好、兴趣偏好等。

以手机商品属性为例，包括品牌、品类、型号、上市时间、价格、颜色、网络、操作系统、分辨率、屏幕尺寸等等。

标签管理体系具有如下特性。

有多种标识方式对用户进行识别，这就像社会生活中的身份证号码一样，只不过换成了网络空间的手机号、Cookie、IMEI、Email、微博、微信账号等，在处理过程中，这些信息都是加密的，机器知道但人不知道。

百分点用户画像逻辑架构如下图所示，通过对电商、社区、移动应用、微博、微信等多种类别的数据源进行采集，然后对用户进行画像，最终在个性化推荐、用户洞察、精准营销等方面进行应用。百分点的数据源多且庞大，服务的客户超过了1500多家，覆盖行业超过了40多个。举例来说，一个网民，他在访问一个电商A，同时又访问了一个电商B，这两个电商本身的知识体系是不一样的。比如说这个用户他访问一双鞋，他在电商A上的品类可能是鞋-男鞋-运动鞋，在网站B上可能是运动-户外-男鞋，品类描述可能是不一样的。所以百分点打造了这么一个系统，叫商品画像系统。通过这个系统，所有的标签就有了一个标签规划，之后就可以去构建这个用户在全网的用户画像标签。用户画像只是一个起点，而不是一个结束。基于此，还可以打造一系列的服务，比如精准营销、个性化推荐等。

下图是用户画像的技术架构图。我们可以看到总共分为五层：第一是数据源；第二层是数据采集服务，百分点有一堆数据采集服务，包括我们的数据探头，能够对用户的行为进行一个实时采集；第三层是数据预处理，主要是结构化；第四层是商品画像，这一块都是我们的用户画像服务。我们可以看到用户画像是分两大块，实时处理更偏重于预测用户画像的需求，离线处理更偏重于用户的长期偏好；第五是统一的数据接口，还有就是集群，上面可以接入各种各样的应用。

下图是用户标签产出流程示例。

https://user.qzone.qq.com/1691589876?\_t\_=0.007478161249309778

用户在互联网上的行为主要分为电商类、社交类和媒体类。每种行为差异很大，电商类行为包括浏览、搜索、添加购物车、收藏、支付等，而社交类则是点赞、转发、评论等。

接着下一步需要对页面标签进行抽取，在做这件事情之前需要训练模型，首先准备训练数据，通过标注和规则生成，再就是对于序列集做一个序列化处理。首先会得到一个弱模型，最终得到一个强模型，然后把自己的参数都保留下来。这个时候我们会加一个决策，如果说效果不太好的话，我们会进行下一轮的优化。当这个模型设置之后，我们就可以去做预测了。我们的预测总共分为四大块，包括输入、输入预处理、预测和产出。也就是说用户这个标签已经有了，这个标签对用户的信誉度是1还是0？这个时候就到了用户行为建模。用户行为建模的背后思想主要可以认为有两大块，成本越高行为权重越高，下单就比浏览更高一些，时间越近的行为权重越高，比如我今天看了一个手机，一定比我一周前看了一次电脑权重要高一些。我们可以按场景去分，首先是产生需求，再就是决策，然后是结束，百分点基于业务考虑，实行标签权重积累的机制。

这是我们的客户某航空公司的案例，项目目的是挖掘高价值旅客，希望通过分析旅客出行偏好优化运力资源。最终百分点帮他构建了5个标签大类，75个标签小类，数万个小标签，以下是当时的一些效果截图。

刚才讲的都是百分点已经做的事情，但是百分点做得还远远不够。接下来可能会在四大方面深入思考和实践：一是不同的场景，也就是说用户在家里和在办公环境下代表的偏好是不一样的；二是用户心理学特征，比如当一个用户看一件女装的时候，她这个时候是无聊去逛还是有目的的逛，反映在标签权重上是不一样的；三是让用户主动反馈反感点，我们强调了许多，一般都是在强调用户喜欢什么，但是用户不喜欢什么，我们做得还不够，我们应该让用户主动告诉我们他不喜欢什么，比如他不喜欢吃葱，他不喜欢吃羊肉串，这样我们预测的时候会准得多；四是用户的兴趣转移快速捕获，一开始我们使用的是一个半衰期的，而且按频率细分，我们是否可以按人去分？比如按访次去分？比如针对品类手机这个标签，对于手机发烧友，可能过了一年他依然会对手机比较感兴趣，但是对于像我这种，只有想购买的时候才去看，可能我两天不看，就表示这个兴趣已经衰减为零了。

4基于用户画像的精准营销

通过建立模型对用户的数据建标签，实现用户画像，再根据推荐算法实现精准营销，模型要考虑精度和稳定度，进行充分的修改、完善。

推荐系统的业务架构如下：

第一层是推荐业务活动层，把推荐结果展示给用户。

第二层是推荐算法层，包括用户画像推荐、情景推荐等。

第三层是索引层，对产品、用户数据建立索引，提高查询速度。

第四层是数据层，存储用户、产品及推荐基础数据如推荐模型。

1场景引擎：个性化的核心，判断用户处于哪个购物环节，有什么样的购物目标；

2规则引擎：业务的核心，结合用户、场景、算法输出数据和业务KPI，决定为用户推荐哪些内容；

3算法引擎：计算用户之间的相似度、商品之间的相似度、用户对商品的评分、用户分群、热门排行……，采用用户协同相关和商品协同相关等算法，根据用户画像、产品画像的数据进行推荐。

4展示引擎：将推荐内容以最佳的展示方式呈现在用户面前。

推荐主要是深挖用户潜在的购物兴趣或学习兴趣，在合适的时间推荐合适的商品，做到有的放矢，从而实现精准营销，提升转化率和效益。推荐的方法是，根据用户的短期兴趣采用用户意图引擎推荐商品，根据用户的长期兴趣采用用户画像引擎推荐商品，根据用户的潜在兴趣采用千人千面引擎基于用户协同相关算法推荐商品；根据合适的场景采用情景推荐引擎推荐商品，根据用户的购物周期采用反向推荐引擎推荐商品，根据最近热点和用户行为推荐商品。推荐可以用到许多方面，如：

