知识图谱与复杂网络概念

Knowledge Graph & Complex Networks

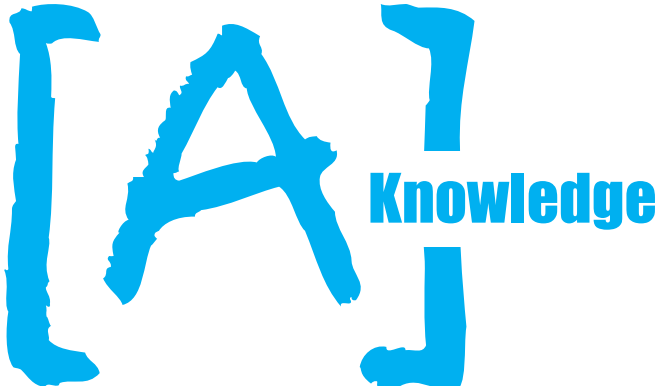
赵亚伟

zhaoyw@ucas.ac.cn

中国科学院大学 大数据分析技术实验室

目录

CONTENTS



知识图谱

复杂网络

节点重要性和相似性

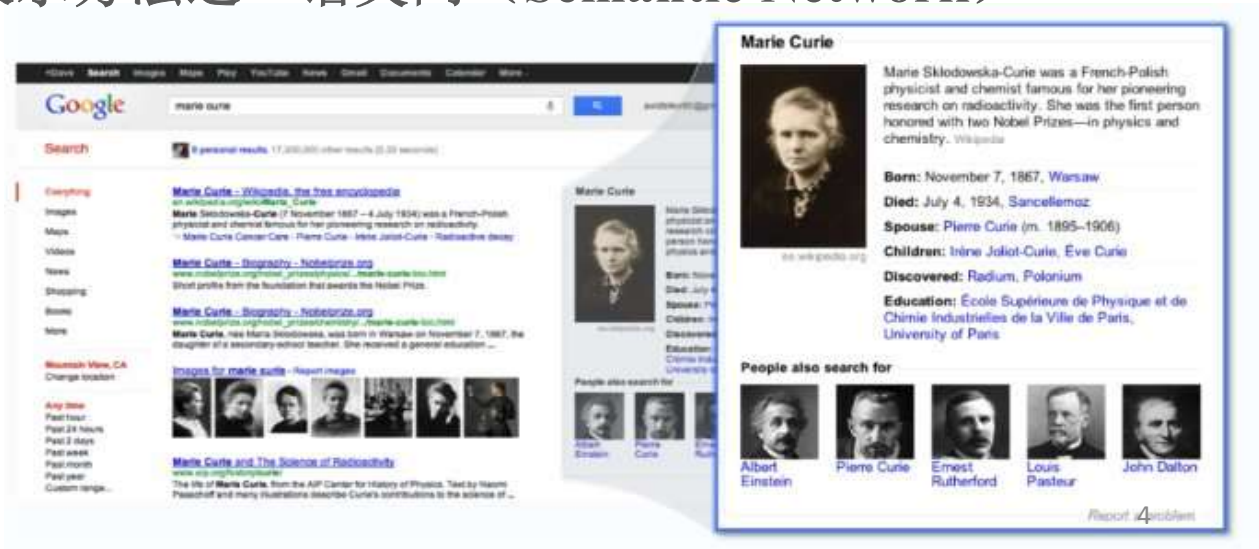
预测算法及模型

案例分析

什么是知识图谱?

- 2012年5月，搜索引擎巨头谷歌在它的搜索页面中首次引入“知识图谱（Knowledge Graph）”：
  - 用户除了得到搜索网页链接外，还将看到与查询词有关的更加智能化的答案
  - 如图所示，当用户输入“Marie Curie”（玛丽·居里）这个查询词，谷歌会在右侧提供了居里夫人的详细信息，如个人简介、出生地点、生卒年月等，甚至还包括一些与居里夫人有关的历史人物，例如爱因斯坦、皮埃尔·居里（居里夫人的丈夫）等。
- 知识图谱源于人工智能的知识表示方法之一语义网（Semantic Network）

目前尚无统一定义



什么是知识图谱?

- 谷歌原高级副总裁艾米特·辛格博士一语道破知识图谱的重要意义所在：
  - “构成这个世界的是实体，而非字符串（**things, not strings**）”

> Google announced its “knowledge graph” today and describes it as “an intelligent model—in geek-speak, a ‘graph’ — that understands real-world entities and their relationships to one another: things, not strings. …

- 谷歌知识图谱一出激起千层浪，美国的微软必应，中国的百度、搜狗等搜索引擎公司在短短的一年内纷纷宣布了各自的“知识图谱”产品，如百度“知心”、搜狗“知立方”等。

百度图谱，知立方


章子怡
章子怡（Zhang Ziyi），1979年2月9日出生于北京，电影演员，2000年毕业于中... 更多 >
Ta的那些事儿
牵手继女过情人节
告别片段曝光
晒女儿满月照
当妈后复出
怀孕损失千万
母女包抢眼


明星人气榜
张学友 NO.66
10543
去拉票：
张学友的关系图谱
亲情
友情
爱情
查看全部
妻子
干爹
干爹
干妹妹
干妈

换一个角度看问题

- 知识图谱为人们提供了另一个视角看问题
- 二维世界生物与三维世界生物
  - 知识图谱提供了一种可以从全局看待问题的方法

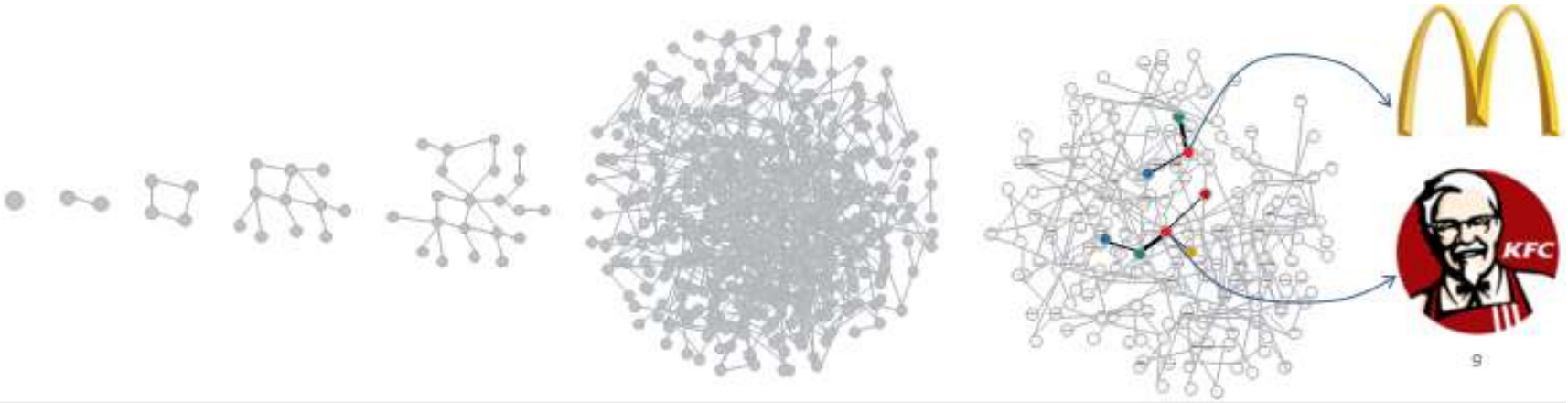
淘宝村的故事

- 从一夜暴富到萧条平淡 山东曹县淘宝村经历了什么？
  - 在山东曹县大集镇，至今流传着这样一个传说：一个村子里的小伙相亲，兄弟们开车助阵，结果，头车出了村，尾车还没进村，阵势十分壮观。姑娘一看，当即就答应了，据说，这个传说就发生在，大集镇东北两公里外的张庄，那个曾因有人一夜暴富，而名动江湖的“中国淘宝村”。

- 然而，如今这些门头无一例外全部关门歇业。街上空无一人，偶尔能看到一两个下地干农活的村民。

淘宝村的故事Cont.

- 电商平台上的各商铺之间是相互影响的，所谓“市场就那么大”，新的商铺开张必然对其他商铺造成冲击
- 销售风险（或机会）在商铺图谱中具有传播效应，商铺的加入和撤出都会对整个网络产生影响
- 几乎所有商户都想通过流量（出货量）的增长最终实现业绩的增长，成为“龙头”商铺，最终形成几个商铺独大局面，而这些“龙头”商铺之间也具有很强的关联性，如同麦当劳和肯德基



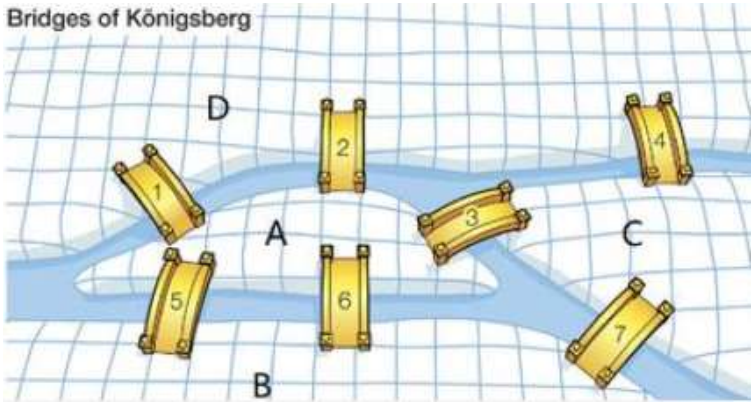
知识图谱（Knowledge Graph）是图（Graph）吗?

知识图谱是网络吗?

建立知识图谱的工作意义本身远大于其科学概念

哥尼斯堡七桥问题

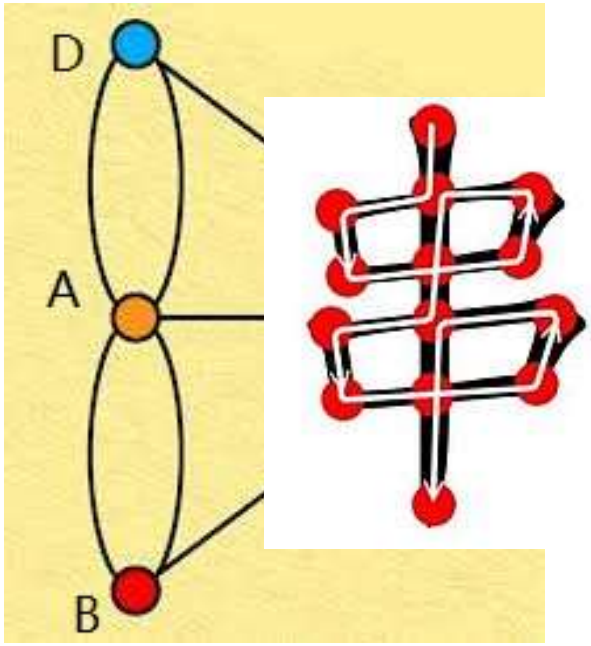
在哥尼斯堡的一个公园里，有七座桥将普雷格尔河中两个岛及岛与河岸连接起来(如图)。问是否可能从这四块陆地中任一块出发，恰好通过每座桥一次，再回到起点？



图论

- 1736年29岁的欧拉向圣彼得堡科学院递交了《哥尼斯堡的七座桥》的论文，在解答问题的同时，开创了数学的一个新的分支——图论与几何拓扑

一笔画问题



图论

- 图论（**Graph Theory**）是数学的一个分支。它以图为研究对象。图论中的图是由若干给定的点及连接两点的线所构成的图形，这种图形通常用来描述某些事物之间的某种特定关系，用点代表事物，用连接两点的线表示相应两个事物间具有这种关系。
- 图是由顶点集合及顶点间的关系集合组成的一种数据结构 **G(V, E)**，图由边和点组成。

图的定义

G = (V, E)

