



数据融合及挖掘

小组成员：李佳钰 李泽宁 李挺 王彪 于明飞

2019年10月24日



数据融合



数据融合示例



数据挖掘



数据挖掘示例

定义

把不同来源和不同时间点的数据自动或半自动地转换成一种形式，这种形式为人类提供有效支持或者做出自动决策。

数据融合



- 信号处理
- 概率统计
- 信息论
- 模式识别
- 模糊数学
- 人工智能

融合方法



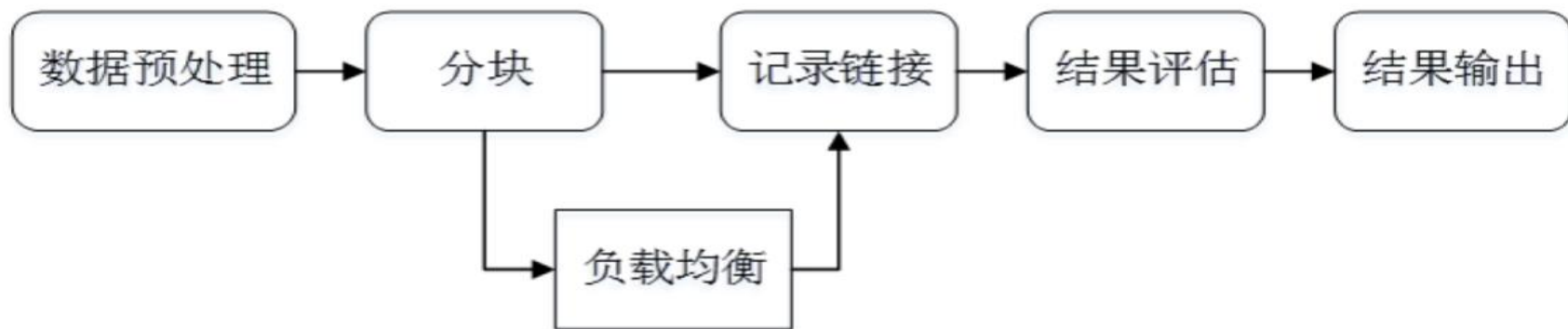
- 加权平均法
- 卡尔曼滤波法
- 多贝叶斯估计
- Dempster-Shafer (D-S)
- 证据推理
- 模糊逻辑理论
- 神经网络