根据用户和购买场景使用storm进行实时计算，给出推荐结果；对大量样本数据使用Spark进行离线机器学习计算，产生模型，用于用户画像权重确定和实时计算。大规模批处理使用Hadoop的mapReduce计算。对用户的搜索也可以使用用户画像和商品画像进行结果展示。用户的行为数据在变化，商品的信息也在变化，用户画像和商品画像需要定时进行修改，比如半个月或一个月。在Hbase中对商品标签和用户标签等建有表保存数据，并根据这些数据用机器学习训练算法模型，模型结果保存在hbase，在具体推荐的时候取近一个月的数据带入模型进行计算，多种推荐结果根据规则计算后的到最佳的推荐结果，再用展示引擎显示给用户。计算的中间结果保存在hbase中。用户购买之后，商品和用户的标签要及时更新。通过用户画像精准营销可以把转化率提升30-500%。

用户画像的推荐算法：

用户画像的算法及其他算法都需要不断优化。

商品、用户的画像标签权重的计算公式如下：

主题标签的权重计算公式：

根据计算公式算出每个标签的weight。

5推荐算法构建方法

1.计算权重系数ɑ

画像权重需要考虑时间、地点即网址、网页内容、事件即用户活动,如购买商品，下面是具体例子。

什么时间：时间包括两个重要信息，时间戳+时间长度。时间戳，为了标识用户行为的时间点，如，1395121950（精度到秒），1395121950.083612（精度到微秒），通常采用精度到秒的时间戳即可。因为微秒的时间戳精度并不可靠。浏览器时间精度，准确度最多也只能到毫秒。时间长度，为了标识用户在某一页面的停留时间。

什么地点：用户接触点，Touch Point。对于每个用户接触点。潜在包含了两层信息：网址 + 内容。

网址：每一个url链接（页面/屏幕），即定位了一个互联网页面地址，或者某个产品的特定页面。可以是PC上某电商网站的页面url，也可以是手机上的微博，微信等应用某个功能页面，某款产品应用的特定画面。如，长城红酒单品页，微信订阅号页面，某游戏的过关页。

内容：每个url网址（页面/屏幕）中的内容。可以是单品的相关信息：类别、品牌、描述、属性、网站信息等等。如，红酒，长城，干红，对于每个互联网接触点，其中网址决定了权重；内容决定了标签。

具体的权重计算公式是：

tag\_tf为用户购买商品的成本，tag\_action为用户活动，tag\_attenuation为用户活动的时间也是衰减因子 ，n为用户购买商品所需要的时间。通过log和数据库的数据可以得到这些数据。

对画像权重的系数ɑ，需要采用逻辑回归的Cost Function进行计算。可以通过matlab软件来计算θ，确定一个标签的值，如价格{22,11,33，。。。}，然后带入公式进行计算，对不同标签分别计算。标签的值有数字的，也有文字的，可以用分类的方法把文字转化为数字，比如地域的东方为1，南方为2。标签的取值数字化后需要进行归一化处理，把数字变为0到1之间的数，这样利于提高回归处理的精度。归一化处理后需要对数据进行去噪，从而保证训练集跟预测集特征分布的一致性。数据其他推荐算法都需要对数据做归一化处理。

逻辑回归的Cost Function可以表示为：

由于y 只能等于0或1，所以可以将逻辑回归中的Cost function的两个公式合并，具体推导如下：

故逻辑回归的Cost function可简化为：

对于这个公式，这里稍微补充一点，注意中括号中的公式正是对逻辑回归进行最大似然估计中的最大似然函数，对于最大似然函数求最大值，从而得到参数(\theta\)的估计值。反过来，这里为了求一个合适的参数，需要最小化Cost function，也就是：

minθJ(θ)

而对于新的变量x来说，就是根据hθ(x)的公式输出结果：

与线性回归相似，这里我们采用梯度下降算法来学习参数θ，对于J(θ):

目标是最小化J(θ)，则梯度下降算法的如下：

对J(θ)求导后，梯度下降算法如下：

注意，这个算法和线性回归里的梯度下降算法几乎是一致的，除了hθ(x)的表示不同。

2．逻辑回归算法构建过程

根据用户的行为日志，建立用户的品牌偏好，并预测他们在将来一个月内对品牌下商品的购买行为。将这个问题转为一个二分类问题，可以采用现有的有监督的分类回归算法（逻辑回归、SVM、随机森林、GBRT）等，最后基于基本的算法模型进行模型融合。将训练集分为“购买过”跟“未购买过”两部分，分别训练，对预测集同样分为两部分，分别预测，最后两部分融合。

· 为了训练集和预测集的信息量差异，将不同时间尺度的训练集预测集分别训练预测，最后融合。

· 逻辑回归模型与树模型采用不一样的特征体系，使得融合效果更好。

解决步骤

· 构建特征

· 构建训练集和预测集

· 数据预处理

· 构建本地评估框架

· 模型融合

特征抽取

训练样本的基本形式是一对有交互的用户品牌对，我们对这样的用户品牌对提取特征。首先，从总的分类上分为用户的特征、品牌的特征、用户对这个品牌的特征；其次，考虑用户行为的周期性，从时间维度上分为最近一个星期、半个月、一个月等不同时间段的特征；然后，根据这次比赛给定的四种行为（点击、购买、收藏、购物车），从类型维度上去提取特征。在多维度考虑的基础上，主要提取的是点击数、天数、订单数这样的特征。

从数据到业务需要算法以及特征体系的驱动，在特征抽取的过程中，最本质的一点是要去让数据“贴合”业务，反过来就要求我们从业务指标出发提取特征。参照一些电商分析网站以及自己对业务的理解，诸如用户购买力、品牌购买成本、转化率、用户对品牌的忠诚度等等指标都可以从给定的数据集得以体现。典型的，比如如何衡量品牌的购买成本，可以使用购买该品牌的用户数、回头客的数量、平均每个回头客的回头次数等等来体现这个指标。