V：顶点集，非空有限；

E：边集，有限；

对**e ∈ E**，记为**(v, w)**

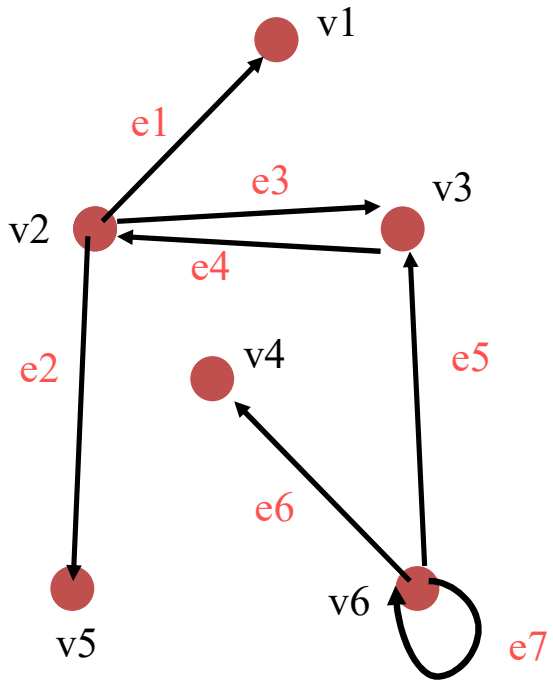
v、**w** 邻接

与**e**相关联

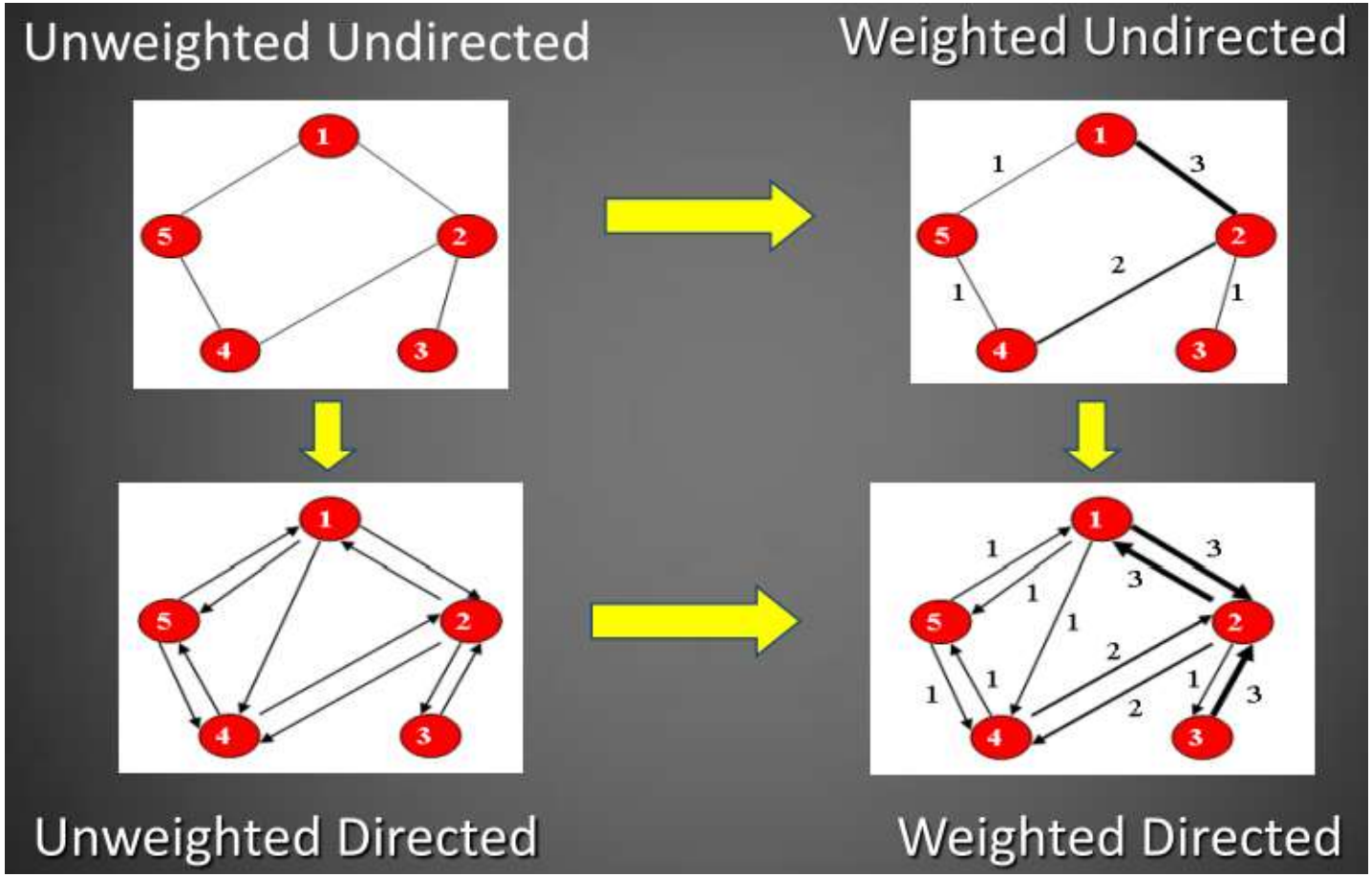
有向图与无向图

有向： **(v, w) ≠ (w, v)**

无向： **(v, w) = (w, v)**



图的分类


Unweighted Undirected
Weighted Undirected
Unweighted Directed
Weighted Directed

表示与存储

Computer Array

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Adjacency matrix

Computer Relational DB

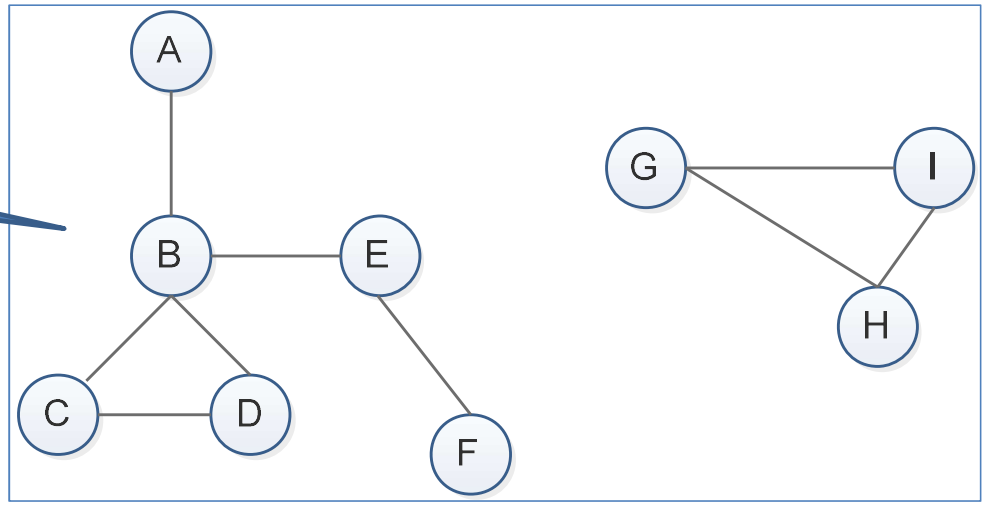
$$\begin{pmatrix} 1 & 2 & 1 \\ 1 & 3 & 0 \\ 1 & 4 & 0 \\ 1 & 5 & 1 \\ 2 & 3 & 1 \\ 2 & 4 & 1 \\ 2 & 5 & 0 \\ 2 & 6 & 0 \\ 3 & 4 & 1 \\ 3 & 5 & 1 \\ 3 & 6 & 1 \\ 4 & 5 & 0 \\ 4 & 6 & 0 \\ 5 & 6 & 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & 2 & 1 \\ & & \\ & & \\ 1 & 5 & 1 \\ 2 & 3 & 1 \\ 2 & 4 & 1 \\ & & \\ & & \\ 3 & 4 & 1 \\ 3 & 5 & 1 \\ 3 & 6 & 1 \\ & & \\ & & \\ & & \end{pmatrix}$$

(a) Normal (b)Pajek

路径及连通性

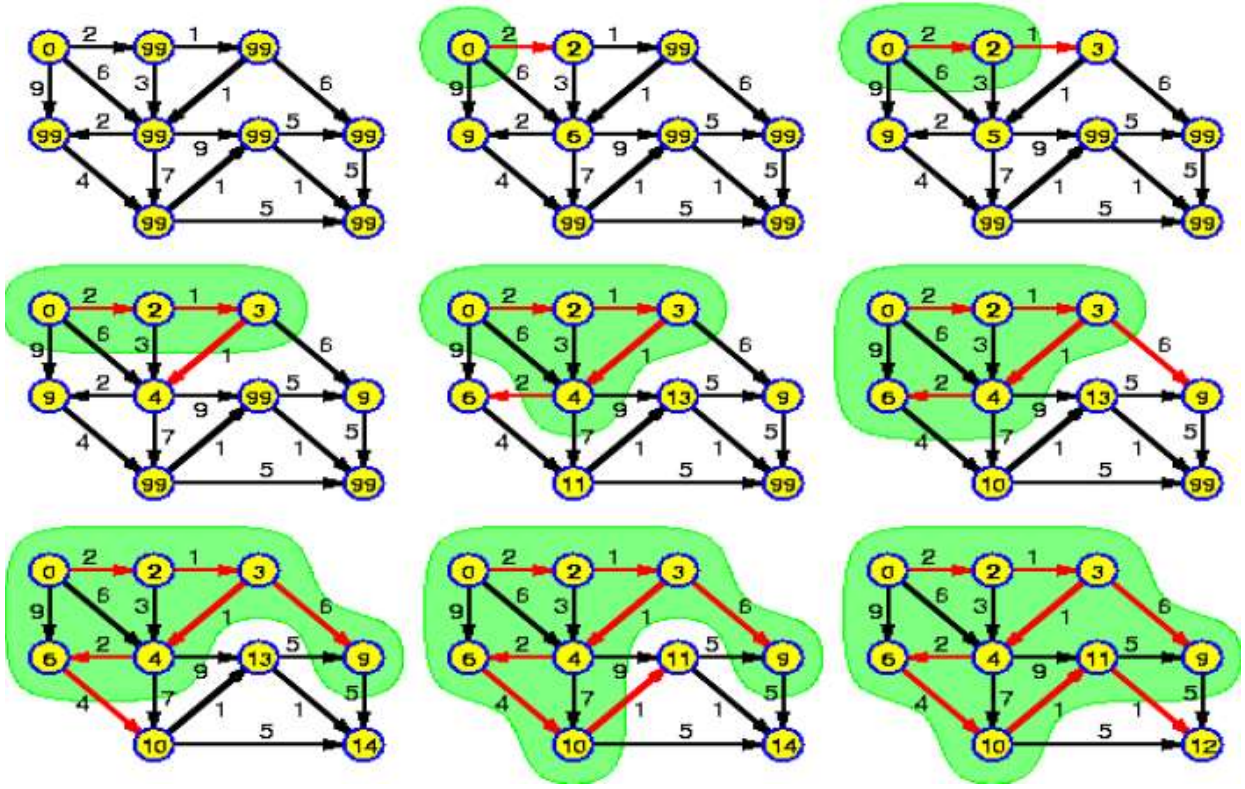
- 路径Path: A sequence of links which connect a sequence of nodes:
  - Path (i, j) = {(i, i_1), (i_1 , i_2), …, (i_k , j)}, with start node i and end node j
- 连通性Connected： at least one path between each pair of nodes

包含两个连通片的非连通图



Dijkstra算法

- 计算加权有向图的最短路径
- Q：如何计算连通图中任意一点*i*到其他所有节点的最短距离？



关于图的计算

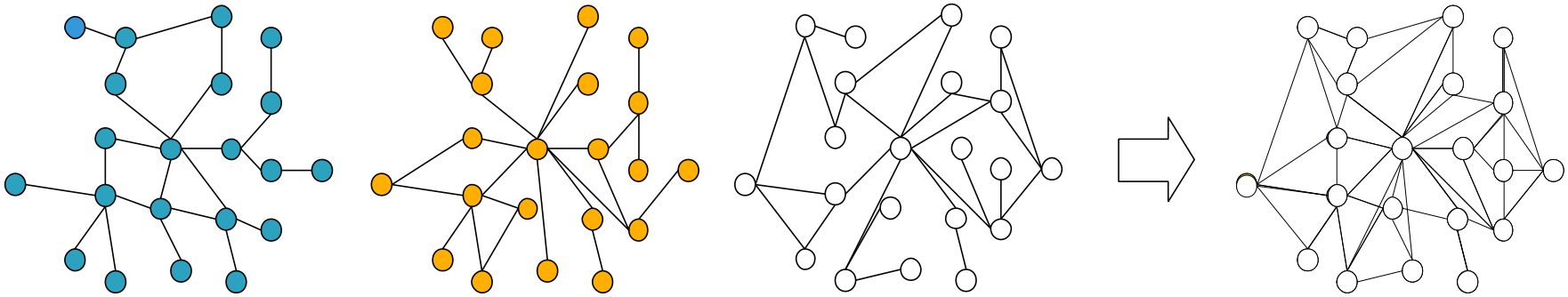
- **Watts（WS小世界模型的提出者）：**

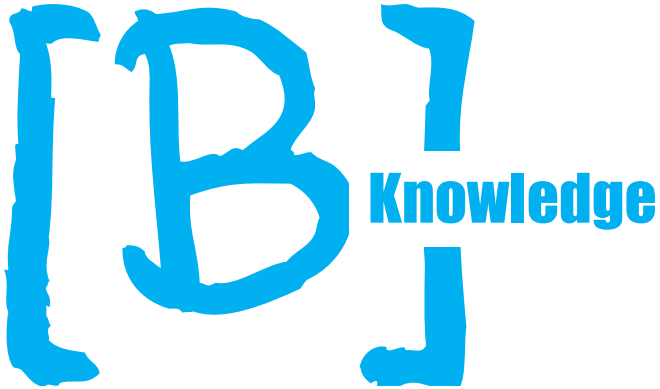
> “图论对我来说是一个谜。作为纯数学的一个分支，图论大体上可分为两部分：一部分几乎是一看就懂，另一部分则晦涩难懂。我从一本教材上学会了一看就懂的部分，但是学了好久也没搞明白晦涩的部分。最后只好承认那部分并不怎么有趣”