定义

知识融合，即合并两个知识图谱(本体)，基本的问题都是研究怎样将来自多个来源的关于同一个实体或概念的描述信息融合起来



基本技术流程



数据预处理阶段，原始数据的质量会直接影响到最终链接的结果，不同的数据集对同一实体的描述方式往往是不相同的，对这些数据进行归一化是提高后续链接精确度的重要步骤。

- 语法正规化

语法匹配：如生日、联系电话的表示方法；

综合属性：如家庭地址的表达方式；

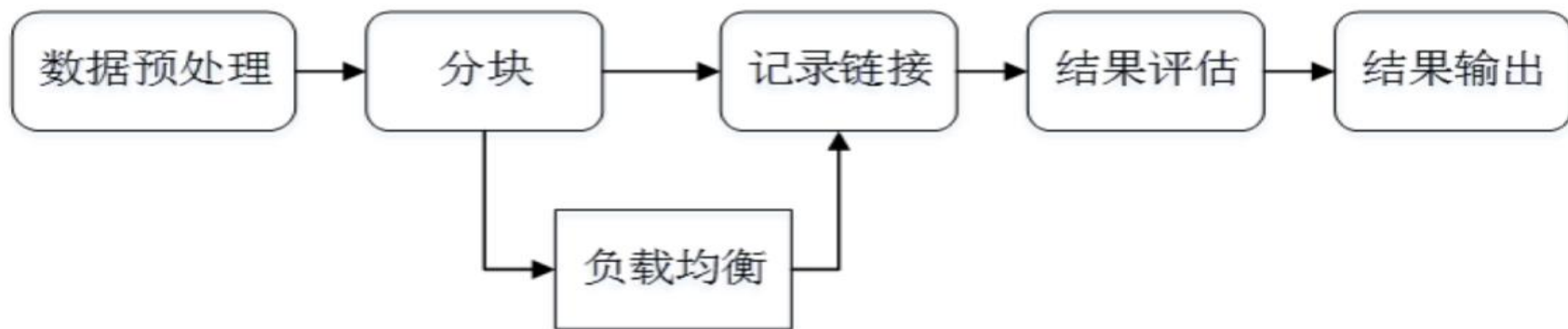
- 数据正规化

移除空格、《》、""、-等符号；

输入错误类的拓扑错误；

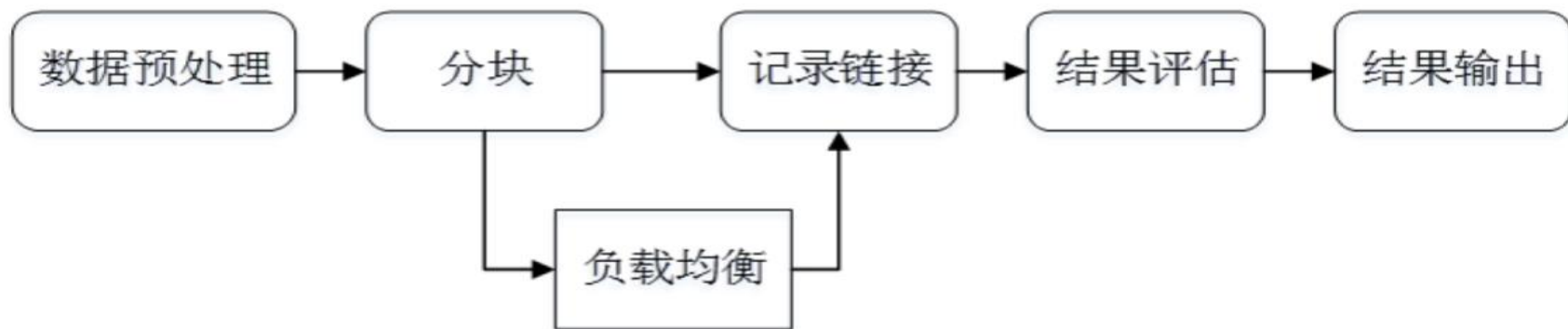
用正式名字替换昵称和缩写等。

基本技术流程



分块 (Blocking)是从给定的知识库中的所有实体对中,选出潜在匹配的记录对作为候选项,并将候选项的大小尽可能的缩小。这么做的原因很简单,因为数据太多了,我们不可能去一一连接,利用分块提高效率。常用的分块方法有基于Hash函数的分块、邻近分块等。

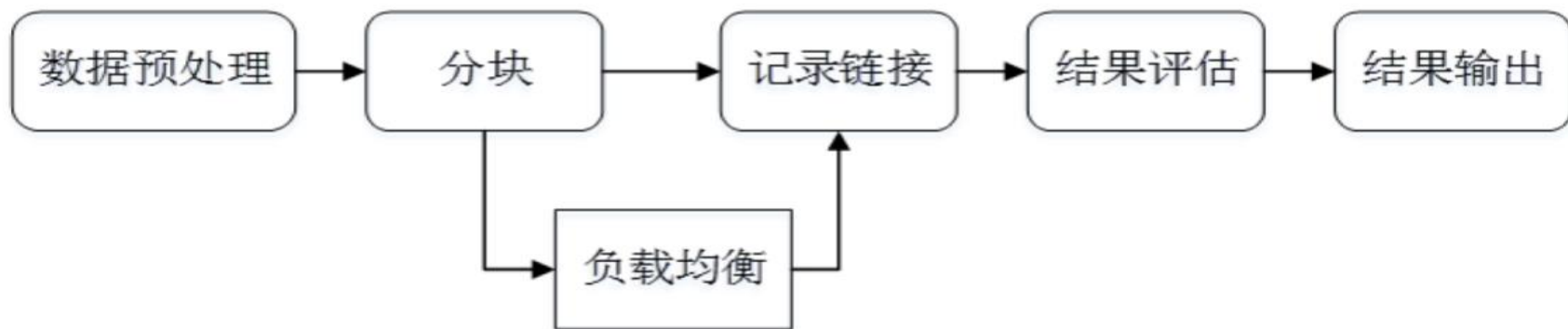
基本技术流程



负载均衡 (Load Balance)来保证所有块中的实体数目相当,从而保证分块对性能的提升程度。最简单的方法是多次Map-Reduce操作。

结果评估是对准确率、召回率、F值以及整个算法的运行时间进行评估。

基本技术流程



假设两个实体的记录 x 和 y , x 和 y 在第 i 个属性上的值是 x_i , y_i , 那么通过如下两步进行记录链接:

属性相似度: 综合单个属性相似度得到属性相似度向量

实体相似度: 根据属性相似度向量得到一个实体的相似度

知识融合

- 属性相似度

编辑距离：Levenshtein Distance、Wagner and Fisher Distance、Edit Distance with affine gaps

集合相似度计算：Dice系数、Jaccard系数

基于向量的相似度计算：TF-IDF

- 实体相似度

聚合：加权平均、手动制定规则、分类器

聚类：层次聚类、相关性聚类、Canopy + K-means

表示学习



数据融合



数据融合示例



数据挖掘



数据挖掘示例

定义 1 Definition

实体

客观存在、相互区别的事物，包括具体的人、事、物、抽象的概念或联系

2

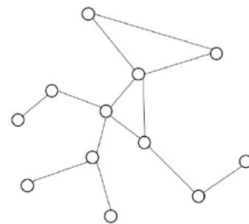
实体对齐

判断相同或不同知识库中两个实体是否表示同一物理对象的过程

例如，互动百科中的实体“刘洋（航天员）”和百度百科中的实体“刘洋（中国首位女航天员）”