特征抽取的最后一部分主要是在之前特征体系的基础上做一些补充，主要有：

1.比值类特征。主要是一些基本特征的相除，这个一般是具有直观含义的，比如用户购买某个品牌的次数除以用户总的购买次数，这可以体现这个品牌在用户整个购买行为中的地位。

2.均值、方差类特征。主要是一些购买次数、活跃天数等统计值的平均值或方差，可以体现某些用户某些行为的平均水平或者波动情况。

3.策略。Season1的时候更多的考虑通过策略来预测，Season2的时候把一部分策略转化为特征，比如最后一次购买之后的点击次数等。

模型训练

逻辑回归、随机森林和GBRT，逻辑回归本质是一个线性模型，如果为了防止过拟合，可以添加正则化项；如果方便特征的筛选，可以使用逐步逻辑回归；逻辑回归在大数据量下精度会下降，可以通过添加大量的特征（比如哑变量的方式）来提高精度，这次比赛来看，跟随机森林和GBRT相比，逻辑回归单模型上存在天然的劣势。随机森林跟GBRT都是一种ensemble方法，随机森林是一种多棵决策树组合，最后bagging的方法；GBRT是一种boosting方法，每棵树学习的是上一棵树的残差。经过学习之后的LR分类器其实就是一组权值w0,w1,...,wm。

对于模型的训练我们团队主要有两点心得：

· 训练集和预测集的构造。我们通过前三个月提取特征，最后一个月标记来构造训练集；在预测的时候，我们使用四个月所有的行为构造预测集，跟训练集相比，时间尺度上不一样，但是这样会带来更多信息量、更多的用户品牌对，比三个月预测的效果好一些。在处理训练集和预测集时间尺度上不一致时，需要进行数据规格化处理，这里我们使用的是归一化方法（仅仅归一化时间尺度不一致的特征）。归一化方法对离群点比较敏感，这个过程中，需要对数据进行去噪，从而保证训练集跟预测集特征分布的一致性。去噪时，我们去除了异常点击量和异常购买量的用户品牌对。

· 健壮的本地测试。本地测试时，我们尝试过两种方法，一种是将原始的训练样本集划分，80%训练，20%用来本地测试，这种方法跟实际线上训练预测的关系不一样，并且训练集跟测试集的特征分布式一致的，很多问题不易察觉。另外一种方法，本地最后一个月有购买行为的用户品牌对用来测试，前三个月的数据按照线上训练预测的方式构造训练集跟预测集，这种方式时间跨度跟少一个月，但是正样本的数量以及跟线上实际情况很接近，我们主要使用这种方法来本地调参以及部分特征的调试。

首先需要获取最佳的评估参数，使得训练得到的LR模型可以获得最佳的分类效果。这个过程也可以看做是一个搜索的过程，即在一个LR模型的解空间内，如何查找一个与我们设计的LR模型最为匹配的解。为了达到能够获取对应的最佳LR模型，我们需要设计一种搜索策略，考虑按照什么样的准则去选择最优的模型。

如何选择最佳的LR模型，直观的想法就是通过预测模型的结果与真实值的匹配程度评价预测模型的好坏。在机器学习领域中，使用损失函数(loss function)或者代价函数(cost function)来计算预测结果与真实值得匹配程度。

选取了牛顿-拉斐森迭代算法以及L-BFGS算法作为LR模型的迭代算法。当学习算法迭代完成之后，我们可以获对应各个属性的权重。接下来的任务我们需要对现有属性与响应变量之间的显著性进行检验，针对已有的训练模型对应的属性集进行验证，删除显著性不符合阈值的特征。由于在构建风险函数的时候，使用了MLE方法，因此可以使用Wald Test对于计算得到的参数，进行显著性验证。在使用Wald Test之前，要确保期望值与评估值之间的差值符合正态分布。Wald统计变量的一般形式：

其中

表示评估值，表示期望值，表示评估值方差。在本次试验中我们将原假设设定为，即表示现有的属性与响应变量无相关性，因此本实验的Wald统计值可以表示为：

其中是实际估计的参数值，是的标准方差。由于Wald统计值对应卡方分布，因此可以利用卡方分布计算P值，如果P值大于指定的阈值，那么可以认为原假设成立，即该属性与响应变量是显著不相关，删除该变量，否则保存该变量。在实际的训练过程中，每次验证属性显著性的时候，只挑选P值最大与人为设定的阈值进行比较；如果选择的P值不大于阈值，那么模型训练完毕；否则删除选择的P值对应的属性，更新预测模型。重新学习更新后的预测模型，推测对应的权重值，然后再次对各个属性进行Wald Test验证。重复上面的过程，直到没有任何变量的Wald Test对应的P值都不大于人为设定的阈值为止。到此整个模型的训练过程结束。

优化画像算法、主题推荐算法等都按照这个方法完成。

三、文本挖掘

1.内容、方法与价值

文本分析是利用自然语言处理（NLP）技术分析文本文档、社交媒体、网页等文本数据的一种应用。随着电子商务、数字营销和大数据技术的高速发展，数据驱动的文件管理、用户体验管理已经成为企业核心竞争力，而文本分析则是用户体验管理的关键应用。而对传统文本文档这些相对增量不大、总量稳定的文本数据进行分析，则凸显其知识、情报、价值挖掘，尤其是对海量文本的简化、标记化、知识化，则是构建专家系统、人工智能、知识图谱的基础。

文本简化为词向量关联分析后的可视化显示

文本分析的目的在于从根本上把所有的非结构化数据整合从而化为结构化数据，从之前被认为难以量化的海量文本中抽取出大量有价值的、有意义的数据或信息。

所以，对文本数据进行分析，我们能得到有价值的商业线索或洞察。

大数据文本分析提取出的主要维度

基于大数据的文本分析能够揭示出潜藏在文本信息当中的趋势和关联，为商业决策、行业趋势研究和热点内容追踪提供有力支持。

大数据文本信息的来源

数据还包括企业本身的数据及其他数据。基于大数据的文本分析能快速准确的识别出企业/品牌/产品自身及竞争对手在互联网上的口碑变化，深度挖掘文本数据价值，在消费者洞察、产品研发、运营管理、市场营销、品牌战略方面，为管理决策提供科学依据。