- 图的计算涉及的内容很多，这里仅介绍了相关的概念以及与图谱、复杂网络相关的内容，其他内容可以参阅相关专业资料

图谱维度

- 相比于图，知识图谱丰富了节点和边的语义信息，更能反映现实世界的实体及实体间的关系。总之，知识图谱就是把所有不同种类的信息（**Heterogeneous Information**）连接在一起而得到的一个语义关系网络。
- 采用知识图谱的方式进行多维度的客户关系建模，可以获得不同维度下的企业之间的关联关系，我们称之为“同构模式关联图谱”
- 不同维度的图谱可以进行叠加，叠加方式可分为部分叠加和全部叠加，我们称叠加后获得的关联图谱为“异构模式关联图谱”





知识图谱

复杂网络

节点重要性和相似性

预测算法及模型

案例分析

网络

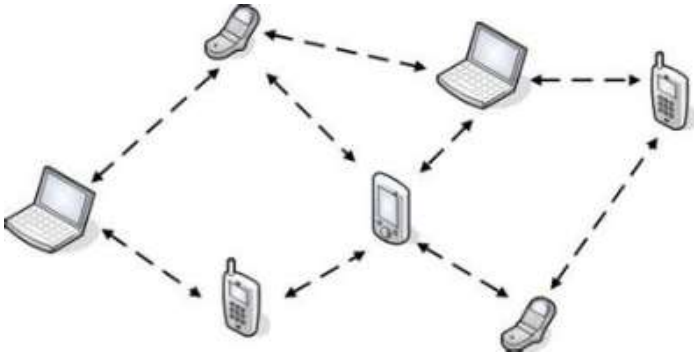
- 网络是由节点和连线构成，表示诸多对象及其相互联系。
- 在数学上，网络是一种图，一般认为是一种有向加权图。
- 网络除了数学定义外，还有具体的物理含义，即网络是从某种相同类型的实际问题中抽象出来的模型。
- 在计算机领域中，网络是信息传输、接收、共享的虚拟平台，通过它把各个点、面、体的信息联系到一起，从而实现这些资源的共享。网络是人类发展史来最重要的发明，提高了科技和人类社会的发展。

复杂网络

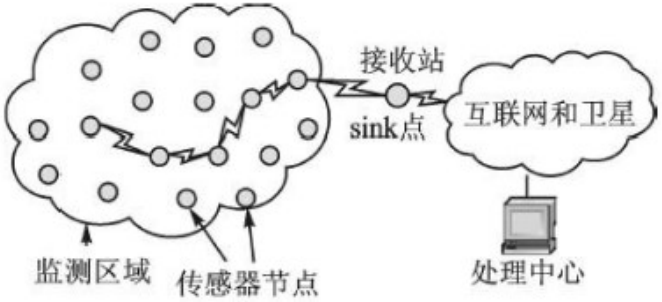
- 复杂网络（**Complex Network**），具有自组织、自相似、吸引子、小世界、无标度中部分或全部性质的网络称为复杂网络。（钱学森）
- 简而言之，复杂网络是呈现高度复杂性的网络
- 研究复杂网络的目的是研究复杂系统的结构以及动力学特性

自组织

- 自组织：如果一个系统靠外部指令而形成组织，就是他组织；如果不存在外部指令，系统按照相互默契的某种规则，各尽其责而又协调地自动地形成有序结构，就是自组织。



移动通信网络



无线传感器网络

自相似

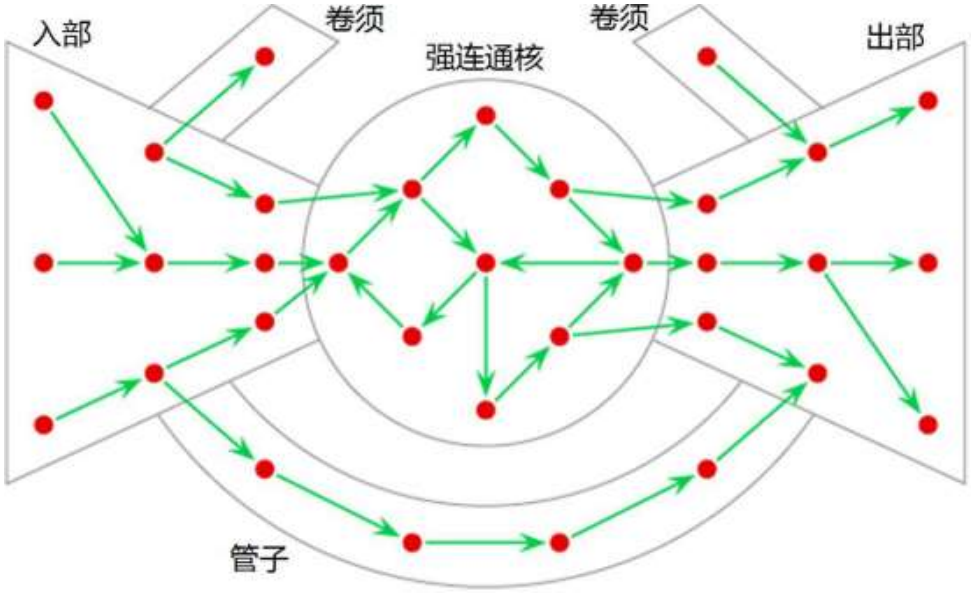
- 自相似：
  - 一种形状的每一部分在几何上相似于整体，一般对分形而言。
  - 基于自相似，根据网络的部分节点特性，可以计算网络其他实体特性

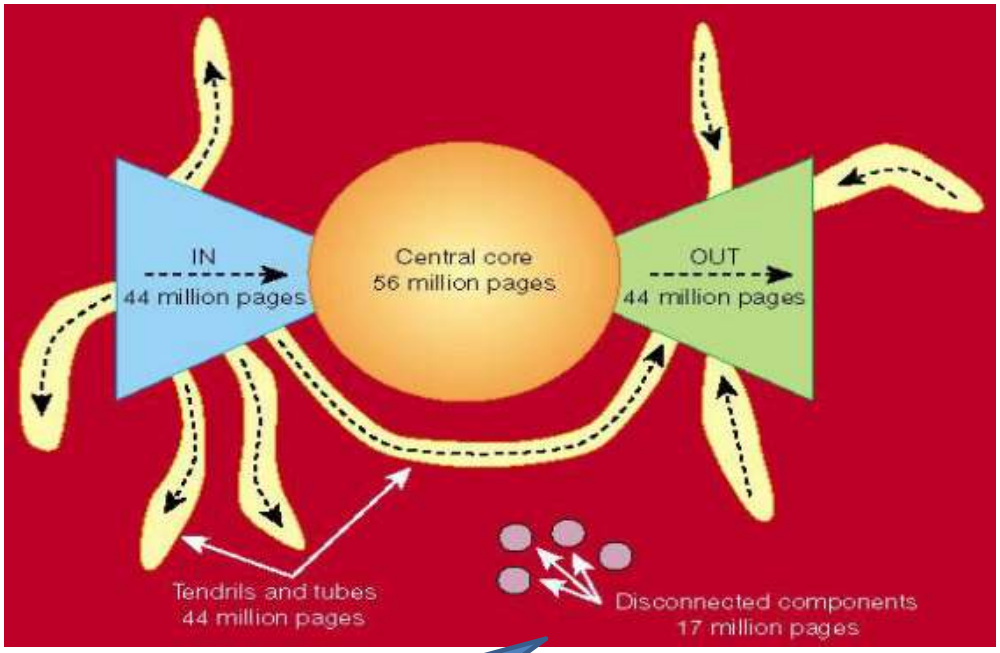
吸引子

- 吸引子：相空间（可以表示出一个系统所有可能状态的空间）中稳定的不动点集。
- 海纳百川，大海就是百川的吸引子
- 落叶归根，树根就是叶子的吸引子
- 热力学系统的平衡态就是该系统的吸引子

复杂网络的连通性

■ 有向网络中的巨片——蝴蝶结

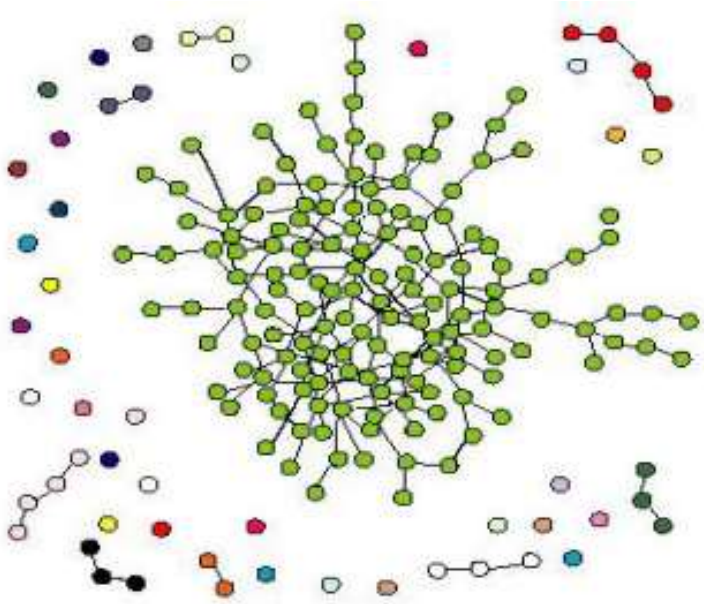




WWW的蝴蝶结结构

巨片

- 大规模网络的连通性其实是一个相当脆弱的性质，因为单个节点或者少部分节点的行为都可能破坏连通性
- 尽管全球社会关系网是不连通的，但是你所在的连通片确实可以非常大
- 经验表明，许多大规模的复杂网络都是不连通的，但总有相当比例的节点形成巨大的连通片，简称“巨片”
- 在分析一个复杂网络时，需要考虑巨片的存在，即重点分析的是连通的巨片

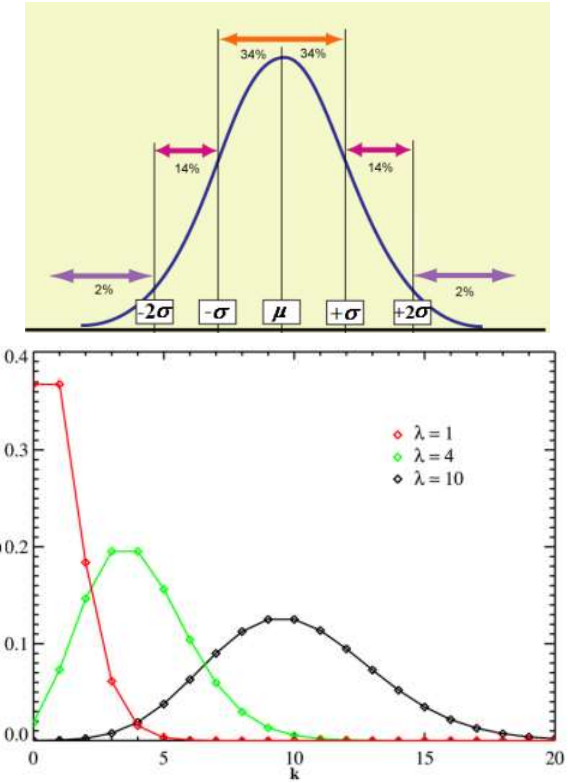


小世界-Bacon数

- 大多数网络尽管规模很大但是任意两个节（顶）点间却有一条相当短的路径
- **Kavin Bacon**在许多部电影中饰演小角色
- **Virginia**大学的计算机专家**Brett Tjaden**设计了一个游戏，他声称电影演员**Kevin Bacon**是电影界的中心
- 在游戏里定义了一个所谓的**Bacon**数：随便想一个演员，如果他（她）和**Kavin Bacon**一起演过电影，那么他（她）的**Bacon**数就为**1**；如果他（她）没有和**Bacon**演过电影，但是和**Bacon**数为**1**的演员一起演过电影，那么他的**Bacon**数就为**2**；依此类推
- 发现: 在曾经参演的美国电影演员中，没有一个人的**Bacon**数超过**4**