图神经网络 (GNN) & 图卷积网络 (GCN)

图神经网络 (GNN)



- 图：由若干个**结点(Node)**及连接两个结点的**边(Edge)**所构成的图形
- GNN：理论基础是**不动点理论**，通过**迭代式更新**所有结点的隐藏状态，来获得每个结点的**图感知的隐藏状态**。

- 隐藏状态的状态更新函数f:

$$\mathbf{h}_v^{t+1} = f(\mathbf{x}_v, \mathbf{x}_{co}[v], \mathbf{h}_n^t e[v], \mathbf{x}_n e[v])$$

其中， \mathbf{x}_v ：结点v特征， $\mathbf{x}_{co}[v]$ ：与结点v相邻边特征， $\mathbf{h}_n^t e[v]$ ：邻居结点在t时刻隐藏状态， $\mathbf{x}_n e[v]$ ：结点v的邻居结点特征

- 局部输出函数g：根据任务需求来确定。
- 方法：通过神经网络来拟合状态更新函数f和局部输出函数g。

图神经网络 (GNN) & 图卷积网络 (GCN)

图卷积网络 (GCN)

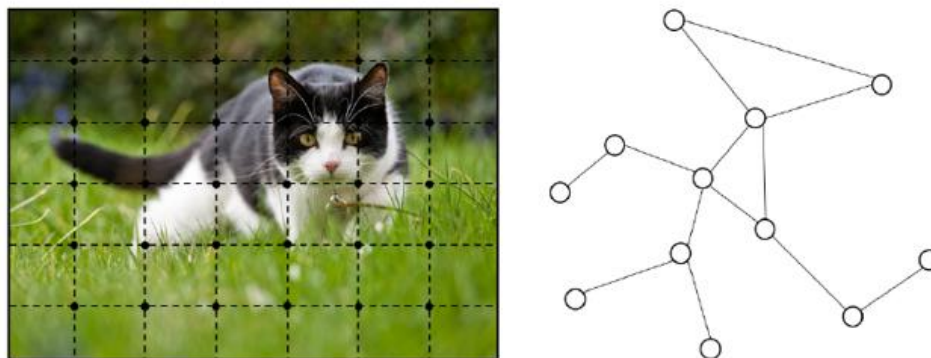
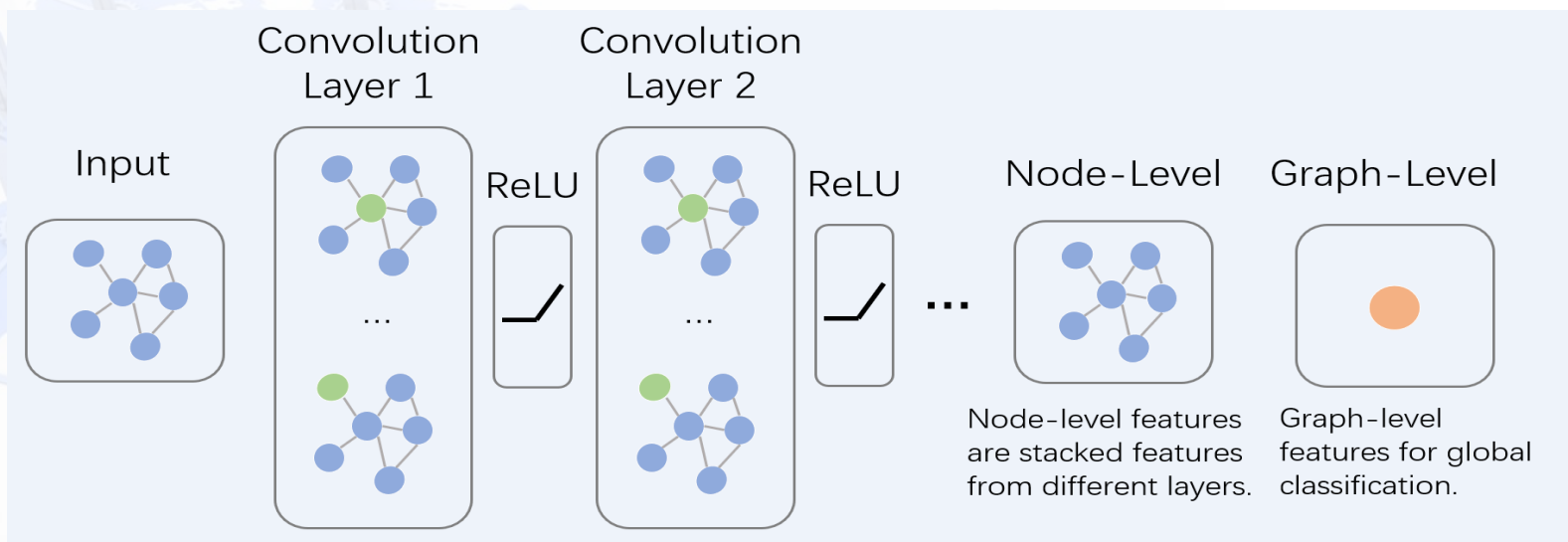


