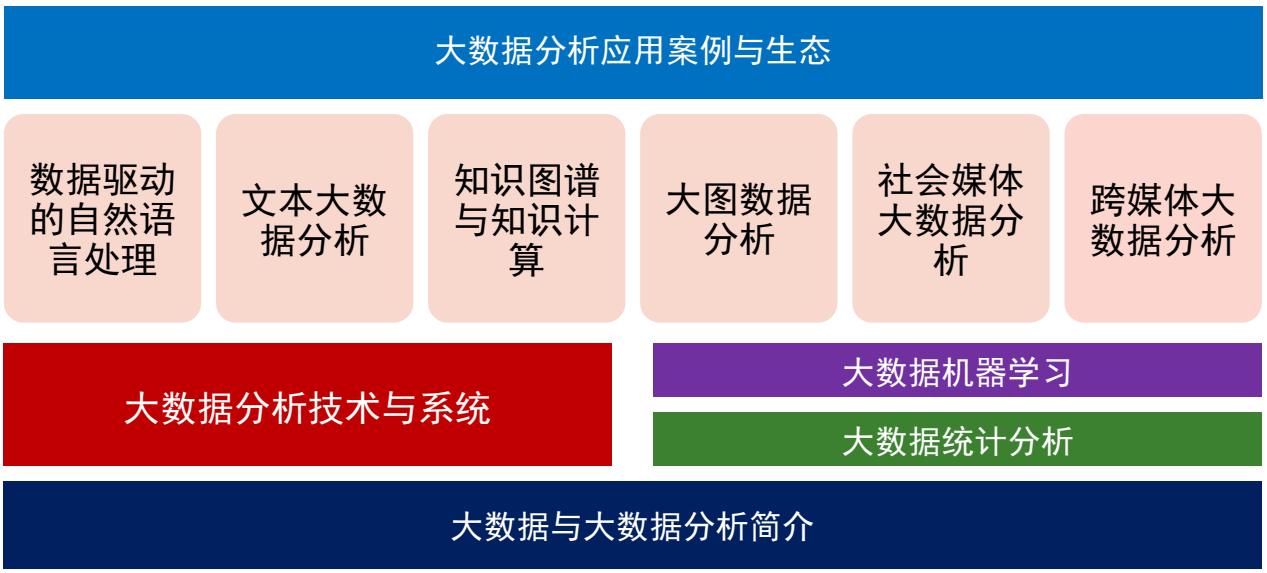
大数据分析

课程总复习-Part2

程学旗
靳小龙
刘盛华

课程内容安排



Outline

- 大数据与大数据分析简介
- 大数据分析技术与系统
- 大数据统计分析
- 大数据机器学习
- 数据驱动的自然语言处理
- 文本大数据分析
- 知识图谱与知识计算
- 大图挖掘与分析
- 社会媒体分析

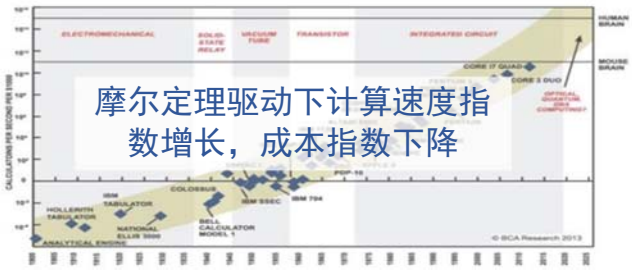
Outline

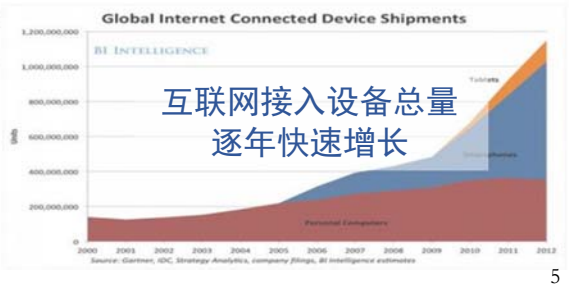
- 大数据与大数据分析简介
- 大数据分析技术与系统
- 大数据统计分析
- 大数据机器学习
- 数据驱动的自然语言处理
- 文本大数据分析
- 知识图谱与知识计算
- 大图挖掘与分析
- 社会媒体分析