舆情监测及分析

利用基于大数据的文本分析，我们可以清晰的知晓事件从始发到发酵期、发展期、高涨期、回落期和反馈期等阶段的演变过程，分析舆情的传播路径、传播节点、发展态势和受众反馈等情报。

了解用户反馈

通过基于大数据的文本分析，企业可以用正确的方式阅读用户散落在网络上的“声音”，企业可以直接读懂自己用户的想法，挖掘出用户对于产品/服务的情绪和态度。比如，大数据文本分析可以回答如下问题：

用户喜欢的是它产品的哪一方面？

比起其他公司的产品来，客户是否更倾向他的产品？

这些偏好会随着时间发展和变化吗？

大数据文本在互联网商业实践中体现其巨大商业价值的。

有文本分析参与的运营分析逻辑流程

用户评论进行文本分析的几大模块

从众多用户评论中提取出跟产品属性相关且有代表性的关键词，并进行相似文本聚类，然后给出相应权重，最后匹配相应的情绪属性和情绪值。最终的结果使得运营者拥有了用户视角，知晓产品有哪些地方表现尚可，哪些地方反映平平，而哪些地方是需要进行改进和完善的。

用户的对于Kindle Voyage评论的主要关注点

另一方面，运营方也要及时了解用户的负向反馈，针对性地改进服务质量。

在这里只出现了“压敏按键”这些负面词汇，点开“压敏按键”一词，看到用户的详细吐槽……“压敏按键在按下的时候出现下陷情形……”虽然这样的反馈不多，但仍要引起高度警惕，具体排查是产品本身的设计问题，还是极个别的产品质量问题，并将此反馈传递给相关负责部门。

地域评分和情绪分析

将用户的反馈数据和评分同步到地图上，通过形成的数据地图可以直观的看到各地对于Kindle Voyage的评价和情绪度，从而对整体的用户反馈情况进行监控，重点“关照”其中的用户差评“重灾区”。

此处选取的是Kindle Voyage在江苏省各市的用户反馈数据，反映出该地区KindleVoyage用户对于产品的评价及情绪倾向性。

其中，数值代表评分高低，颜色反映情绪值，绿色代表良好，蓝色是一般，红色代表处于警报状态，需要重点关注，排查问题发生的具体原因。

大数据文本分析对于产品、设计、营销和运营的巨大价值，它的重要性不亚于传统的结构性数据分析。用正确的方式阅读这些海量的文本数据，我们就可以直接读懂用户的想法，获得强有力的决策支持，从而使产品研发、营销推广和日常运营更贴近消费者需求，最终在用户心中形成良好的品牌形象。

2.文本挖掘技术

存储信息使用最多的是文本，所以文本挖掘被认为比数据挖掘具有更高的商业潜力，当数据挖掘的对象完全由文本这种数据类型组成时，这个过程就称为文本数据挖掘，事实上，最近研究表明公司信息有80%包含在文本文档中。

1）文本分类

文本分类指按照预先定义的主题类别，为文档集合中的每个文档确定一个类别。这样用户不但能够方便地浏览文档，而且可以通过限制搜索范围来使文档的查找更容易、快捷。目前，用于英文文本分类的分类方法较多，用于中文文本分类的方法较少，主要有朴素贝叶斯分类（Naïve Bayes），向量空间模型（Vector Space Model）以及线性最小二乘LLSF(Linear Least Square Fit)。

2）文本聚类

聚类与分类的不同之处在于，聚类没有预先定义好的主体类别，它的目标是将文档集合分成若干个簇，要求同一簇内文档内容的相似度尽可能的大，而不同簇之间的相似度尽可能的小。

3）文本结构分析

其目的是为了更好地理解文本的主题思想，了解文本表达的内容以及采用的方式，最终结果是建立文本的逻辑结构，即文本结构树，根结点是文本主题，依次为层次和段落。

4）Web文本数据挖掘

在Web迅猛发展的同时，不能忽视“信息爆炸”的问题，即信息极大丰富而知识相对匮乏。据估计，Web已经发展成为拥有3亿个页面的分布式信息空间，而且这个数字仍以每4-6个月翻1倍的速度增加，在这些大量、异质的Web信息资源中，蕴含着具有巨大潜在价值的知识。人们迫切需要能够从Web上快速、有效的发现资源和知识的工具。

文本挖掘目前面临的问题有挖掘算法的效率和可扩展性、遗漏及噪声数据的处理、私有数据的保护与数据安全性等。

文本挖掘不但要处理大量的结构化和非结构化的文档数据，而且还要处理其中复杂的语义关系，因此，现有的数据挖掘技术无法直接应用于其上。对于非结构化问题，一条途径是发展全新的数据挖掘算法直接对非结构化数据进行挖掘，对于数据非常复杂，导致这种算法的复杂性很高；另一条途径就是将非结构化问题结构化，利用现有的数据挖掘技术进行挖掘，目前的文本挖掘一般采用该途径进行。对于语义关系，则需要集成计算语言学和自然语言处理等成果进行分析。我们按照文本挖掘过程介绍其涉及的主要技术及其主要进展。

1）文本数据预处理技术

预处理技术包括Stemming（英文）/分词（中文）、特征表示和特征提取。与数据库中的结构化数据相比，文本具有有限的结构，或者根本就没有结构。此外，文档的内容是人类所使用的自然语言，计算机很难处理其语义。文本信息源的这些特殊性使得数据预处理技术在文本挖掘中更加重要。

(1)分词技术

在对文档进行特征提取前，需要先进行文本信息的预处理，对英文而言需要进行Stemming处理，中文的情况则不同，因为中文词与词之间没有固定的间隔符（空格），需要进行分词处理。目前主要有基于词库的分词算法和无词典的分词技术两种。