Fig. 1. Left: image in Euclidean space. Right: graph in non-Euclidean space



实验描述

- 任务：把实体与它在其他语言的对应实体进行匹配
- 数据集：DBP15K-en_zh

代码思想

- GCN把来自于不同语言的实体嵌入到一个统一的向量空间
- 属性和结构都学习一个表示
- 实体向量的相似度

技术方案

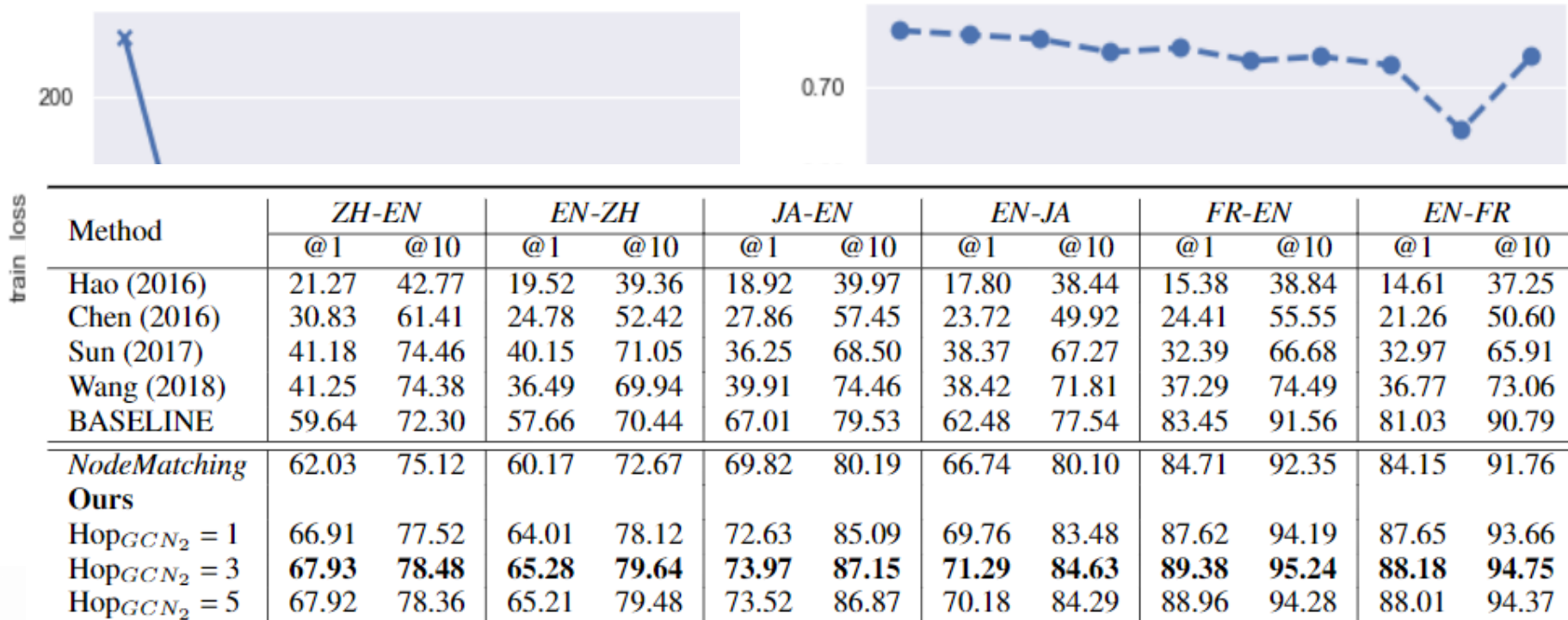
- **Input Representation Layer**
- **Node-Level (Local) Matching Layer**
- **Graph-Level (Global) Matching Layer**
- **Prediction Layer**

技术方案

- **激活函数: ReLU**
- **损失函数: margin函数**
- **优化器 : Adam**
- **输出函数: Softmax**
- **评测标准: Top-K**

跨语言KG实体对齐

实验结果



测试集上:

- Top1 准确率: **66.2%**
- Top10 准确率: **77.6%**



数据融合



数据融合示例



数据挖掘



数据挖掘示例

概念

- 处理大量粗糙的、不清晰的数据
- 挖掘出隐藏在数据中的未知而有意义的知识及信息的过程
- 来源于统计分析，借鉴了统计领域、人工智能、模式识别的相关技术
- 包括文本挖掘、图像挖掘、万维网挖掘、预测分析等众多标准术语