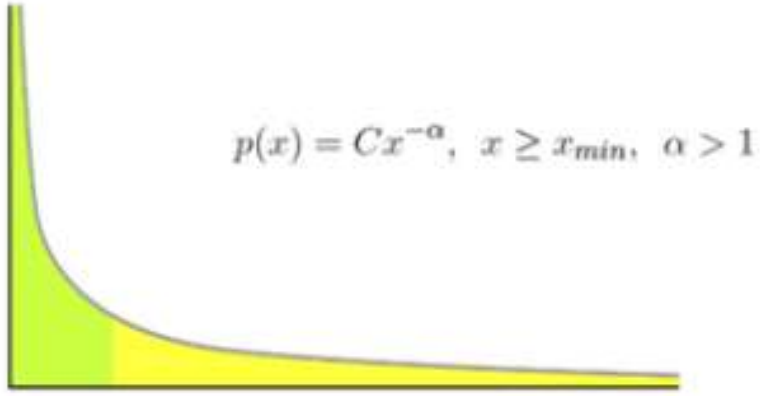
度分布

- 常见的分布为正态分布
  - $\xi - N(\mu, \sigma^2)$
  - $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-u)^2}{2\sigma^2}}$
- 泊松分布
  - $p(k) = \frac{\lambda^k e^{-\lambda}}{k!}$
- 幂律分布
  - $P(k) \sim k^{-\gamma}$ $P(k) = Ck^{-\gamma}$



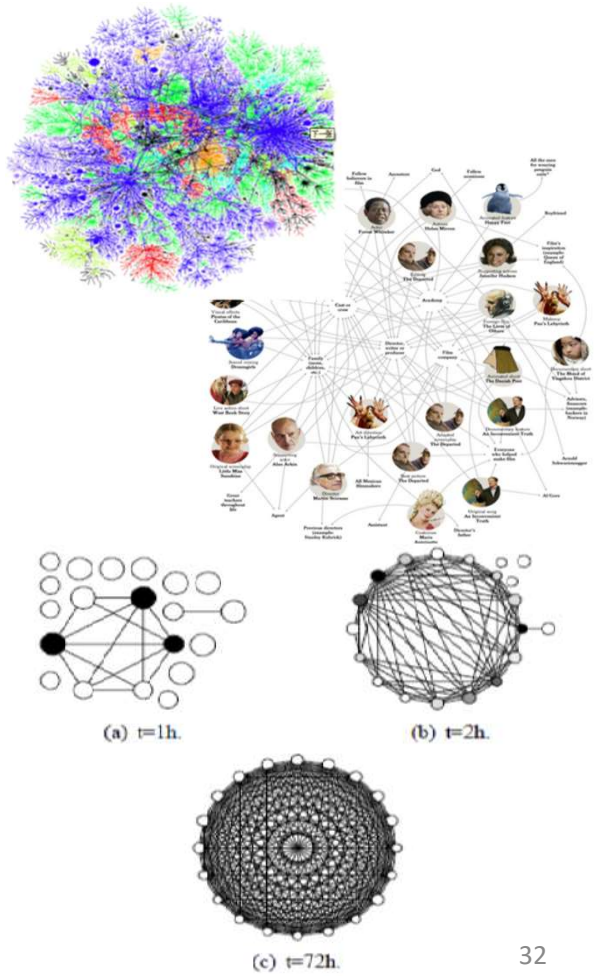
无标度

- 现实世界的网络大部分都不是随机网络，少数的节点往往拥有大量的连接，而大部分节点却很少，一般而言他们符合**Zipf**定律（也就是**80/20**马太定律，最简单的幂律分布）。
- 将度分布符合幂律分布（长尾分布）的复杂网络称为无标度网络。



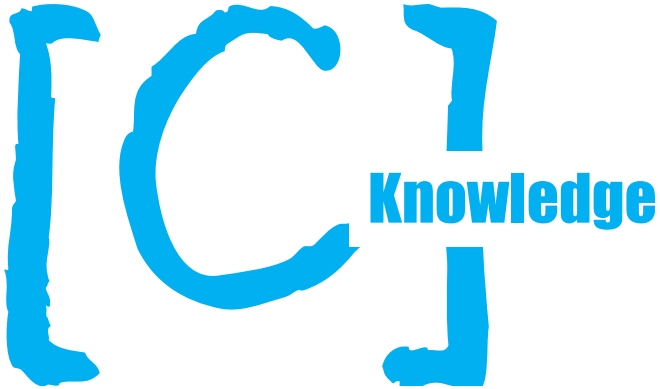
复杂性的表现

- 结构复杂：表现在节点数目巨大，网络结构呈现多种不同特征。
- 网络进化：表现在节点或连接的产生与消失。例如**World Wide Web**，网页或链接随时可能出现或断开，导致网络结构不断发生变化。
- 连接多样性：节点之间的连接权重存在差异，且有可能存在方向性。
- 节点多样性：复杂网络中的节点可以代表任何事物。例如，人际关系构成的复杂网络节点代表单独个体，万维网组成的复杂网络节点可以表示不同网页。
- 动力学复杂性：节点集可能属于非线性动力学系统，例如节点状态随时间发生复杂变化。



结论

- 复杂网络是一种网络，网络用图描述
- 知识图谱是图，是一种带有语义（知识）的图
- 知识图谱是复杂网络吗？只要满足复杂网络定义的知识图谱就是，否则就不是！
- 知识图谱是节点和边带有丰富语义的复杂网络



知识图谱

复杂网络

节点重要性和相似性

预测算法及模型

案例分析

节点重要性和相似性分析

- 知识图谱中总有一些节点起到核心作用，为了有效控制，需要将这些节点找出来，即节点重要性分析；同样，当某个节点出现问题（如风险），需要将类似的节点找出来，便于及时采取措施，需要进行相似性分析
- 节点重要性
  - HITS算法
  - **PageRank算法**
  - 度中心性
  - 介数中心性
- 节点相似性
  - 距离（余弦距离）

PageRank算法背景

- Google的网页排名算法，发明专利
- 解决的问题：网页搜索结果的排名不好
- 关键问题：网页权重计算

- PageRank并不是唯一的网页排名算法
  - 雅虎的目录分类

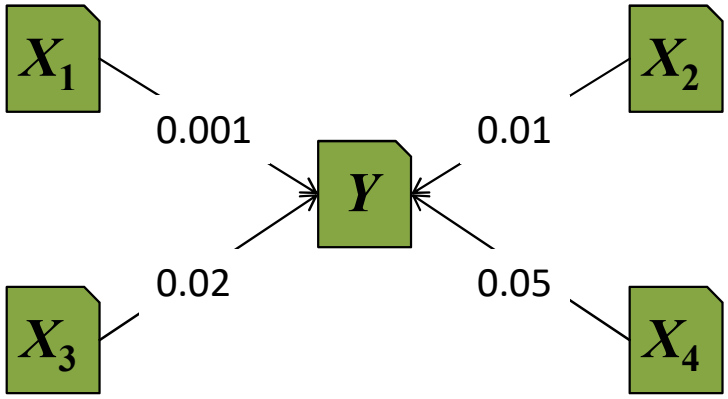
- 互联网是一个网络，一个关于网页的网络，网页和网页之间通过锚文本进行连接

- 引申：计算复杂网络的节点重要性

PageRank的基本思想

- 基本思想：
  - 一个网页的重要性（即权重）是由链接到这个网页的上游网页的重要性决定的
  - 一个网页 Y 的排名应该来自于所有指向这个网页的其他网页 X_1, X_2, … , X_K 的权重之和，如下图，网页 Y 的排名

    $pagerank = 0.001 + 0.01 + 0.02 + 0.05 = 0.081$



PageRank的计算方法

- 假定向量

  $\mathbf{B} = [b_1, b_2, \ldots, b_N]^T$

  为第一、第二、…第N个网页的网页排名（pagerank，即权重）

- 矩阵

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} & \cdots & a_{1M} \\ \cdots & & & & \cdots \\ a_{m1} & \cdots & a_{mn} & \cdots & a_{mM} \\ \cdots & & & & \cdots \\ a_{M1} & \cdots & a_{Mn} & \cdots & a_{MM} \end{bmatrix}$$

为网页之间链接的数目，其中a_{mn}代表第m个网页指向第n个网页的链接数（注意：链接的权重由上游节点权重决定）。A是已知的，B是未知的

PageRank的计算方法Cont.

- 现在计算B，B算出来，排名就出来了，即网页权重排序
- 假定B_i是第i次迭代的结果，那么
  - $B_i = A \cdot B_{i-1}$ （1）
- 初始假设：所有网页的权重都是1/N，即
  - $B_0 = [\frac{1}{N}, \frac{1}{N}, \cdots, \frac{1}{N}]^T$
- 那么，根据公式（1）就可以计算得到 B_1，B_2，…。当两次迭代结果B_i 和 B_{i-1} 之间差很小并接近于0时，则停止迭代
- 一般来讲，只要10次左右的迭代基本就收敛了

PageRank的计算方法Cont.

- 平滑处理
  - 由于网页之间的链接数量相比互联网的规模非常稀疏，A中会出现0链接数量或极小（相对互联网）链接数量
  - A中不允许出现0，出现0意味着该网页的重要性为0，这不符合实际情况
  - 所以，需要进行平滑处理，基本方法是将A中的0用一个小的常数替代。
  - 于是，上述公式（1）就演变为

$$B_i = [\frac{\alpha}{N} \cdot I + (1-\alpha)A] \cdot B_{i-1} \qquad (2)$$

其中N为互联网网页的数量，α是一个较小的常数，I为单位矩阵

PageRank的计算方法Cont.

- 最后一个问题
  - 如果将互联网的所有页面都考虑进去，上述公式（1）或（2）的计算量都是十分庞大的
  - 如何提升计算性能？MapReduce方法！这是Google进行大数据处理的又一重要方法，是一个并行的编程模型
  - 由于PageRank算法中采用了矩阵计算，因此，很容易采用MapReduce对计算任务进行划分，并采用并行的方式提升计算性能

节点相似性分析

- 在知识图谱中，两个节点是否相似？
- 节点相似性计算有很多方法，最直接的方法是利用节点的属性进行距离计算，距离越小越相似
  - 例如，节点如果描述的是人，则计算两个节点是否相似性可以利用各自的年龄、性别、职业、兴趣等
  - 计算过程中需要进行量化和归一化处理
- 基于拓扑结构也可以计算两个节点的相似性
  - 例如，先计算两个节点的度、余度、介数、聚类系数等，然后计算二者之间的距离

余弦相似度

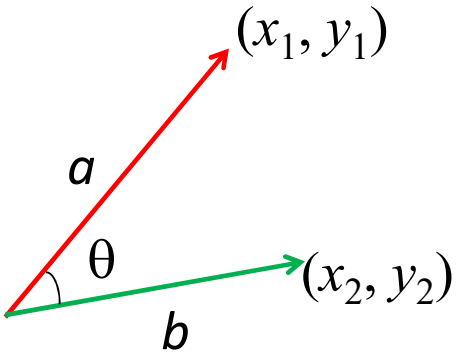
- 在计算两个节点的相似性过程中，不可避免地出现属性是文本描述的情况，那么如何计算两段文字的相似性呢？
  - 句子A：这只皮靴号码大了。那只号码合适
  - 句子B：这只皮靴号码不小，那只更合适
- 如何度量句子A和句子B的相似度？

余弦相似度Cont.

- **基本思路**
  - 如果这两句话的用词越相似，它们的内容就应该越相似。因此，可以从词频入手，计算它们的相似程度。
- **第1步，分词。**
  - 句子A：这只/皮靴/号码/大了。那只/号码/合适。
  - 句子B：这只/皮靴/号码/不/小，那只/更/合适。
- **第2步，列出所有的词。**
  - 这只，皮靴，号码，大了。那只，合适，不，小，很
- **第3步，计算词频。**
  - 句子A：这只1，皮靴1，号码2，大了1。那只1，合适1，不0，小0，更0
  - 句子B：这只1，皮靴1，号码1，大了0。那只1，合适1，不1，小1，更1
- **第4步，写出词频向量。**
  - 句子A：(1，1，2，1，1，1，0，0，0)
  - 句子B：(1，1，1，0，1，1，1，1，1)

余弦相似度Cont.

- 向量a和向量b的夹角 的余弦计算如下



$$\cos(\theta) = \frac{a \cdot b}{\|a\| \times \|b\|}$$

$$= \frac{(x_1,y_1)\cdot(x_2,y_2)}{\sqrt{x_1^2+y_1^2}\times\sqrt{\chi_2^2+y_z^2}}$$

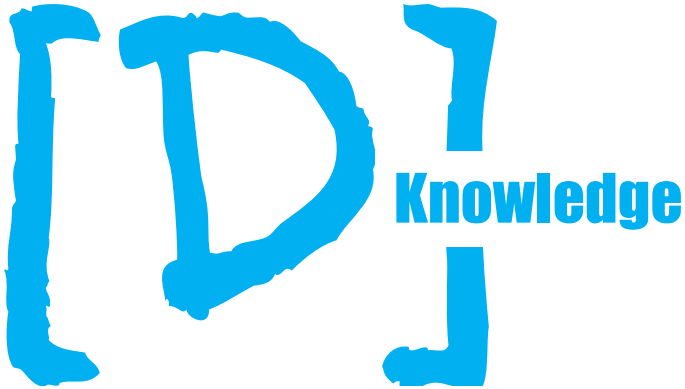
$$= \frac{x_1x_2+y_1y_2}{\sqrt{x_1^2+y_1^2}\times\sqrt{x_2^2+y_2^2}}$$

余弦相似度Cont.