基于词库的分词算法包括正向最大匹配、正向最小匹配、逆向匹配及逐次遍历匹配法等。这类算法的特点是易于实现，设计简单；但分词的正确性很大程度上取决于所建的词库。因此基于词库的分词技术对于歧义和未登录词的切分有很大的困难。

基于无词典的分词技术的基本思想是：基于词频的统计，将原文中任意前后紧邻的两个字作为一个词进行出现频率的统计，出现的次数越高，成为一个词的可能性就越大，在频率超过某个预先设定的阈值时，就将其作为一个词进行索引。这种方法能够有效地提出未登录词。

(2)特征表示

文本特征指的是关于文本的元数据，分为描述性特征（如文本的名称、日期、大小、类型等）和语义性特征（如文本的作者、机构、标题、内容等）。特征表示是指一定特征项（如词条或描述）来代表文档，在文本挖掘时只需对这些特征项进行处理，从而实现对非结构化的文本处理。这是一个非结构化向结构化转换的处理步骤。特征表示的构造过程就是挖掘模型的构造过程。特征表示模型有多种，常用的有布尔逻辑型、向量空间模型（Vector Space Model，VSM）、概率型以及混合型等。W3C近来制定的XML，RDF等规范提供了对Web文档资源进行描述的语言和框架。

(3)特征提取

用向量空间模型得到的特征向量的维数往往会达到数十万维，如此高维的特征对即将进行的分类学习未必全是重要、有益的（一般只选择2%-5%的最佳特征作为分类数据），而且高维的特征会大大增加机器的学习时间，这便是特征提取所要完成的工作。

特征提取算法一般是构造一个评价函数，对每个特征进行评估，然后把特征按分值高低排队，预定数目分数最高的特征被选取。在文本处理中，常用的评估函数有信息增益（Information Gain）、互信息（Mutual Information）、文本证据权（The Weight of Evidence for Text）和词频。

2）文本挖掘分析技术

文本转换为向量形式并经特征提取后，便可以进行挖掘分析了。常用的文本挖掘分析技术有：文本结构分析、文本摘要、文本分类、文本聚类、文本关联分析、分布分析和趋势预测等。

(1)文本结构分析

其目的是为了更好地理解文本的主题思想，了解文本所表达的内容以及采用的方式。最终结果是建立文本的逻辑结构，即文本结构树，根节点是文本主题，依次为层次和段落。

(2)文本摘要

文本摘要是指从文档中抽取关键信息，用简洁的形式对文档内容进行解释和概括。这样，用户不需要浏览全文就可以了解文档或文档集合的总体内容。

任何一篇文章总有一些主题句，大部分位于整篇文章的开头或结尾部分，而且往往是在段首或段尾，因此文本摘要自动生成算法主要考察文本的开头、结尾，而且在构造句子的权值函数时，相应的给标题、子标题、段首和段尾的句子较大的权值，按权值大小选择句子组成相应的摘要。

(3)文本分类

文本分类的目的是让机器学会一个分类函数或分类模型，该模型能把文本映射到已存在的多个类别中的某一类，是检索或查询的速度更快，准确率更高。训练方法和分类算法是分类系统的核心部分。用于文本分类的分类方法较多，主要有朴素贝叶斯分类、向量空间模型、决策树、支持向量机、后向传播分类、遗传算法、基于案例的推理、K-最邻近、基于中心点的分类方法、粗糙集、模糊集以及线性最小二乘等。

厉宇航等指出传统特征提取的方法是基于词形的，并不考察词语的意义，忽略了同一意义下词形的多样性，不确定性以及词义间的关系，尤其是上下位关系。该文的方法在向量空间模型的基础上，以“概念”为基础，同时考虑词义的上位关系，使得训练过程中可以从词语中提炼出更加概括性的信息，从而达到提高分类精度的目的。

(4)文本聚类

文本分类是将文档归入到已经存在的类中，文本聚类的目标和文本分类是一样的，知识实现的方法不同。文本聚类是无教师的机器学习，聚类没有预先定义好的主题类别，它的目标是将文档集合分成若干个簇，要求同一簇内文档内容的相似度尽可能大，而不同簇间的相似度尽可能小。Hearst等人的研究已经证明了“聚类假设”，即与用户查询相关的文档通常会聚类的比较靠近，而远离与用户查询不相关的文档。

(5)关联分析

关联分析是指从文档集合中找出不同词语之间的关系。Feldman和Hirsh研究了文本数据库中关联规则的挖掘，提出了一种从大量文档中发现数千本在Amazon网站上找不到的新书籍；Wang Ke等以Web上的电影介绍作为测试文档，通过使用OEM模型从这些半结构化的页面中抽取词语项，进而得到一些关于电影名称、导演、演员、编剧的出现模式。

(6)分布分析与趋势预测

分布分析与趋势预测是指通过对文档的分析，得到特定数据在某个历史时刻的情况或将来的取值趋势。Feldman R等使用多种分布模型对路透社的两万多篇新闻进行了挖掘，得到主题、国家、组织、人、股票交易之间的相对分布，揭示了一些有趣的趋势。Wuthrich B等通过分析Web上出版的权威性经济文章对每天的股票市场指数进行预测，取得了良好的效果。

(7)可视化技术

数据可视化（Data Visualization）技术指的是运用计算机图形学和图像处理技术，将数据转换为图形或图像在屏幕上显示出来，并进行交互处理的理论、方法和技术。它涉及到计算机图形学、图像处理、计算机辅助设计、计算机视觉及人机交互技术等多个领域。国内外学者已经对信息可视化技术进行了大量的研究，运用最小张力计算、多维标度法、语义分析、内容图谱分析、引文网络分析及神经网络技术，进行了信息和数据的可视化表达。

3文本挖掘热点难点问题

显然，目标不同，文本挖掘的过程也不尽相同。但不论何种目标，都不可忽视如下几个方面的研究：

1)文本建模

向量空间模型，也称为“词袋”法，是目前文本处理的标准模式。简答讲，就是提取文本高频词构成特征向量来表达文本特征的方法，该方法有效描述了词与文档间的频率关系。面对复杂繁琐的自然语言文本，向量空间模型是目前最为简便有效的文本表示方法。