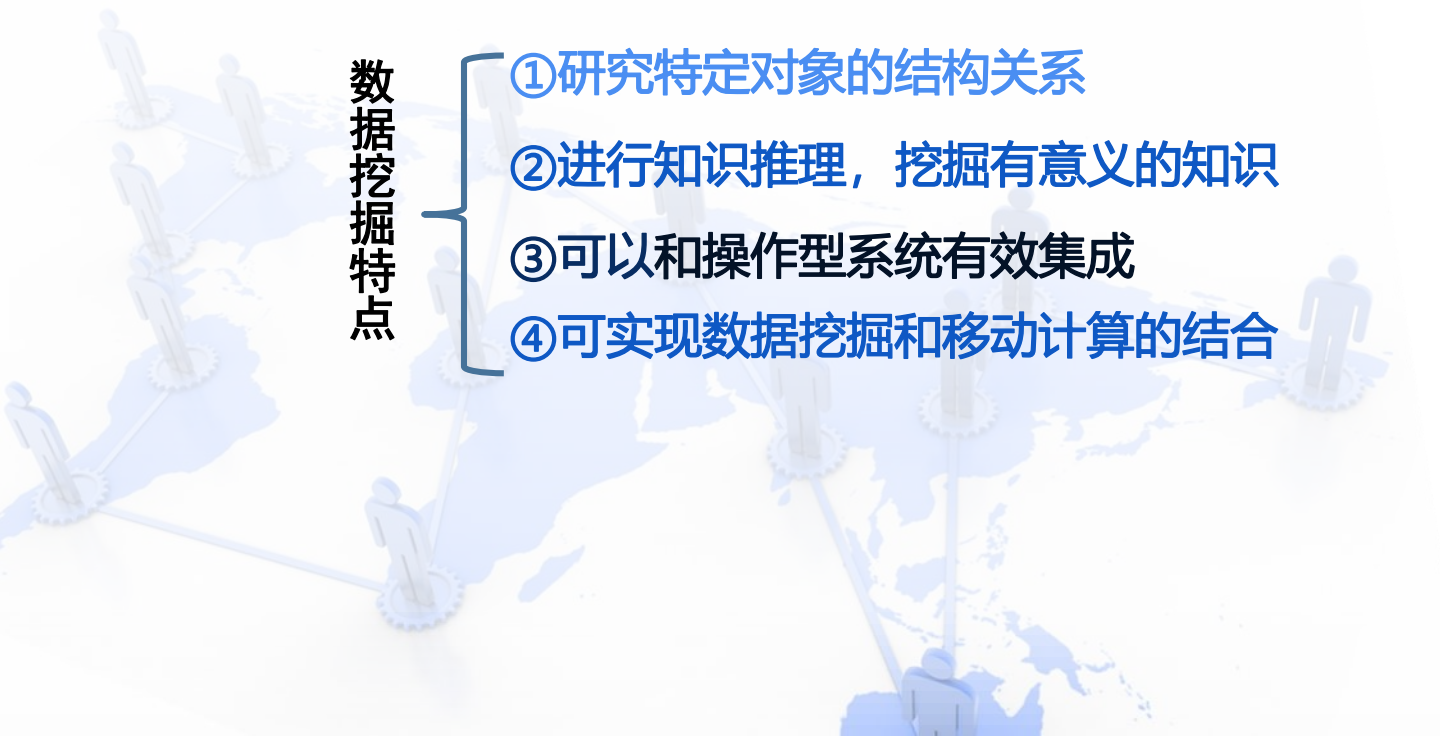


数据挖掘

特点

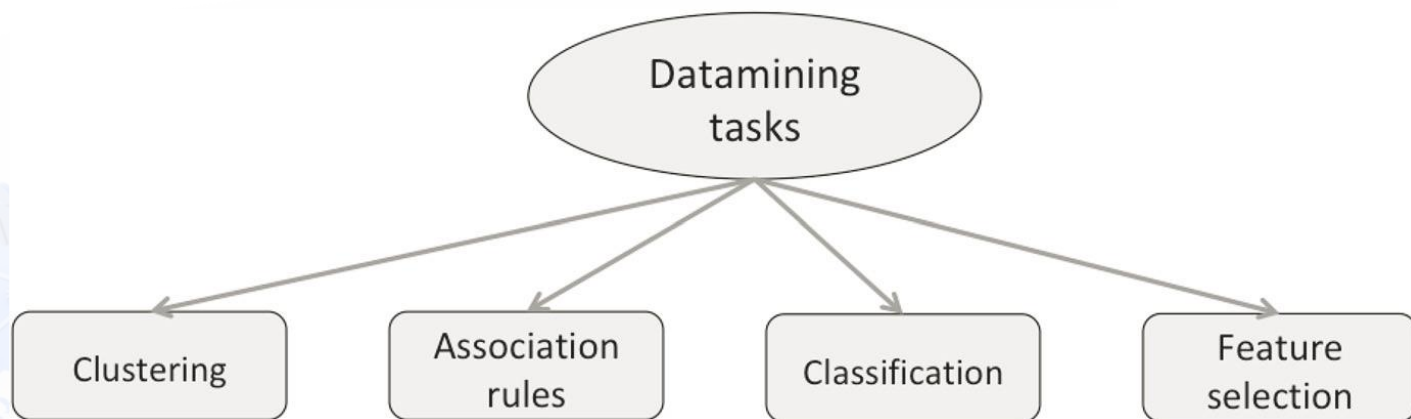
- 以海量的数据为基础
- 侧重于模型和算法
- 其具体任务包括分类与回归/相关/聚类分析，关联规则挖掘和异常检测等