- 如果向量a和b不是二维而是n维，上述余弦的计算法仍然正确。假定a和b是两个n维向量，a是$[a_1, a_2, ..., a_n]$，b是$[b_1, b_2, ..., b_n]$，则a与b的夹角 的余弦等于：

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i^2)} \times \sqrt{\sum_{i=1}^{n}(y_i^2)}}$$

计算结果：71%



知识图谱

复杂网络

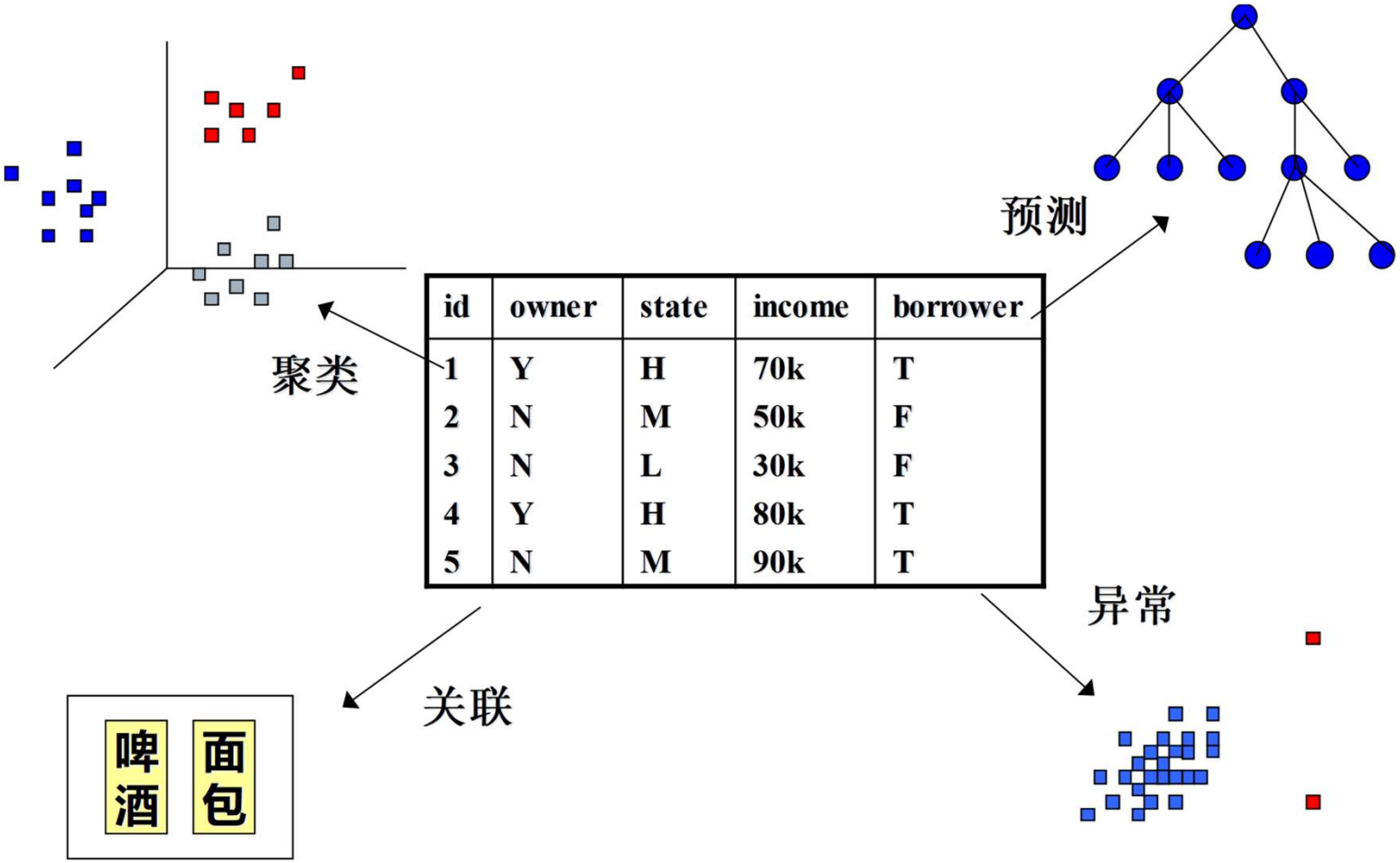
节点重要性和相似性

预测算法及模型

案例分析

机器学习四类算法

- 每一种算法后面都有一个严格数学模型支持

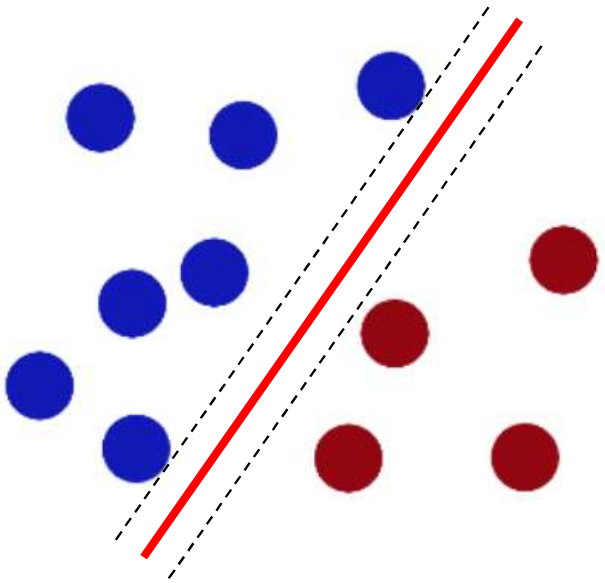


关于预测算法及模型

- 预测的本质是分类问题，在分类算法的基础上加一个时间维度，就可以作为预测算法，因此，分类算法=预测算法
- 常用的分类算法
  - 逻辑回归
  - SVM
  - 神经网络
  - 深度学习
  - 时间序列
  - 隐马尔可夫
  - 贝叶斯网络
  - 随机森林
  - ……

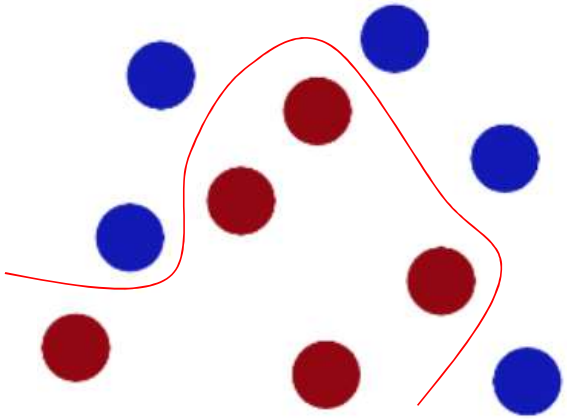
线性分类

- 用一个线性函数进行划分的分类（二分类）
- 右图，一个线性可分问题，可以用一条直线进行划分
- 有无数条直线可以完成这个任务，到底该选哪一条最合适？
  - 很显然中间红色的那条线最合适，即离两边的最近的点最远
  - “学术点”：所有正分类点到该直线的距离与所有负分类点到该直线的距离的总和达到最大，这条直线就是最优分类直线
- 显然，只要是线性就好办！



非线性分类

- 右图，用线性函数搞不定分类了！
- 这是一个非线性分类问题
- 采用一个曲线可以将两种实心圆分开



- 针对这个问题可以用一个非线性函数解决分类问题
- 但是，如果每个问题都找一个曲线函数，这太不科学了！

Logistic Regression算法

- 逻辑回归(Logistic Regression, LR)模型其实仅在线性回归的基础上，套用了一个逻辑函数，也就由于这个逻辑函数，使得逻辑回归模型成为了机器学习领域一颗耀眼的明星
- 这个逻辑函数是Sigmoid函数

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

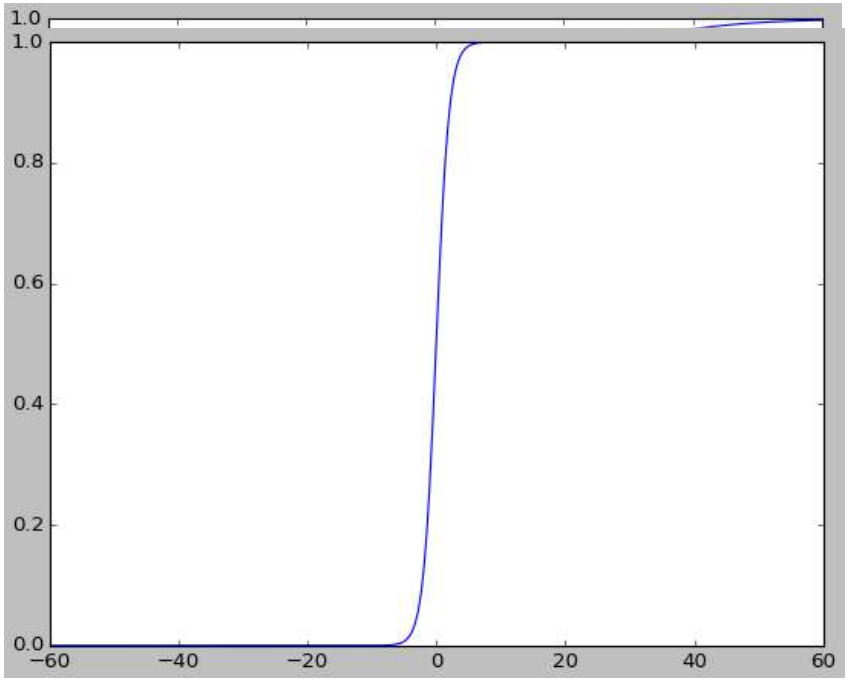
- 对比线性函数

$$\sigma(z) = az + b$$

Logistic算法核心：Sigmoid函数

- 逻辑回归算法使用的函数是Sigmoid（类似单位阶跃函数，但是单位阶跃函数的瞬间跳跃过程很难控制），是一种非线性函数
- 当x的取值范围扩大，Sigmoid函数与单位阶跃函数类似

Sigmoid函数



Sigmoid函数

- **Sigmoid函数**

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

  - 自变量：z
  - 因变量：$\sigma(z)$表示的是概率
- 其中
  - $z = w^T x = w_0 x_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + \ldots + w_n x_n$
  - 向量x是输入，向量w是参数，z是输出（注意z在这里可以理解为单值向量）
- 在训练过程中，向量x和z是已知的，需要确定的是向量w，即分类器的参数
- 如何计算最佳的向量w？

Sigmoid函数 Cont.

- 重新分析Sigmoid函数

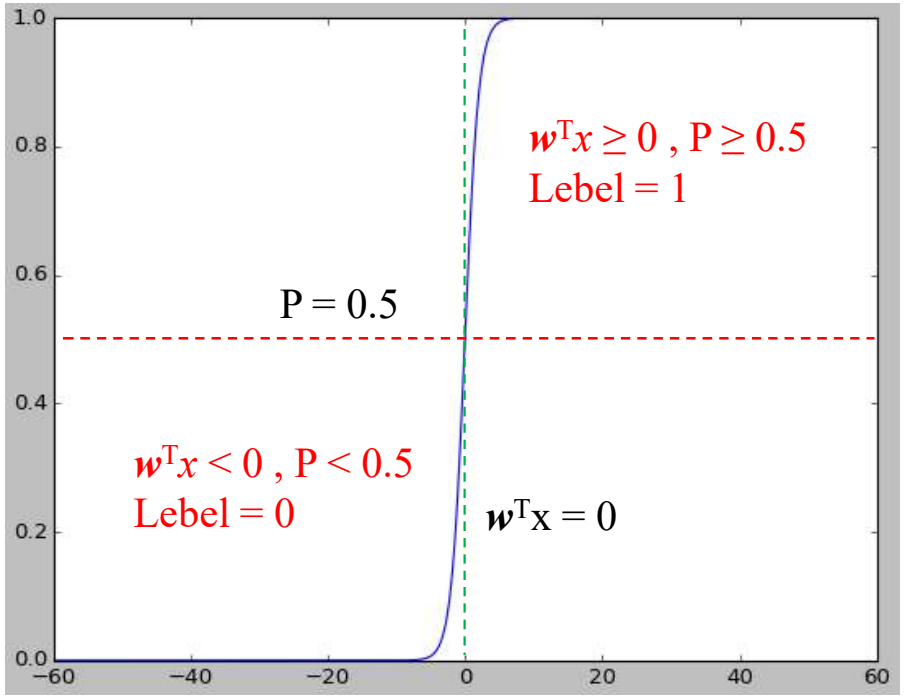
$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

- 等价于

$$\sigma(z) = P(y = 1 \mid x) = \frac{1}{1 + e^{-w^{T}x}}$$

- 即，$\sigma(z)$表示的是x取某个值$y = 1$的概率

Sigmoid函数 Cont.