但向量空间模型建模方法最大的问题就是忽略了词在文本中承担的语法和语义上的作用，同时忽略了词与词之间的顺序关系，丢失了大量有用信息，从而减弱了高频词向量表达文本特征的可信度。同时，向量空间模型在处理真实文本数据时形成的特征向量的高维性也严重影响了后续文本挖掘的效率和结果的准确性。

此外，建模前的文本预处理工作作为整个文本挖掘过程的基础尤为重要。而不同的语言处理又常常不同。如何解决多语言混合如中英文混合情况下的文本处理和建模工作日益重要。同时，不同的语言有不同的切词处理方式。并且存在着大量多词同义、一词多义的现象。

2)特征降维

文本模型的高维特性制约了文本挖掘的效果。不论何种语种，由于语言本身的非结构特性以及建模后的高维特性，使得后续挖掘过程中都面临严重的效率问题。因此有效的降维是进行后续文本挖掘的重要一环。

目前的文本降维方法主要采用基于奇异值分解的潜在语义分析技术。该技术通过分析特征词之间的语义相关性来减少特征向量的维数，通过将词—文档的高维表示投影在低维潜在语义空间中，降低空间的维数，进而得到词—文档的不再稀疏的低维表示。并且，由词袋模型在进行奇异值分解后得到的子空间不再是仅仅反映出词汇出现的频率和分布关系，而进一步揭示了词汇或文档之间的语义联系。

3)挖掘算法的选择

模型创建成功并且进行了有效的降维处理之后，就可以进行具体的挖掘操作了。从狭义的角度理解，也可以说这部分才是真正的挖掘。而广义上来说，整个过程才构成文本挖掘的全部过程。

文本挖掘算法并不是一个新的领域，通常就是数据挖掘方法在文本数据上的应用。因此多数挖掘方法来自机器学习、统计学习、自然语言处理、信息抽取、信息检索以及知识管理等领域，最终目标就是对建模后的文本数据进行分析和处理，找到其中潜在的有用信息。

根据不同的应用目标，挖掘出的知识种类不尽相同，由此可以对文本挖掘的技术和算法进行如下的分类：如根据发现关联规则、聚类、趋势、差异等知识的不同，分别对应不同领域的算法选择。

任何算法技术的研究和设计都离不开开始实验的仿真和具体实例的验证。文本数据挖掘过程亦是如此。由于文本数据的复杂多样性，导致文本数据的挖掘过程相对其他结构化数据要复杂繁琐的多，对数据的敏感性更为严重，在很多情况下，面临对开放语料的实验结果不理想的问题。因此选择更好的评价方法，克服现有语料手工分类不准确带来的误差，以更好地对算法做出评价，同样重要。

4) 模型的理解及可视化表达

多数文本挖掘应用实例的目标同数据挖掘类似，通常是要辅助用户的决策和判断，因此从用户的角度来看，文本挖掘所发现结果的可理解至关重要。而对于各种方法挖掘出的模式、规则等结果，提高可理解性的解决方法通常有两种：一种是以生成人类易于理解的自然语言的方式进行呈现，如对文档进行摘要的方法；另一种方式则是以图形界面方式展示结果，通过提供相对少量的规则，利用计算机图形学、图像处理等可视化技术将结果更加直观的呈现给用户。

5)文本的预处理

大数据分析技术，不仅仅在于建模和算法，数据的质量至关重要，文本NLP也是如此。包括分词标注、实体抽取、词频统计、文本分类、情感分析、关键词提取、相关词分析、依存文法、简繁转换、自动注音和摘要提取等。有的文本挖掘采用很先进的神经网络结果并不理想，原因在于没有做好文本的预处理，文本中无用的字符、单词、符号没有排除，影响到效果。

文本需要先采用高质量的分词软件分词，比如jeba就不错，然后使用正则模式去除数字等非汉字符号，再移除停用词，去除低频率字词，然后用TF-IDF训练文本，用spark的hashingTF、IDF()分析字词权重，区别出key word，再对分词的实体词token字词使用word2vector算法根据文本字词的空间结构来获取token的vector向量表示，或用doc2vector根据上下文context来获取token的vector。然后采用贝叶斯算法、KNN、SVM等算法对token进行分类，再用LDA、KMEANS等算法进行聚类，提取文本的关键词，进行情绪识别，自动做主题归类，对文字进行画像，结合用户画像、情景模式完成智能推荐。还可采用tensorflow深度强化学习，提高处理的精度。今日头条的文本处理就是以上原理，

在新媒体时代异军突起，笑傲江湖。

近年来，可视化技术作为展示结果的关键一环逐渐成为文本挖掘过程中日益重要的一个分支。大量的研究结合语义分析、内容图谱分析、最小张力计算、神经网络技术、多维标度法等数据分析和处理方法进行了结果的可视化表达。

四、 常用算法

大数据有许多数据挖掘的经典算法，涉及到了决策分类，聚类，回归、链接挖掘，关联挖掘，模式挖掘等等方面。其中经典十大算法为：C4.5，K-Means，SVM，Apriori，EM，PageRank，AdaBoost，KNN，NB和CART。一般数据挖掘算法分为两种，有监督和无监督算法，其中有监督算法主要有逻辑回归、决策树、神经网络等，无监督学习主要包括聚类、最邻近距离、支持向量机等。一般需要综合运用多种算法进行分析，运用较多算法的有逻辑回归、SVM、K-Means、NLP自然语言理解等。

1.C4.5算法。C4.5算法与ID3算法一样，都是数学分类算法，C4.5算法是ID3算法的一个改进。ID3算法采用信息增益进行决策判断，而C4.5采用的是增益率。

详细介绍链接：http://blog.csdn.NET/androidlushangderen/article/details/42395865

2.CART算法。CART算法的全称是分类回归树算法，他是一个二元分类，采用的是类似于熵的基尼指数作为分类决策，形成决策树后之后还要进行剪枝，我自己在实现整个算法的时候采用的是代价复杂度算法，