数据挖掘特点

- 
- A faint background illustration shows a network of stylized human figures connected by lines, overlaid on a map of China. The figures are positioned at various points across the map, with lines connecting them to form a web-like structure.
- ①研究特定对象的结构关系
 - ②进行知识推理，挖掘有意义的知识
 - ③可以和操作型系统有效集成
 - ④可实现数据挖掘和移动计算的结合

主要任务

数据挖掘任务可以分为两类:预测(或监督)和描述(或非监督)任务, 即学习数据进行新数据预测, 和描述数据现有的关系。



聚类: 数据集分解或分组, 令组中元素间相似且尽可能不同于其他组中元素

关联规则挖掘: 提取描述不同特征之间的关系的元素

(监督)分类: 建立模型预测某未知值或从其他特性的已知值中获取目标特性(类)

特征选择: 通过选择有意义的特征来减少数据集的大小。

数据挖掘过程

数据挖掘过程要全面考虑多方面因素，包括挖掘模式种类、复杂问题解决能力、操作性能、数据存取能力等。总体上，挖掘过程可分为以下几步：

分析结果并改进：
对学习出的模型进
行分析，使得模型
达到高准确性。

数据的选取：
根据实现目标
选取合适的样
本集

数据建模：对数
据进行学习训练
，最终构造一个
训练模型。

数据的预处理：
确定样本集后对
数据进行预先处
理，使数据可用

数据转化：提取数据典型的特征，来使得提
取的特征可以准确描述数据。



数据挖掘算法

国际学术组织the IEEE International Conference on DataMining (ICDM)
2006年12月从18个数据挖掘领域影响深远的经典算法中评选出了十大经典算法：

C4.5 k-Means SVM Apriori EM PageRank
AdaBoost kNN Naive Bayes CART



IEEE

数据挖掘算法

C4.5 算法

C4.5算法是一种产生决策树的算法，它继承了ID3算法的优点并在以下方面对ID3算法进行改进：

- 1) 用信息增益率来选择属性；
- 2) 在树构造过程中进行剪枝；
- 3) 能够完成对连续属性的离散化处理；
- 4) 能够对不完整数据进行处理。

算法产生的分类规则易于理解，准确率较高。但在构造树的过程中需要对数据集进行多次顺序扫描和排序，导致算法的低效。此外，C4.5只适用于能够驻留于内存的数据集，当训练集无法在内存容纳时程序无法运行。

数据挖掘算法

Apriori算法

Apriori算法是最有影响的挖掘布尔关联规则频繁项集的算法，其核心是基于两阶段频集思想的递推算法。该关联规则在分类上属于单维、单层、布尔关联规则，所有支持度大于最小支持度的项集称为频繁项集(频集)。

Apriori演算法所使用的前置统计量包括了：

最大规则物件数：规则中物件组包含的最大物件数量

最小支援：规则中物件或是物件组必须符合的最低案例数

最小信心水准：计算规则所必须符合的最低信心水准门槛

数据挖掘算法

Apriori算法

算法的基本思想是，首先找出所有频集，其出现的频繁性至少等于预定义的最小支持度。由频集产生强关联规则，须满足最小支持度和最小可信度。使用第一步找到的频集产生只包含集合的项的所有规则，每一条规则的右部只有一项。规则生成后，只有大于用户给定的最小可信度的规则才被留下来。

算法的缺点在于可能产生大量候选集，且可能需要重复扫描数据库。

数据挖掘算法

最大期望算法

最大期望 (EM) 算法是在概率模型中寻找参数最大似然估计的算法，其中概率模型依赖于无法观测的隐藏变量。

EM算法由两个步骤交替进行计算，一是计算期望 (E)，将隐藏变量像能够观测到一样包含在内从而计算最大似然的期望值；另一步是最大化 (M)，也就是最大化在 E 步上找到的最大似然的期望值从而计算参数的最大似然估计，计算过程不断交替进行。