关于输入向量x

- 如果 σ (z) 表示的是x取某个值 $y=1$的概率，而 $z=w^{T}x=w_0x_0+w_1x_1+w_2x_2+w_3x_3+\ldots+w_nx_n$，那么向量 x 表示指标集，而 w 则表示对应的权重向量
- 如

w_0x_0表示指标0（如最小贷款额度）的权重与值的积，w_0x_0是一个常数，$x_0=1$，所以，$z=w^{T}x=w_0+w_1x_1+w_2x_2+w_3x_3+\ldots+w_nx_n$

w_1x_1表示指标1（如贷款周期）的权重与值的积

w_2x_2表示指标2（如资产）的权重与值的积

.....

w_nx_n表示指标n（如上月销售收入）的权重与值的积

最佳向量w的计算

- 最佳向量w的计算实质上是一个优化问题
- 常用的优化方法（训练算法）：
  - 梯度上升法：要找到某函数的最大值，最好的方法是沿着该函数的梯度方向探寻，凹函数
  - 梯度下降法：同上，找最小值，沿着梯度方向，凸函数

关于梯度

- 在向量微积分中，标量场的梯度是一个向量场。标量场中某一点上的梯度指向标量场增长最快的方向，梯度的长度是这个最大的变化率。更严格的说，从欧几里得空间**Rn**到**R**的函数的梯度是在**Rn**某一点最佳的线性近似。在这个意义上，梯度是雅可比矩阵的一个特殊情况。
- 在单变量的实值函数的情况，梯度只是导数，或者，对于一个线性函数，也就是线的斜率。
- 梯度一词有时用于斜度，也就是一个曲面沿着给定方向的倾斜程度。可以通过取向量梯度和所研究的方向的点积来得到斜度。梯度的数值有时也被称为梯度。

方向导数

- 方向导数是在特定方向上函数的变化率，方向导数在各个方向上的变化一般是不一样的，那到底沿哪个方向最大呢？沿哪个方向最小呢？为了研究方便，就有了梯度的定义。
- 很明显梯度实际上就是以对x的偏导为横坐标，以对y偏导数为纵坐标的一个向量，而方向导数就等于这个向量乘以指定方向的单位向量。
- 根据向量乘积的定义可知，对于一个给定的函数，它的偏导是一定的（当然是在同一个点），所以当给定方向与梯度方向一致时，变化最快
- 总的来说，梯度的定义是为了研究方向导数的大小更方便而定义的。

梯度上升法

- 如果梯度记为▽，则函数$f(x,y)$的梯度由下式梯度算子表示

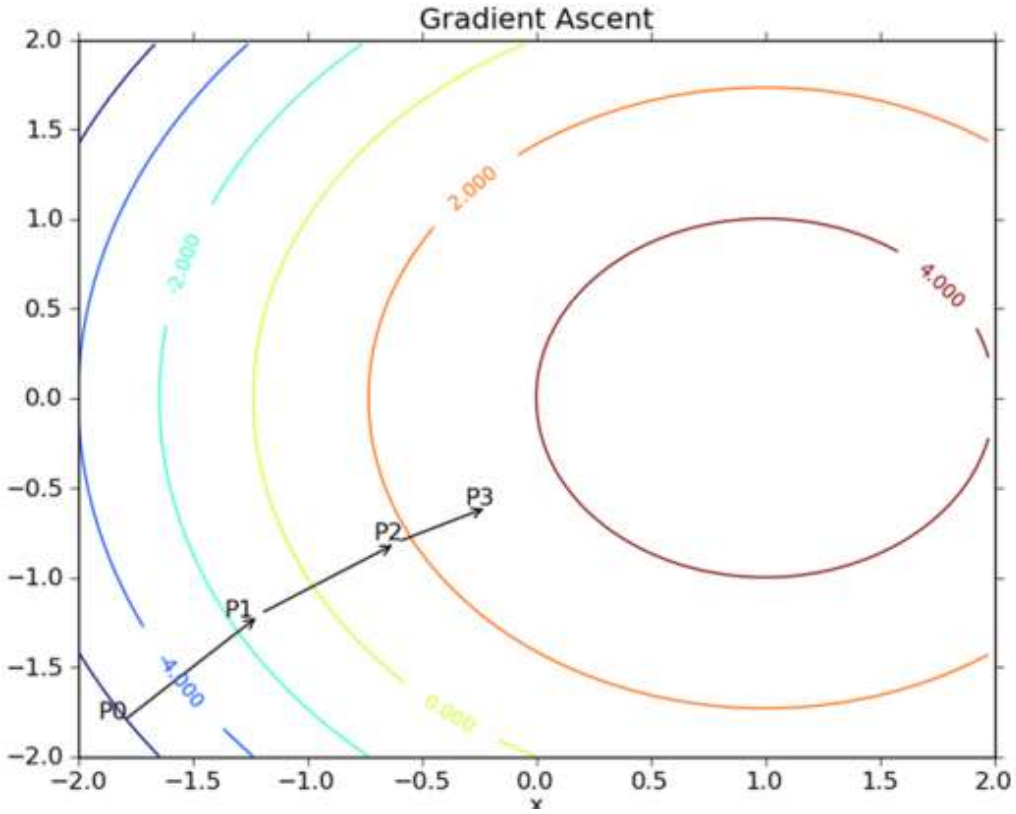
$$\nabla f(x,y) = \begin{pmatrix} \frac{\partial f(x,y)}{\partial x} \\ \frac{\partial f(x,y)}{\partial y} \end{pmatrix}$$

- 沿x方向移动$\frac{\partial f(x,y)}{\partial x}$，则沿y方向就要移动$\frac{\partial f(x,y)}{\partial y}$，数学上证明，这样走下去，必然最快走到收敛点
- 计算向量w

  $w := w + \alpha \nabla_w f(w)$， α 表示步长
- 该公式一直迭代执行，直到某个停止条件为止，如达到某个误差范围

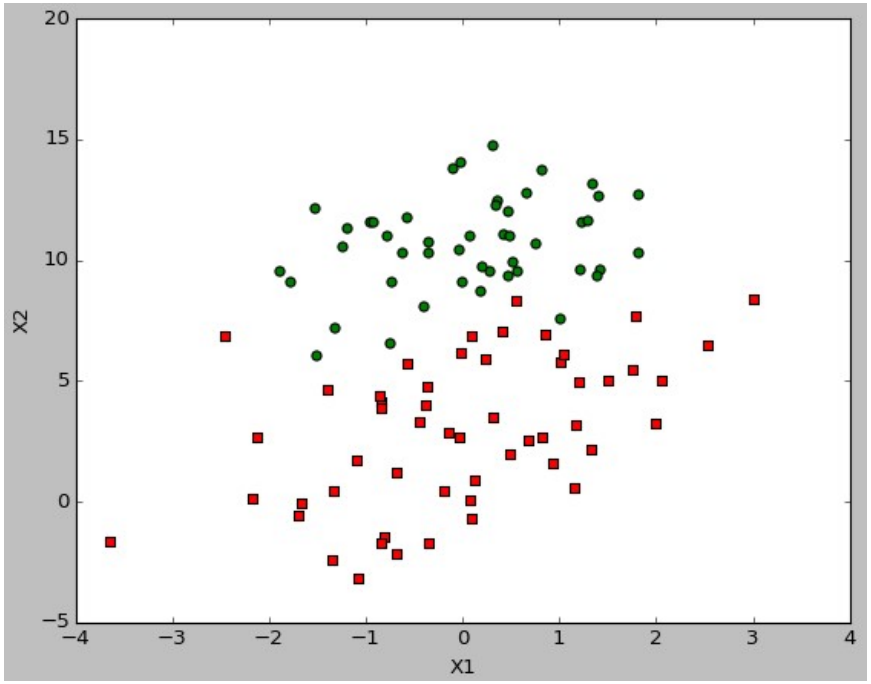
梯度上升

- 梯度上升算法到达每个点后都会重新估计移动方向。从**P0**开始，根据上述梯度算子∇计算完**P0**的梯度，函数根据梯度移动到下一点**P1**，在**PI**点继续计算，重复迭代，直至满足停止条件。
- 迭代过程中，梯度算子总能保证我们能选取最佳的移动方向



举个栗子

- 训练数据集
  - 有100个样本，每个点包含两个数值型特征：X1和X2，一个类别标签



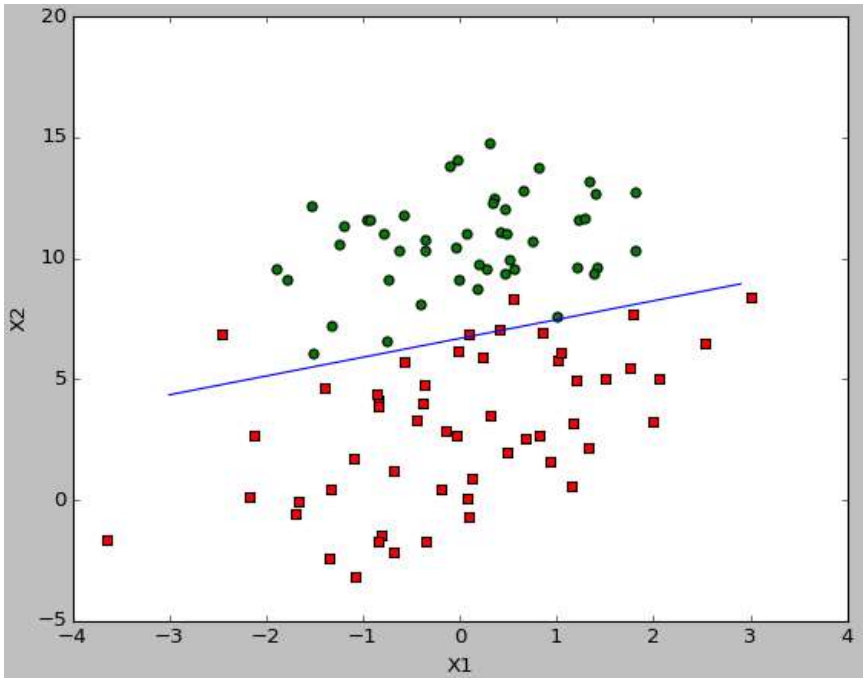
-0.017612	14.053064	0
-1.395634	4.662541	1
-0.752157	6.538620	0
-1.322371	7.152853	0
0.423363	11.054677	0
0.406704	7.067335	1
0.667394	12.741452	0
-2.460150	6.866805	1
0.569411	9.548755	0
-0.026632	10.427743	0
0.850433	6.920334	1
1.347183	13.175500	0
1.176813	3.167020	1
-1.781871	9.097953	0
-0.566606	5.749003	1
0.931635	1.589505	1
-0.024205	6.151823	1
-0.036453	2.690988	1
-0.196949	0.444165	1
1.014459	5.754399	1
1.985298	3.230619	1
-1.693453	-0.557540	1
-0.576525	11.778922	0
-0.346811	-1.678730	1
-2.124484	2.672471	1
1.217916	9.597015	0
-0.733928	9.098687	0
-3.642001	-1.618087	1
0.315985	3.523953	1
1.416614	9.619232	0
-0.386323	3.989286	1
0.556921	8.294984	1
1.224863	11.587360	0

举个栗子

- 该函数将$\mathbf{X_0}$的值设为1，即第一项为常数

$$z = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} = w_0 + w_1x_1 + w_2x_2$$

- 那么，需要确定的是w_0、w_1、w_2三个系数
- 采用梯度上升优化算法计算权值，划分边界如图
- 效果很不错！



如何利用LR解决非线性问题?

