数学之美

- 文字和语言，信息和数字
  - 翻译能达成的条件
    - 不同的文字系统在记录信息上的能力是等价的
    - 文字只是载体，而非信息本身
  - 罗塞塔石碑的指导意义
    - 信息冗余是信息安全的保障
    - 语言的数据，即语料，尤其是双语或者多语语料对翻译至关重要，是我们从事机器翻译研究的基础
  - 罗马数字，小数字出现在大数字前为减，右边为加。如IV 5-1 VII 5+2
  - 从象形文字到楔形文字，从具体到抽象。常用字笔画少，生僻字笔画多，完全符合信息论中的最短编码原理；
- 自然语言处理——从规则到统计
  - 图灵测试，验证机器是否有智能
    - 让人和机器进行交流，如果人无法判断自己交流的对象是人还是机器，就说明这个机器有智能了
  - 误区
    - 20世纪50年代到70年代
    - 基于规则的语言处理
    - 原因
      - 因为陷入了人对于自然语言的认知
      - 认为想让机器完成翻译或者语音识别，必须先让计算机理解自然语言，而这必须先让计算机有智能
      - 后来被称为"鸟飞派"
      - 我们只需要理解空气动力学，就能像鸟一样飞，而不是依靠仿生学
  - 1970年之后，基于统计学的处理方法诞生
    - 近30年取得突破性的发展
  - 但是这两者之争扔持续了15年
    - 新生的需要一定时间成熟
    - 必须等原有的一批看科学家退休
- 统计语言模型
  - 马尔科夫模型
    - 只和前一个单词有关的为二元模型，和前面n个相关的为n元模型
    - 一般为3元模型，Google采用4元模型
- 分词
  - 对于西方来说，词之间有分界符。一些亚洲语言，词之间没有明显的分界符。
  - 因此需要对句子进行分词，才能做进一步自然语言处理
  - 现在由于手机平板的出现，西方国家也有手写体，这时就需要借鉴中文分词法

- 隐含马尔科夫模型
  - 通信的本质是编解码和传输的过程
  - 雅各布森通信六个要素
    - 发送者(信息源)
    - 信道
    - 接收者，
    - 信息
    - 上下文
    - 编码
  - 19世纪，概率论的发展从对(相对静态的)随机变量的研究发展到对随机变量的时间序列s1,s2,s3,...,st,...，即随机过程的研究
  - 隐含马尔可夫模型最早的成功应用是语音识别。李开复博士坚持采用隐含马尔可夫模型的框架，成功研发出世界上第一个大词汇量连续语音识别系统Sphinx
  - 隐含马尔可夫模型有三个基本问题
    - 给定一个模型，如何计算某个特定的输出序列的概率(Forward-Backword算法)
    - 给定一个模型和某个特定的输出序列，如何找到最可能产生这个输出的状态序列(维特比算法)
    - 给定足够量的观测数据，如何估计隐含马尔可夫模型的参数(鲍姆-韦尔奇算法)
  - 隐含马尔可夫模型最初应用于通信领域，继而推广到语音和语言处理中，成为连接自然语言处理和通信的桥梁。也是机器学习的主要工具之一。和所有的机器学习的模型一样，他需要一个训练算法(鲍姆-韦尔奇算法)和使用时的解码算法(维特比算法)
- PageRank
  - 将互联网看成一个整体
  - 一个网页被很多其他网页所链接，说明他受到承认和依赖，那么他的排名就高
  - 权重的想法应该来自佩奇，破除后面迭代问题来自布林
  - 后来出现并行计算工具, MapReduce,由原来的半自动化变成全自动化
- TF-IDF
  - 影响搜索质量的因素
    - 完备的索引
    - 对网页质量的度量， 如pageRank
    - 用户偏好
    - 确定一个网页和某个查询的相关性
  - TF-IDF
    - 单文本词频/逆文本频率指数
    - 为每一个词给一个权重，这个权重设定满足的条件
      - 一个词预测主题的能力越强，权重越大。如原子能＞应用
      - 停用词的权重为0
    - 最早是由剑桥大学 斯巴克·琼斯提出的，但是并没有解释为啥是对数函数而不是其他函数。后来剑桥大学的罗宾逊也解释，但是没有说清。后来康奈大学的萨尔顿解释清楚了这个在信息检索中的用途
- 信息的度量和作用
  - 信息熵
    - 一条信息的信息量与其不确定性有些直接的关系