数据挖掘算法

PageRank算法

Google算法的重要内容，以专利人拉里·佩奇命名。算法根据网站的内外部链接的数量和质量衡量网站的价值。其背后概念是，每个到页面的链接视作对页面的投票，被链接的越多就是被其他网站投票越多，即所谓的“链接流行度”，衡量多少站点愿意和目标网站挂钩。



Google有一套自动化方法来计算这些投票，PageRank分值从0到10，10为最高，不同级别之间的差距很大。

但算法不完全基于外部链接，将PR值与排名脱钩，并降低了PageRank对更新频率。但无论如何，PR值依旧是网站质量的重要参考。



数据挖掘算法

XGBoost算法

一种可拓展树提升系统且隶属于Boosting算法，常用来解决监督学习任务。Boosting算法的思想是学习器的积弱成强，常用的弱学习器有决策树、分类回归树等。

XGBoost算法思想是不断添加分类回归树，不断根据树节点上某特征值进行分裂来构建树模型，直到达到终止条件；添加树便是学习一个新函数，来拟合上次预测结果的残差。

数据挖掘算法

XGBoost算法

算法为了防止过拟合，除在目标函数中加入惩罚项外，还提出了收缩和列特征抽样。

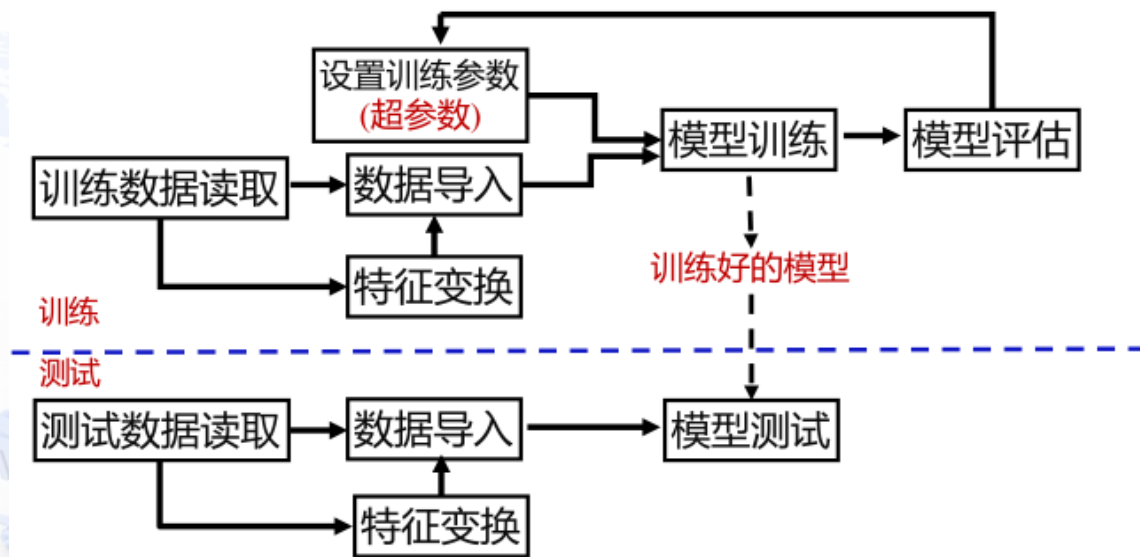
收缩方法是在迭代过程中对每个叶子节点的预测分数乘以权重系数，降低了每棵树的权重系数，为优化后面新生成树留出空间。

列特征采样有两种方案，一是按不同层进行随机采样，二是对所有特征随机选择。

数据挖掘算法

XGBoost算法

XGBoost建模过程与处理数据科学任务流程基本相同。训练数据已进行预处理，模型评估函数与ARIMA模型使用的评估函数相同，使用均方根误差和平均绝对百分误差。



数据挖掘应用现状

商务智能 (BI) : MIS系统在商业的普遍使用, 可以为有关方提供有效决策, 促进销售, 提高竞争力。

Web搜索引擎: 可以产生智能搜索引擎, 给用户提供高效准确的Web检索工具。

金融领域: 突出表现在信用评估和防止欺诈等方面, 以及投资评估和股票交易市场预测。

数据挖掘还可用于工业、农业等其它行业, 也可应用于决策支持和数据库管理系统中; 而作为决策支持和分析工具, 能用来构造知识库。在DBMS中, 数据挖掘可以用于语义查询优化、完整性约束和不一致检验等。



数据融合



数据融合示例



数据挖掘



数据挖掘示例

任务简介:

- 推荐系统是数据挖掘领域的经典问题，可以为企业带来巨大的经济效益。
- 关联规则法是解决推荐问题的经典算法之一，通过生成一些规则来实现推荐功能，这些规则的形式如下：