- 上述的方法仍然是线性分类问题
- 如何利用LR解决非线性问题？
  - 关键在于 z 函数，上述采用的是线性函数
  - 如果解决非线性分类问题，则z函数需要采用非线性函数，例如平方关系、对数关系、指数关系、三角函数关系等等
  - 但是，为了针对$\sigma(z)$应用梯度上升法，需要将非线性的 z 函数映射为线性的 z 函数，这就需要一个映射函数（kernel函数），这就有SVM算法的味道了

$$z = \theta^T x' = \theta_0 x'_0 + \theta_1 x'_1 + \theta_2 x'_2 = \theta_0 x_0 + \theta_1 x_1^2 + \theta_2 x_2 = x_1^2 + x_2$$

$$x'_0 = \phi_0(x_0), x'_1 = \phi_1(x_1), \ldots x'_n = \phi_n(x_n)$$

知识图谱

复杂网络

节点重要性和相似性

预测算法及模型

案例分析

案例1：EMC² Hackathon: Mars Challenge

- 题目描述：
  - *You and your team just landed on Mars. As you prepare your base of operations, you receive word that massive Sun storms are coming your way. Now radio contact with Earth has been lost. Your base has protective electromagnetic shields that can protect you from the radiation, but can only be running for a few minutes at a time without recharging.*
  - *Your only chance of survival is to monitor the current temperature and radiation levels in the planet's atmosphere to detect sun flares and activate your base shields for protection.*
  - *You only have a few hours to implement a sensor array, build and deploy the monitoring application to engage/disengage your shields, then fine tune an algorithm based on your data analysis that decides when to charge your shields and when to engage them for protection. Will you and your team survive?*
  - *You and your team will have at your disposal the necessary tools to survive and win the challenge, however you will need all wits and skills to work together and implement a solution that allows you to survive and outlast other teams.*

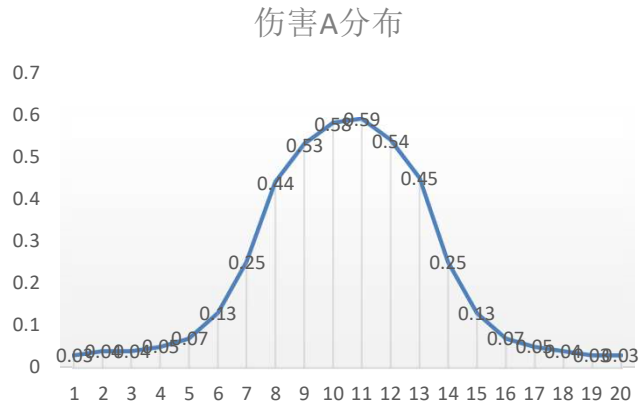
装备

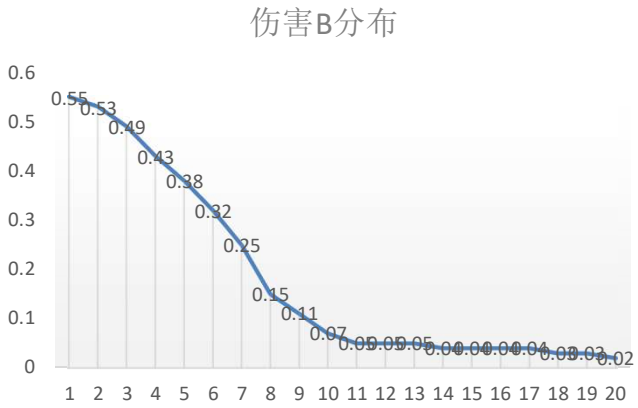
- SensorSuite Service
- Aggregator Service
- Team Command & Control
- Game Controller
- Game Controller Dashboard
- RaspBerry Pi Sensor Setup

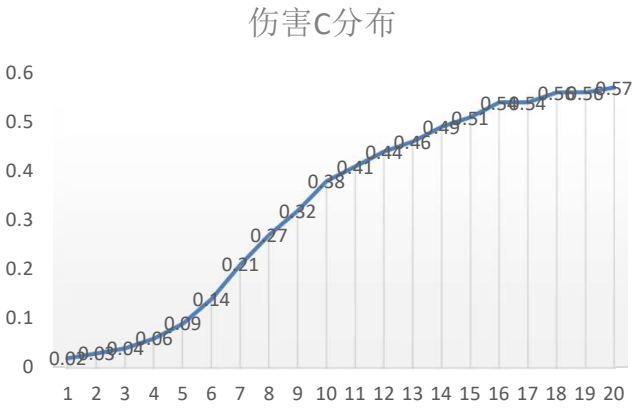
Aggregator	聚合器，整合者; 汇集者，聚合
Command	命令
Control Game	游戏控制
Controller	管理者; 控制者; 控制器
Dashboard	仪表板; 仪表盘; 仪表的控制盘
RaspBerry Pi	树莓派
Sensor	传感器，灵敏元件

算法描述

- 规则：
  - 打开传感器需要电力，消耗“生命力”
  - 不开传感器，无法躲避辐射等伤害，同样消耗“生命力”
  - 假设开1次传感器，生命力-1，躲过1次伤害，生命力+2
  - 假设不开传感器，被伤害1次，生命力-2
- 根据上面描述，可以罗列出可能的伤害事件，如辐射、风暴、地震等，假设有3种伤害类型，每种伤害类型的概率分布计算可得







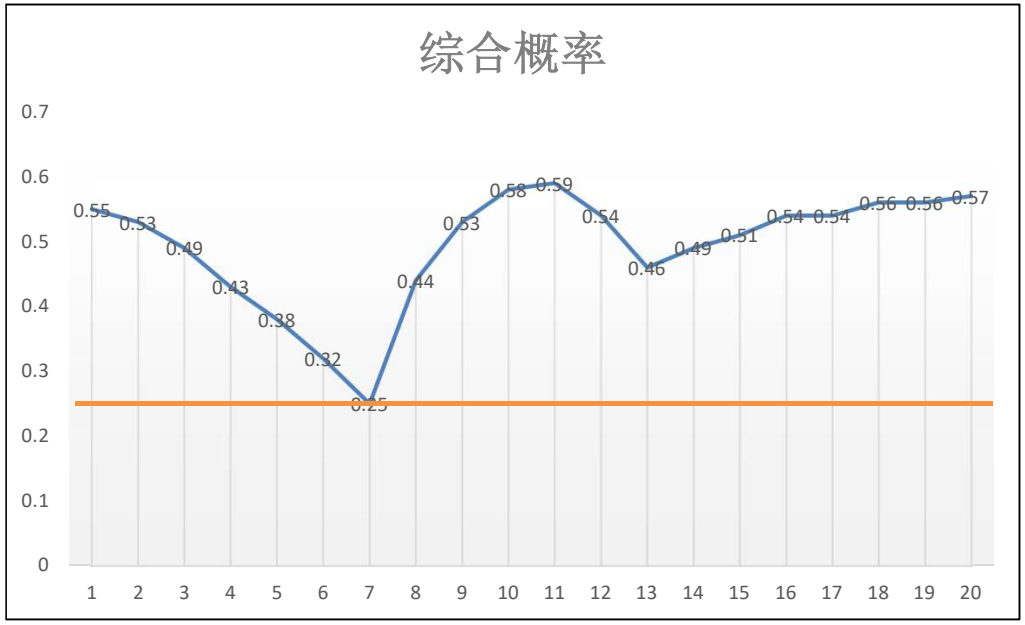
综合概率计算

- 3种伤害的概率叠加后应该如何计算？
- 假设伤害A、B、C的概率分布分别用 $f(A)$、$f(B)$、$f(C)$表示，用 $f(\Omega)$表示综合计算后的概率分布，那么

1. $f(\Omega)=f(A)+f(B)+f(C)$
2. $f(\Omega)=f(A)\times f(B)\times f(C)$
3. $f(\Omega)=f(A)$、$f(B)$、$f(C)$的最大值
4. $f(\Omega)=f(A)$、$f(B)$、$f(C)$的平均值

策略

- 树莓派开启还是关闭是一个策略问题，其实就是根据 $f(\Omega)$ 做一个选择，当 $f(\Omega) \geq \mu$ (阈值)时开启，否则关闭。
- 如何确定μ值呢？可以做一个惩罚函数，即每次生命力增加了，奖励一个分值，否则惩罚一个分值，这样反复训练，就得到了一个较好的策略模型



获奖团队

案例2:系客户关联分析及风险计量

– 客户识别的挑战：随着企业投资集团化、供应链、贸易链等关系愈加复杂，经营模式愈加多样化、隐蔽化，传导化，现有模型技术难以有效评估

引入大数据，设计客户风险体系框架

- 引入大数据思维，创建新型企业信用风险全景描述的体系框架；
- 深度梳理分析与客户相关的各种内部数据和外部数据（银监会、征信、工商、海关、法院等）。

建立客户关联全景视图及风险传导模型

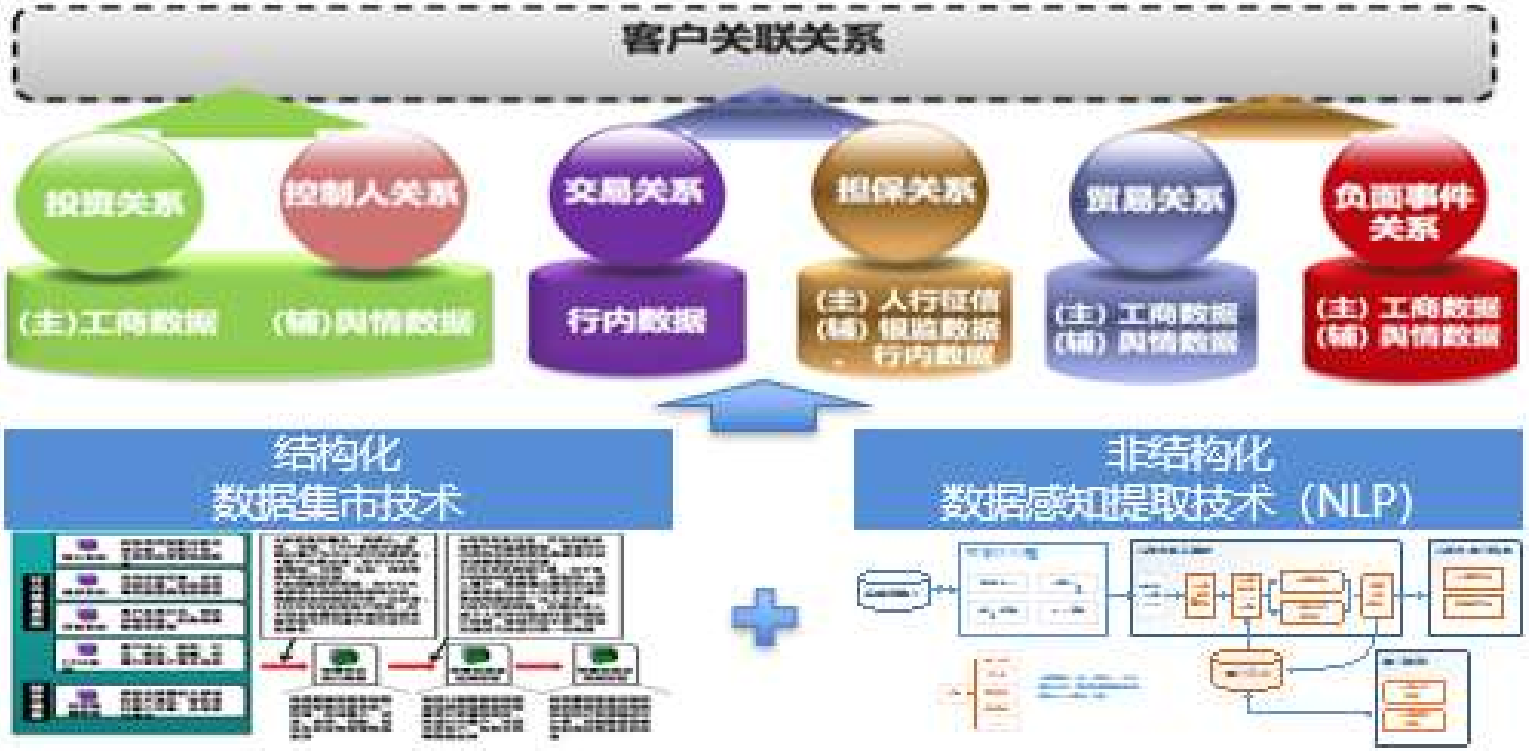
- 有效整合不同来源客户数据的方法和关联维度分类体系，构建客户关联关系全景视图，通过对各类关联维度上的风险传导方式、途径的综合分析，形成风险传导模型。

构建客户综合风险评估量化模型

- 通过数据挖掘技术构建基于大数据的客户风险评分模型，将行内、行外各维度的风险数据输入风险评分模型，对企业客户进行全面风险量化评估。

图谱构建

利用大数据，对企业关联关系进行聚合、去重、补充



系客户识别

基于整合的工商、银监、征信、行内等数据，识别系客户龙头企业，根据龙头企业识别投资链下全部客户，同时拟合担保、控制人、资金和地址关系并控制外延，最终确定以龙头企业为起点的系客户边界。

系客户分析成果

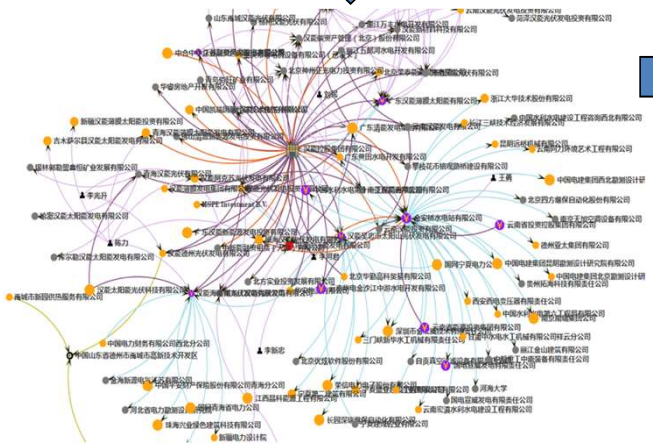
1. 行内有效客户数量：101536家，其中属于龙头企业的有5389家，全部投资数据中涉及龙头企业有7581家
2. 确定的系7418个，其中最大的系包含企业有1111家，系客户数量超过10个 的系有2714个，超过50个的有360个，超过100个的有118个。
3. 根据现有数据目前确定担保圈数量21733个，同一注册地址3502家