        - 信息熵就是度量信息，量化信息作用
    - 不同语言的冗余度差别很大，而汉语在所有语言中冗余度是相对小的
    - 信息的作用——消除不确定性
        - 如果没有信息，任何公式或者数字的游戏都无法排除不确定性
        - 当引入相关信息，就能减少不确定性
        - 从搜索引擎上来看
            - 如果只给定一个信息"中国"，那么会有很多相关信息，你并不能确定哪些是你需要的，但是再多引入一些信息，那么可能在页面的前几个就是你需要的信息了
    - 互信息
        - 香农在信息论中提出了这个概念作为两个随机事件"相关性"的量化度量
- 贾里尼克和现代语言处理
    - 年少教育
        - 小学生和中学生其实没必要花那么多时间读书，而他们的社会经验，生活能力以及在那时树立起的志向将帮助他们的一生
        - 大学以后理解能力增强，很多中学努力学的东西，可能只需要十之一二的时间就能学会
        - 学习是终生的事
        - 书本的内容可以早学，我可以晚学，但是错过了成长阶段却是无法补回来的
    - 贾里尼克在约翰·霍普金斯大学建立了世界著名的CLSP实验室
    - 做事严谨，学生淘汰率高，但是却为每一个学生争取最大便利。闲不住，70多岁仍每天按时上班
- 布尔代数和搜索引擎
    - 搜索 原子能 应用，不包含原子弹
        - 将前两个搜索结果与，并且和第三个结果非。得到搜索结果
    - 布尔代数将逻辑和数学合二为一，给了一个看待世界的全新视角
- 有限状态机和动态规划
    - 解决输入路径匹配，自动识别输入地址
        - 能识别错别字
        - 能识别不太标准的语句
    - 动态规划解决地图中最短路径
- GoogleAK-47的设计者阿米特·辛格博士
    - 简单即是真理
    - 要求对于搜索质量提高的改进方法需要能说清楚理由，必须能对机器学习出来的参数和公式给出合理的解释，否则不能上线
    - 先解决用户80%的问题，再解决剩下的20%
- 余弦定理
    - 对文本进行分类
        - 计算文章中所有实词的TF-IDF值，
        - 将这些词按照对应的实词在词汇表中的位置依次排列，得到一个向量
        - 利用余弦定理计算两个向量夹角
        - 夹角越小，两个文章越相似

- 矩阵运算和文本处理的两个分类问题
  - 比余弦定理时间短的粗分类——奇异值分解
    - 一开始将所有信息列入一个矩阵，每一个值表示TF-IDF
    - 将一个矩阵分解为三个小矩阵相乘
    - 第一个矩阵的每一行表示一个词，每一列表示一个语义相近的词
    - 最后一个矩阵是对文本分类的结果，每一列对应一篇文章，每一行对应一个主题
    - 中间的矩阵表示词的类和文章的类之间的相关性
    - 并行算法由后来Google中国的张智威博士带领的团队解决
  - 一般先进行奇异值分解粗分类，再进行余弦定理细分类
- 图论和网络爬虫
  - 图论的起源可以追溯到欧拉时代
    - 论证了如果一个图能够从一个顶点出发，每条边不重复地遍历回到这个顶点，那么每一个顶点的度必须为偶数
  - 网络爬虫大致的细节
    - 使用BFS还是DFS，大部分用BFS
    - 页面的分析和URL的提取
    - 记录哪些网页已经下载过的小本本——URL表
- 信息指纹及其应用
  - 任何一段信息（图像，语音，文字），都可以对应一个不太长的随机数，作为区别这段信息和其他信息的指纹
  - 产生的这段指纹，只要算法设计得好，任意两段信息的指纹都很难重复
  - 使用伪随机数产生算法（MD5, SHA-1），将其生成为一段定长（128或160)的随机二进制数
  - 用途
    - 网页搜索查询两个用词是否完全相同
    - 判断两个网页是否基本相同，是否存在抄袭
    - YouTube反作弊
- 密码学的基本原理
  - 可追溯至两千年前，凯撒使用密码传送情报
  - 根据信息论，密码学的最高境界是敌方在截取密码后，对我方的所知没有任何增加，用信息论的术语就是信息量没有增加
  - 一般来讲，密码之间分布均匀并且统计独立时，提供的信息最少
  - 公开密钥的好处
    - 简单
    - 可靠
    - 灵活
- 搜索引擎的反作弊和搜索结果权威性
  - 搜索中的作弊可以理解为通信中的噪音
  - 通信中解决噪音干扰的模型在搜索反作弊依然适用
    - 从信息源出发，加强通信（编码）自身的抗干扰能力
      - 利用余弦定理，计算网站的出链向量
    - 从传输来看，过滤掉噪音，还原信息