如果用户喜欢某些物品，那么他们也会喜欢这个物品

- 使用Apriori-关联规则算法在MovieLens数据集上实验，实现电影的推荐功能



数据准备



数据挖掘



结果评价

MovieLens: 明尼苏达大学公开的用于测试推荐算法的数据集，包含了100万条电影评分数据，每条数据有以下四个字段：

UserID	MovieID	Rating	Datetime
196	242	3	1997-12-04 15:55:49
186	302	3	1998-04-04 19:22:22
22	377	1	1997-11-07 07:18:36
244	51	2	1997-11-27 05:02:03
166	346	1	1998-02-02 05:33:16

要实现推荐功能，就要确定用户是否喜欢某一部电影，因此要对数据进行预处理。

增加新特征Favorable，若用户的评分大于3，则认为该用户喜欢这部电影，该条记录的Favorable为真。处理后的数据如下图所示：

UserID	MovieID	Favorable
62	257	False
286	1014	True
200	222	True
210	40	False
224	29	False

删去Favorable为False的记录，数据预处理完成

- Apriori算法查找频繁项集

频繁项集：支持度大于给定值的电影集合

支持度：数据集中规则应验的次数，在本实验中指喜欢该电影的用户数量。

取最小支持度为50，在数据集上执行Apriori算法后，得到2968个频繁项集

- 根据频繁项集抽取关联规则
关联规则：如果用户喜欢某些电影，那么他们也会喜欢另一部电影。

频繁项集{1, 2, 3, 4}, 可以产生4条规则：
喜欢某些电影{1, 2, 3}的用户，也会喜欢电影4
喜欢某些电影{1, 2, 4}的用户，也会喜欢电影3
喜欢某些电影{1, 3, 4}的用户，也会喜欢电影2
喜欢某些电影{2, 3, 4}的用户，也会喜欢电影1

- 实验结果：共产生15285条候选规则

- 计算每条候选规则的置信度，删除掉置信度小于1的候选规则后，得到5152条关联规则
前五条如下所示：

Rule #1

Rule: 评论了 `frozenset({64, 98, 56, 50, 7})` 的人，他也会评论 174
- 置信度Confidence: 1.000

Rule #2

Rule: 评论了 `frozenset({98, 100, 172, 79, 50, 56})` 的人，他也会评论 7
- 置信度Confidence: 1.000

Rule #3

Rule: 评论了 `frozenset({98, 172, 181, 174, 7})` 的人，他也会评论 50
- 置信度Confidence: 1.000

Rule #4

Rule: 评论了 `frozenset({64, 98, 100, 7, 172, 50})` 的人，他也会评论 174
- 置信度Confidence: 1.000

Rule #5

Rule: 评论了 `frozenset({64, 1, 7, 172, 79, 50})` 的人，他也会评论 181
- 置信度Confidence: 1.000

结果评价

- 交叉验证法：前200名用户对电影的评价作为训练集，剩下的数据作为测试集
- 置信度：衡量规则的准确率，统计方式为：当前规则的出现次数/条件相同的规则数量

把电影编号替换为电影名称后，前五条规则的评估结果如下：



结果评价

Rule #1

Rule: 评论了 Shawshank Redemption, The (1994), Silence of the Lambs, The (1991), Pulp Fiction (1994), Star Wars (1977), Twelve Monkeys (1995) 的人, 他也会评论 Raiders of the Lost Ark (1981)

- 训练集上的置信度: 1.000
- 测试集上的置信度: 0.909

Rule #2

Rule: 评论了 Silence of the Lambs, The (1991), Fargo (1996), Empire Strikes Back, The (1980), Fugitive, The (1993), Star Wars (1977), Pulp Fiction (1994) 的人, 他也会评论 Twelve Monkeys (1995)

- 训练集上的置信度: 1.000
- 测试集上的置信度: 0.609

Rule #3

Rule: 评论了 Silence of the Lambs, The (1991), Empire Strikes Back, The (1980), Return of the Jedi (1983), Raiders of the Lost Ark (1981), Twelve Monkeys (1995) 的人, 他也会评论 Star Wars (1977)

- 训练集上的置信度: 1.000
- 测试集上的置信度: 0.946

Rule #4

Rule: 评论了 Shawshank Redemption, The (1994), Silence of the Lambs, The (1991), Fargo (1996), Twelve Monkeys (1995), Empire Strikes Back, The (1980), Star Wars (1977) 的人, 他也会评论 Raiders of the Lost Ark (1981)

- 训练集上的置信度: 1.000
- 测试集上的置信度: 0.971

Rule #5

Rule: 评论了 Shawshank Redemption, The (1994), Toy Story (1995), Twelve Monkeys (1995), Empire Strikes Back, The (1980), Fugitive, The (1993), Star Wars (1977) 的人, 他也会评论 Return of the Jedi (1983)

- 训练集上的置信度: 1.000
- 测试集上的置信度: 0.900

Thanks for Listening