系客户	原集团	识别后
英利系	7	英利能源（中国）有限公司（19/45家，有贷款余额7家）
建龙系	4	北京建龙重工集团有限公司（21/37家，，有贷款余额7家）
中钢系	2	中国中钢集团公司（8/73家,有贷款余额2家）
汉能系	6	汉能控股集团有限公司（27/54家,有贷款余额5家）
庆华系	8	中国庆华能源集团有限公司（27/51家,有贷款余额8家）

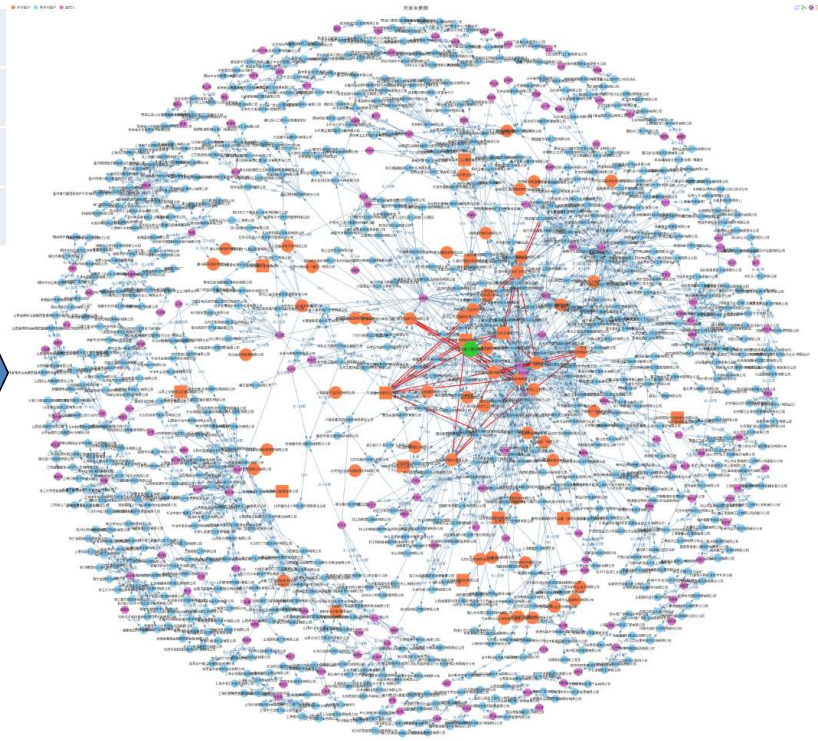
大数据“系客户”识别与计量

银行数据

北京**系	天津**钢铁实业有限公司
	唐山**实业有限公司
	承德**特殊钢有限公司
	天津市**科技有限公司



征信数据



全网数据

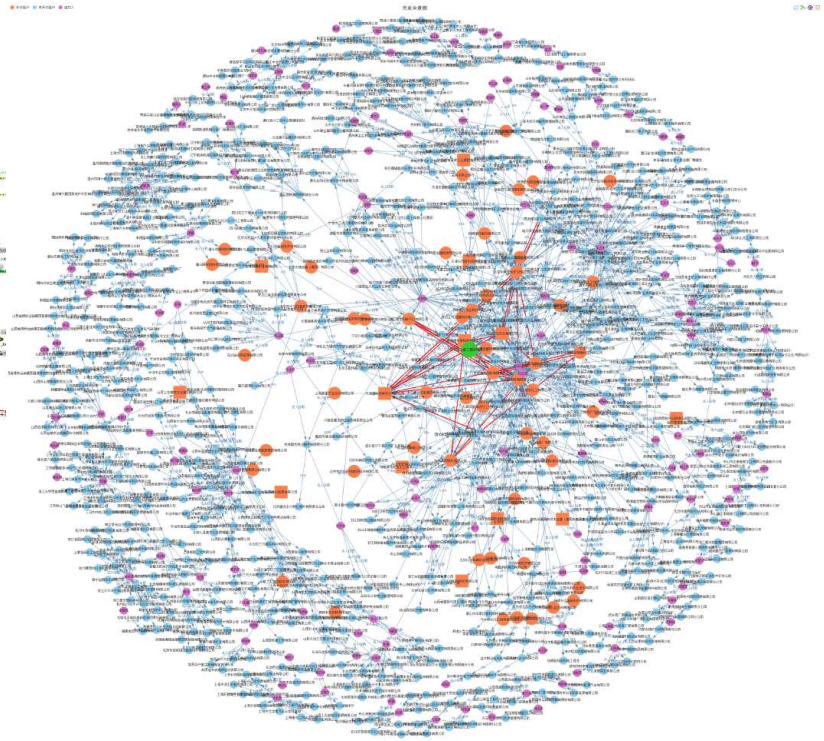
“建龙系”分析成果

- 关系复杂，易于形成系统性风险。
  股权投资的企业有321个，控制人关系的企业有537个、担保关系有243个，去重后“建龙系”关联控制企业共有886个。

- 在我行贷款项目多、涉及分行广
  该系客户族谱内属于开行客户有71个，分布在18家分行，其中在开行有贷款余额的客户有15个

- 核心企业存在过度担保、循环担保。
  行内全部71家客户担保关系总数243个，形成担保环总数94个。

- 外部风险事件
  根据舆情信息基于机器学习的数据挖掘分析，建龙系风险突发出口可能会集中在“钢铁产能过剩”、“银行抽贷”、“核心人物涉及政治风险”等热点。



预测模型构建

1）银监会反馈的全国范围内的关联关系数据

2）银监会反馈同业不良、预警；法人违约、个人违约

3）行内客户基本信息，贷款明细、账户信息，风险记录信息

4）行内客户集团关系信息，关联关系数据

5）征信披露客户欠息、垫款、资产处置等信息

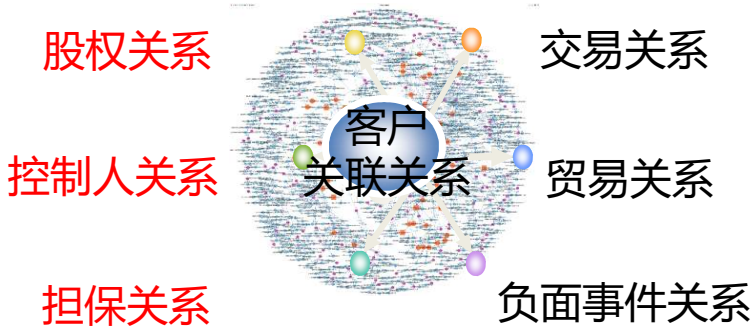
时间：201104~201604，共60月的时序数据

客户类型：一般公司类有贷客户

业务类型：所有业务类型

贷款发放状态：所有发放状态

客户区域：全国各区域客户数据



知识图谱计算

风险信号库

指标大类	指标示例	数量
客户基本属性类	包括企业规模、所属行业、首次建立业务关系时间、是否小企业等	20
客户贷款状况类	如贷款期限、贷款余额、欠息金额、逾期情况、担保金额等	72
客户贷款行为类	包括近3个月欠息次数、近6个月逾期天数、近6个月逾期凭证数等	81
关联企业投向行业类	当某贷款客户的关联企业都投向限制类行业时，则该客户在关联群中很可能充当“提款机”角色	23
关联企业风险信息类	包括关联企业不良贷款率、同一法人企业逾期贷款率、关联违约企业数等指标	95
关联关系个人类	包括法人代表、法人配偶的是否有零售贷款违约、法人零售违约笔数、客户个人股东个数等指标	9

机器学习建模

关键风险变量

指标大类	指标小类	入选个数
行内自身信贷风险	贷款管理	2
	风险缓释	1
	交易行为	1
征信交易行为风险	交易行为	2
同业交易行为	交易行为	1
关联风险	关联风险	4
	同一法人风险	3
	担保风险	1

模型验证和比较

总的违约样本数159，前100名捕获60个实际发生违约的客户，占比38%，前500名捕获83个客户，占比52%

项目	预测未来3个月		
	自身	自身+关联	模型提升值
正确率	15.94%	38.98%	144.54%
提升度	66.38	162.35	155.58%
前10名	10	7	
前20名	15	16	
前30名	23	23	
前40名	27	30	
前50名	29	34	
前60名	33	41	
前70名	35	48	
前80名	40	51	
前90名	42	57	
前100名	45	60	33.33%
前200名	51	69	
前300名	55	74	
前500名	71	83	16.90%

模型按月预测效果统计

预测月份 / 排名 / 数量	2015年								2016年			
	5月	6月	7月	8月	9月	10月	11月	12月	1月	2月	3月	4月
前10	4	4	3	2	2	1	2	7	1	1	3	6
前20	4	5	3	2	2	1	2	7	2	2	4	10
前30	5	5	4	2	2	2	3	8	2	2	4	11
前40	7	7	5	2	2	2	3	9	2	2	4	11
前50	7	7	5	2	2	3	4	9	2	2	4	11
前100	7	9	6	4	4	3	7	9	2	3	7	13
前200	7	9	7	4	6	3	7	9	3	3	8	14
前500	7	9	8	5	7	4	7	9	4	4	8	14
总风险客户数	10	9	13	9	7	10	10	9	6	6	14	14

平均每月发生风险客户13个，模型前10名捕获3个违约客户，前50名捕获4.83个，前100名捕获6.17个客户

业务验证

企业名称	泰安科诺型钢股份有限公司
组织机构代码	751773119
行内客户编号	019406
法定代表人	钱风国
企业类型	股份有限公司
成立日期	2003年6月27日
注册资金	28000万元（2014年）
企业沿革	2003年6月27日由钱占绪、尹延春、钱占武、周绪昌、钱占勇、钱占财6名自然人共同以货币资金出资组建处理泰安市科诺型钢有限责任公司。2004年9月，注册资本为27600万元，到2007年12月，公司由泰安市科诺型钢有限责任公司整体变更为泰安科诺型钢股份有限公司，注册资本28000万元，公司法定代表人钱占绪为公司实际控制人。
股东	中投亿能（北京）投资有限公司、北京中金国盈投资发展中心（有限合伙）、上海明昱投资有限公司、北京黑马汽车资讯有限公司、北京瑞银投资咨询有限公司、张继祥、张文广、赵平海、周凯、王文建、张斌、周来顺、禹世勇、唐德刚、北京瑞银投资咨询有限公司、周坤、张海琳、钱占绪、孙超、杨继平、张传祥、刘栋、巩振波、张建东、刘兴合、钱继家、赵平顺、周美海、钱广金、尹延春、钱占武、周绪昌、钱占勇、钱占财、赵勇、钱风武、张复海、李俊涛、李岩
与我行信贷关系	人求我 次级类贷款2250万元（20160630），合同有效期至2017年6月25日
担保人	中鸿联合融资担保有限公司，以保证担保形式提供担保，合同有效期：2009年6月26日-2017年6月25日
信用等级	BBB

历史预警信息

行内数据时点	同业数据时点	征信数据时点	客户编号	风险评分	风险排名	
20150531	20150430	20150531	019406	4.37%	470	
20150630	20150531	20150630	019406	24.75%	61	突变
20150731	20150630	20150731	019406	32.60%	24	
20150831	20150731	20150831	019406	36.81%	23	
20150930	20150831	20150930	019406	36.96%	25	
20151031	20150930	20151031	019406	84.05%	7	突变
20151130	20151031	20151130	019406	86.32%	6	
20151231	20151130	20151231	019406	90.17%	4	
20160131	20151231	20160131	019406	91.65%	4	
20160229	20160131	20160229	019406	95.36%	2	
20160331	20160229	20160331	019406	97.43%	1	
20160430	20160331	20160430	019406	99.78%	1	
20160531	20160430	20160531	019406	99.82%	1	

实际资产质量分类情况

数据批次	当前资产质量级别	资产质量分类大类
201503	5	关注
201504	5	关注
201505	5	关注
201506	5	关注
201507	5	关注
201508	5	关注
201509	5	关注
201510	5	关注
201511	8	关注
201512	8	关注
201601	8	关注
201602	8	关注
201603	8	关注
201604	8	关注
201605	8	关注
201606	10	次级
201607	10	次级
201608	10	次级
201609	10	次级

资产质量降为不良的原因：

1.资金周转难导致差您供应不足，加剧公司财务风险。周转资金量少、速度慢导致原材料库存骤减，开机率仅40%，以签订单不能按时供货，大量订单流失，导致市场竞争力减弱，资金回笼难，形成恶性循环，加剧财务风险。2015年前三季度，公司实现销售收入11.25亿元，较去年同期价绍38%；实现利润5678万元，利润下滑59%。

2.关联企业占款增加流动性风险。2015年10月22日，科诺型钢进行股权变更，实际控制人由钱占绪变更为钱风国（为钱占绪之子）。变更前以钱占绪为实际控制人的关联企业之间存在关联交易，关联企业之间存在大量占款，其中短期借款5.56亿元，应付票据1.45亿元，长期借款0.3亿元，债权银行共计13家。经与企业沟通了解到，科诺型钢7.73亿元负债中涉及关联企业占款约4.7亿元。在商业银行对科诺型钢及其关联企业进行抽贷的情况下，资金链急剧紧张，加剧流动性风险。

Thanks