详细介绍链接：http://blog.csdn.net/androidlushangderen/article/details/42558235

2. KNN(K最近邻)算法。给定一些已经训练好的数据，输入一个新的测试数据点，计算包含于此测试数据点的最近的点的分类情况，哪个分类的类型占多数，则此

3. 测试点的分类与此相同，所以在这里,有的时候可以复制不同的分类点不同的权重。近的点的权重大点，远的点自然就小点。

详细介绍链接：http://blog.csdn.Net/androidlushangderen/article/details/42613011

4.Naive Bayes(朴素贝叶斯)算法。朴素贝叶斯算法是贝叶斯算法里面一种比较简单的分类算法，用到了一个比较重要的贝叶斯定理，用一句简单的话概括就是条件概率的相互转换推导。

详细介绍链接：http://blog.csdn.net/androidlushangderen/article/details/42680161

5.SVM(支持向量机)算法。支持向量机算法是一种对线性和非线性数据进行分类的方法，非线性数据进行分类的时候可以通过核函数转为线性的情况再处理。其中的一个关键的步骤是搜索最大边缘超平面。

详细介绍链接：http://blog.csdn.net/androidlushangderen/article/details/42780439

6.EM(期望最大化)算法。期望最大化算法，可以拆分为2个算法，1个E-Step期望化步骤,和1个M-Step最大化步骤。他是一种算法框架，在每次计算结果之后，逼近统计模型参数的最大似然或最大后验估计。

详细介绍链接：http://blog.csdn.net/androidlushangderen/article/details/42921789

7.Apriori算法。Apriori算法是关联规则挖掘算法，通过连接和剪枝运算挖掘出频繁项集，然后根据频繁项集得到关联规则，关联规则的导出需要满足最小置信度的要求。

详细介绍链接：http://blog.csdn.net/androidlushangderen/article/details/43059211

8.FP-Tree(频繁模式树)算法。这个算法也有被称为FP-growth算法，这个算法克服了Apriori算法的产生过多侯选集的缺点，通过递归的产生频度模式树，然后对树进行挖掘，后面的过程与Apriori算法一致。

详细介绍链接：http://blog.csdn.net/androidlushangderen/article/details/43234309

9.PageRank(网页重要性/排名)算法。PageRank算法最早产生于Google,核心思想是通过网页的入链数作为一个网页好快的判定标准，如果1个网页内部包含了多个指向外部的链接，则PR值将会被均分，PageRank算法也会遭到Link Span攻击。

详细介绍链接：http://blog.csdn.net/androidlushangderen/article/details/43311943

10.HITS算法。HITS算法是另外一个链接算法，部分原理与PageRank算法是比较相似的，HITS算法引入了权威值和中心值的概念，HITS算法是受用户查询条件影响的，他一般用于小规模的数据链接分析，也更容易遭受到攻击。

详细介绍链接：http://blog.csdn.net/androidlushangderen/article/details/43311943

11.K-Means(K均值)算法。K-Means算法是聚类算法，k在在这里指的是分类的类型数，所以在开始设定的时候非常关键，算法的原理是首先假定k个分类点，然后根据欧式距离计算分类，然后去同分类的均值作为新的聚簇中心，循环操作直到收敛。

详细介绍链接：http://blog.csdn.net/androidlushangderen/article/details/43373159

12.BIRCH算法。BIRCH算法利用构建CF聚类特征树作为算法的核心，通过树的形式，BIRCH算法扫描数据库，在内存中建立一棵初始的CF-树，可以看做数据的多层压缩。

详细介绍链接：http://blog.csdn.net/androidlushangderen/article/details/43532111

13.AdaBoost算法。AdaBoost算法是一种提升算法，通过对数据的多次训练得到多个互补的分类器，然后组合多个分类器，构成一个更加准确的分类器。

详细介绍链接：http://blog.csdn.net/androidlushangderen/article/details/43635115

14.GSP算法。GSP算法是序列模式挖掘算法。GSP算法也是Apriori类算法，在算法的过程中也会进行连接和剪枝操作，不过在剪枝判断的时候还加上了一些时间上的约束等条件。

详细介绍链接：http://blog.csdn.net/androidlushangderen/article/details/43699083

15.PreFixSpan算法。PreFixSpan算法是另一个序列模式挖掘算法，在算法的过程中不会产生候选集，给定初始前缀模式，不断的通过后缀模式中的元素转到前缀模式中，而不断的递归挖掘下去。

详细介绍链接：http://blog.csdn.net/androidlushangderen/article/details/43766253

16.CBA(基于关联规则分类)算法。CBA算法是一种集成挖掘算法，因为他是建立在关联规则挖掘算法之上的，在已有的关联规则理论前提下，做分类判断，只是在算法的开始时对数据做处理，变成类似于事务的形式。

详细介绍链接：http://blog.csdn.net/androidlushangderen/article/details/43818787

17.RoughSets(粗糙集)算法。粗糙集理论是一个比较新颖的数据挖掘思想。这里使用的是用粗糙集进行属性约简的算法，通过上下近似集的判断删除无效的属性，进行规制的输出。

详细介绍链接：http://blog.csdn.net/androidlushangderen/article/details/43876001

18.gSpan算法。gSpan算法属于图挖掘算法领域。，主要用于频繁子图的挖掘，相较于其他的图算法，子图挖掘算法是他们的一个前提或基础算法。gSpan算法用到了DFS编码，和Edge五元组，最右路径子图扩展等概念，算法比较的抽象和复杂。

19神经网络算法。BP（Back Propagation）神经网络是1986年由Rumelhart和McCelland为首的科学家小组提出，是一种按误差逆传播算法训练的多层前馈网络，是目前应用最广泛的神经网络模型之一。BP网络能学习和存贮大量的输入-输出模式映射关系，而无需事前揭示描述这种映射关系的数学方程。它的学习规则是使用最速下降法，通过反向传播来不断调整网络的权值和阈值，使网络的误差平方和最小。BP神经网络模型拓扑结构包括输入层（input）、隐层(hidden layer)和输出层(output layer)。