        - 图论。
            - 如果有几个节点两两互相都连接在一起，他们被称为一个Clique。因为需要这样提高自己的排名
    - 搜索结果的权威性
        - pageRank只能从链接的网络质量数量来判断网页内容质量。但是像很多八卦网站，内容未必权威
        - 如何度量
            - 引入一个新的概念 "提及"(Mention)
                - 但是这个隐含在文章的自然语句中，需要通过自然语言处理方法分析，即使有好的算法，计算量仍然很大
                - 而且提及的组织，机构必须是和主题相契合的
                - 拥有云计算大数据技术，计算权威才有可能
            - 计算权威的步骤
                - 对每一个网页正文(包括标题)中的每一句话进行句法分析，然后找出涉及到主题的短语，以及对信息源的描述
                - 利用互信息，找到主题短语和信息源的相关性
                - 需要对主题短语进行聚合，用矩阵运算的方法
                - 需要对一个网站中的网页进行聚合，比如吧一个网站下面的网页按照子域或者子目录进行聚合
- 数学模型的重要性
    - 对于很多问题，完美的数学模型应当是最简单的
    - 托勒密
        - 2000多年前的罗马时代
        - 计算出诸多天体运行轨迹，于天文学作用堪比欧几里得之于几何学，牛顿之于物理学
        - 但是他提出的是地心说
        - 贡献
            - 发明球坐标
            - 定义了包括赤道和零度经线在内的经纬线
            - 提出了黄道
            - 发明了弧度制
            - 对地心说模型的完善
                - 采用40–60个在大圆上面套小圆的方法，精确计算出所有行星的运行轨迹
            - 制定了儒略历
                - 每年为365天，每四年增加一个闰年
                - 后来经过1500年，误差多出了10天，最后教皇格里高利在儒略历的基础上删除10天，并将每世纪最后一年的闰年改成平年
    - 结论
        - 正确的数学模型应当在形式上是简单的(托勒密显然太复杂)
        - 一个正确的模型一开始可能不如一个精雕细琢过的错误模型(哥白尼的日心说)
        - 大量准确的数据对研发很重要
        - 正确的模型可能受噪音干扰，而显得不准确；这时不应该用一种凑合的修正方法加以弥补，而是要找到噪音的根源，也许能通往重大的发现

- 最大熵模型
  - 原理
    - 对一个随机事件的概率分布进行预测时，我们的预测应当满足全部已知的条件，而对未知的情况不要做任何主观假设
    - 不要把所有鸡蛋放在一个篮子里。当我们我们遇到不确定性时，就要保留各种可能性
  - 最大熵模型在形式上是最漂亮、最完美的统计模型。
  - 应用
    - 自然语言处理
      - 拉纳帕提做出来当时世界上最好的词性标识系统和句法分析器
    - 金融
      - 贾里尼克和达拉·皮垂兄弟还有做语音识别系统的同事到文艺复兴技术公司
      - 1988年创立至今，该基金的净回报率高达34%；08年金融危机，全球股市暴跌，回报率高达80%
- 拼音输入法的数学原理
  - 早期汉字输入法
    - 使用拼音，后来改成双拼
    - 虽然敲击的次数少了，但是严重影响思维。而且很难记住
  - 后来，使用拼音
    - 不需要专门练习
    - 输入自然，不会中断思维，找键时间短
    - 因为编码长，有信息冗余量，容错性好
  - 香农第一定理
    - 对于一个信息，任何编码的长度都不小于他的信息熵
  - 拼音转汉字的算法
    - 动态规划
    - 与导航相似
    - 每一个音节可以对应多少个汉字，把一个拼音串对应的汉字从左到右连起来，就是一张有向图，她被称为网格图
    - 拼音输入法就是要根据上下文在给定拼音条件下找到一个最优的句子，对应到图中就是要找从起点到终点的一条最短路径
- 自然语言处理的教父马库斯
  - 运用自己的影响力，推动自然科学基金会和DARPA出资立项，联络了多所大学和研究机构，建立了数百个标准的语料库组织(LDC)
  - 他的影响力很大一部分是靠弟子传播出去的
    - 管理相对宽松，所以很多弟子性格迥异。凭借自己的经验和见识，避免自己的学生做无用功
    - 迈克尔·柯林斯
      - 追求完美
      - 做出了世界上最好的分析器
    - 艾里克·布莱尔
      - 简单才美