20线性回归，logistic回归，泊松回归

算法类似性

根据算法的功能和形式的类似性，我们可以把算法分类，比如说基于树的算法，基于神经网络的算法等等。当然，机器学习的范围非常庞大，有些算法很难明确归类到某一类。而对于有些分类来说，同一分类的算法可以针对不同类型的问题。这里，我们尽量把常用的算法按照最容易理解的方式进行分类。

回归算法

回归算法是试图采用对误差的衡量来探索变量之间的关系的一类算法。回归算法是统计机器学习的利器。在机器学习领域，人们说起回归，有时候是指一类问题，有时候是指一类算法，这一点常常会使初学者有所困惑。常见的回归算法包括：最小二乘法（Ordinary Least Square），逻辑回归（LogisticRegression），逐步式回归（Stepwise Regression），多元自适应回归样条（Multivariate Adaptive Regression Splines）以及本地散点平滑估计（Locally Estimated Scatterplot Smoothing）.

基于实例的算法

基于实例的算法常常用来对决策问题建立模型，这样的模型常常先选取一批样本数据，然后根据某些近似性把新数据与样本数据进行比较。通过这种方式来寻找最佳的匹配。因此，基于实例的算法常常也被称为“赢家通吃”学习或者“基于记忆的学习”。常见的算法包括 k-Nearest Neighbor(KNN), 学习矢量量化（Learning Vector Quantization， LVQ），以及自组织映射算法（Self-Organizing Map ， SOM）

正则化方法

正则化方法是其他算法（通常是回归算法）的延伸，根据算法的复杂度对算法进行调整。正则化方法通常对简单模型予以奖励而对复杂算法予以惩罚。常见的算法包括：Ridge Regression， Least Absolute Shrinkageand Selection Operator（LASSO），以及弹性网络（Elastic Net）。

决策树学习

决策树算法根据数据的属性采用树状结构建立决策模型， 决策树模型常常用来解决分类和回归问题。常见的算法包括：分类及回归树（Classification And Regression Tree， CART）， ID3 (Iterative Dichotomiser 3)， C4.5， Chi-squared Automatic Interaction Detection(CHAID), Decision Stump,随机森林（Random Forest）， 多元自适应回归样条（MARS）以及梯度推进机（Gradient Boosting Machine， GBM）

贝叶斯方法

贝叶斯方法算法是基于贝叶斯定理的一类算法，主要用来解决分类和回归问题。常见算法包括：朴素贝叶斯算法，平均单依赖估计（Averaged One-Dependence Estimators， AODE），以及Bayesian Belief Network（BBN）。

基于核的算法

基于核的算法中最著名的莫过于支持向量机（SVM）了。基于核的算法把输入数据映射到一个高阶的向量空间， 在这些高阶向量空间里， 有些分类或者回归问题能够更容易的解决。 常见的基于核的算法包括：支持向量机（Support Vector Machine， SVM）， 径向基函数（Radial Basis Function ，RBF)， 以及线性判别分析（Linear Discriminate Analysis ，LDA)等。SVM适用于数据量小、维度高、数据不相关的场景。

聚类算法

聚类，就像回归一样，有时候人们描述的是一类问题，有时候描述的是一类算法。聚类算法通常按照中心点或者分层的方式对输入数据进行归并。所以的聚类算法都试图找到数据的内在结构，以便按照最大的共同点将数据进行归类。常见的聚类算法包括 k-Means算法以及期望最大化算法（Expectation Maximization， EM）。

关联规则学习

关联规则学习通过寻找最能够解释数据变量之间关系的规则，来找出大量多元数据集中有用的关联规则。常见算法包括 Apriori算法和Eclat算法等。

AI人工神经网络

人工神经网络算法模拟生物神经网络，是一类模式匹配算法。通常用于解决分类和回归问题。人工神经网络是机器学习的一个庞大的分支，有几百种不同的算法。（其中深度学习就是其中的一类算法，我们会单独讨论），重要的人工神经网络算法包括：感知器神经网络（Perceptron Neural Network）, 反向传递（Back Propagation）， Hopfield网络，自组织映射（Self-Organizing Map, SOM）。学习矢量量化（LearningVector Quantization， LVQ）,DNN、CNN、RNN、GAN等人工智能算法。

深度学习

深度学习算法是对人工神经网络的发展。在近期赢得了很多关注， 特别是百度也开始发力深度学习后， 更是在国内引起了很多关注。 在计算能力变得日益廉价的今天，深度学习试图建立大得多也复杂得多的神经网络。很多深度学习的算法是半监督式学习算法，用来处理存在少量未标识数据的大数据集。常见的深度学习算法包括：受限波尔兹曼机（Restricted Boltzmann Machine， RBN）， Deep Belief Networks（DBN），卷积网络（Convolutional Network）, 堆栈式自动编码器（Stacked Auto-encoders）。

降低维度算法

像聚类算法一样，降低维度算法试图分析数据的内在结构，不过降低维度算法是以非监督学习的方式试图利用较少的信息来归纳或者解释数据。这类算法可以用于高维数据的可视化或者用来简化数据以便监督式学习使用。常见的算法包括：主成份分析（Principle Component Analysis， PCA），奇异值分解SVD、偏最小二乘回归（Partial Least Square Regression，PLS）， Sammon映射，多维尺度（Multi-DimensionalScaling, MDS）, 投影追踪（ProjectionPursuit）等。

集成算法

集成算法用一些相对较弱的学习模型独立地就同样的样本进行训练，然后把结果整合起来进行整体预测。集成算法的主要难点在于究竟集成哪些独立的较弱的学习模型以及如何把学习结果整合起来。这是一类非常强大的算法，同时也非常流行。常见的算法包括：Boosting， Bootstrapped Aggregation（Bagging）， AdaBoost，堆叠泛化（Stacked Generalization， Blending），梯度推进机（Gradient Boosting Machine, GBM），随机森林（RandomForest）。

http://xinsheng.huawei.com/cn/blog/detail\_79549.html