        - 基于变换规则的机器学习方法
    - 大卫·雅让斯基
    - 拉纳帕提
- 布隆过滤器
    - 一个很长的二进制向量和一系列随机映射函数
    - 步骤
        - 先建立一个很长的二进制向量，然后用随机数产生器产生8个信息指纹，
        - 再用一个随机数产生器把这8个信息指纹映射到二进制比特位中的8个自然数，
        - 把这8个位置的比特位全部设置为1
        - 之后再用相同的随机数产生器对邮件地址进行比对
    - 特点
        - 不会漏掉黑名单中的任何一个
        - 但是会误杀
- 贝叶斯
    - 贝叶斯是马尔科夫链的拓展，马尔科夫链是贝叶斯的特例
    - 计算概率
    - 谷歌工程师利用贝叶斯建立了 Phil cluster
        - 前期数据不够，模型不好
        - 后期改进，并更名为Rephil
- 条件随机场、文法分析器
    - 文法分析
        - 演变
            - 由一开始基于规则变成基于统计
            - 尤金·查尼亚克搭建了桥梁
            - 后来马库斯的高足拉纳帕提又建立了新的统计方法
    - 条件随机场
        - 条件随机场是马尔可夫模型的一种拓展
    - 条件随机场应用
        - 文法分析器
        - 预防犯罪
        - 机器学习
        - 模式识别
- 维特比和维特比算法
    - 维特比
        - 创立了高通公司，3G的CDMA
        - 发明维特比算法
    - 维特比算法
        - 动态规划算法
    - CDMA技术
        - 3G移动通信的基础
        - 之前的频分多址和时分多址

        - 因为是根据密码进行加密和解密，所以被称为码分多址
        - 只需要通过密码过滤掉无法解码的信号，留下和自己密码对应的信号即可
- 期望最大化算法——上帝的算法
    - 步骤
        - 随机选取k个点，作为起始的中心
        - 计算所有点到这些中心的距离，将这些点归到最近的一类中
        - 重新计算每一类的中心
        - 重复上述过程，直到每次新的中心和旧的中心之间的偏移非常小，即过程收敛
    - 如果是凸函数，则能找到全局最优
    - 但是文本分类中的余弦距离都不保证是凸函数，因此有可能给出的是局部最优
- 逻辑回归和搜索广告
    - 搜索广告三个阶段
        - 早期Overture和百度的广告系统，价高者得
        - 谷歌的CTR，关键技术是点击率预估
        - 进一步的全局优化
    - 逻辑回归
        - 采用逻辑回归函数，自变量无穷，值域零到壹
        - 拥有一个常量和k个自变量，自己k个自变量系数
        - 是一种将影响概率的不同因素结合到一起的指数函数
- 云计算
    - google采用MapReduce，
    - 本质上是分治的思想
- Google大脑和人工神经网络
    - 人工神经网络和贝叶斯差不多
        - 相同点
            - 都是有向图，遵循马尔科夫链假设
            - 训练方法相似
            - 模式分类上，这两种方法在效果上相似，很多用神经网络能解决的，贝叶斯网络也能解决，反之亦然
            - 计算量大
        - 不同点
            - 人工神经网络完全标准化，没有贝叶斯灵活
            - 贝叶斯更容易考虑上下文前后的关系，人工神经网络相对孤立
                - 因此贝叶斯可以用来解码一个输入序列，比如讲一段语音识别成文字，或者将英语句子翻译成中文
                - 人工神经网络可以识别一个个字，但是很难处理一个序列。因此主要用于估计一个概率模型的参数
    - Google大脑采用人工神经网络的原因
        - 算法稳定
        - 有很好的通用性
        - 容易并行化运算

- 大数据的威力
  - 数据的重要性
    - 从有文明就有数据，以前是对过往经验的总结，得出的结论
    - 数据得出的结论有时候和我们"以为"有很大出入
  - 大数据的"大"
    - 不只是数据量大，更多的是维度大还有完备性
  - 案例
    - Google仅做了一年的自然语言处理就领先其他公司很多年。算法没有多大改进，就是拥有大数据进行分析训练
    - Google的搜索前期比Bing好，因为数据集大，便于总结分析。百度搜索比搜狗有道好也是同样道理
  - 为什么需要大数据
    - 除了最热的IT行业，现在医疗技术很多都基于大数据，比如分析基因缺陷是否会导致疾病的概率升高。为每个病人个性化制作抗癌药物
    - 其他行业也会越来越依赖大数据