互联网及其延伸导致大数据现象

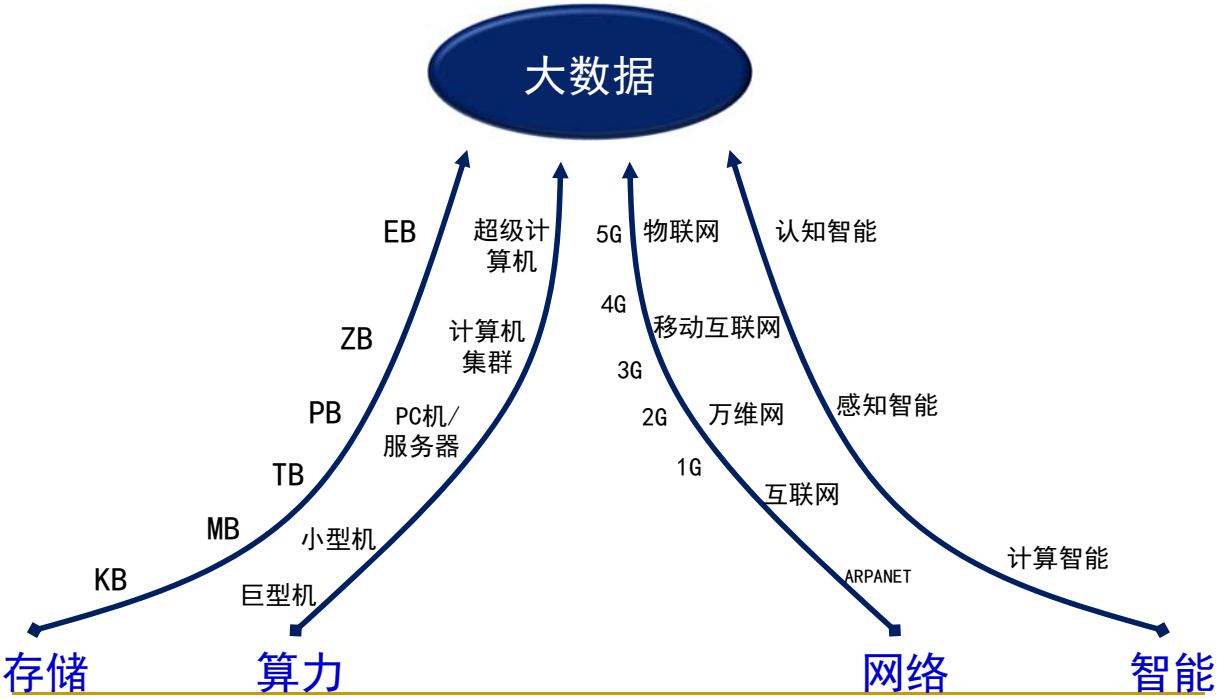
大数据源于信息技术的不断廉价化与互联网及其延伸所带来的无处不在的信息技术应用。四个驱动：

- 摩尔定律驱动的指数增长模式；
- 技术低成本化驱动的万物数字化；
- 宽带、移动、泛在互联驱动的人机物广泛联接；
- 云计算模式驱动的数据大规模汇聚；





大数据是存储、算力、网络、智能发展的产物



何为大数据？

技术能力的视角

- 大数据指的是规模超过现有数据处理工具获取、存储、管理和分析能力的数据集。并不是超过某个特定数量级的数据集才是大数据。

——麦肯锡《大数据：创新、竞争和生产力的下一个前沿领域》

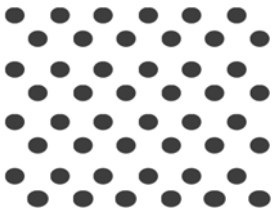
大数据内涵的视角

- 大数据是具备海量、高速、多样、可变等特征的多维数据集，需要通过可伸缩的体系结构实现高效的存储、处理和分析。

——NIST《大数据白皮书》

大数据的基本特征

Volume



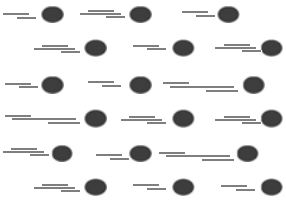
体量巨大

From terabytes to exabyte to zetabytes of existing data to process

IDC Analyze the Future

到2020年，数据总量达40ZB，人均5.2TB

Velocity



速度极快

Batch data, near-time data, real time data, streaming data, miliseconds to seconds to respond

facebook

分享的内容条目超过25亿个/天，增加数据超过500TB/天

Variety

模态多样

Structured, semi-structured, unstructured, text, pictures, multimedia

Veracity

真伪难辨

Uncertainty due to data inconsistency & incompleteness, ambiguities, deception, model approximation

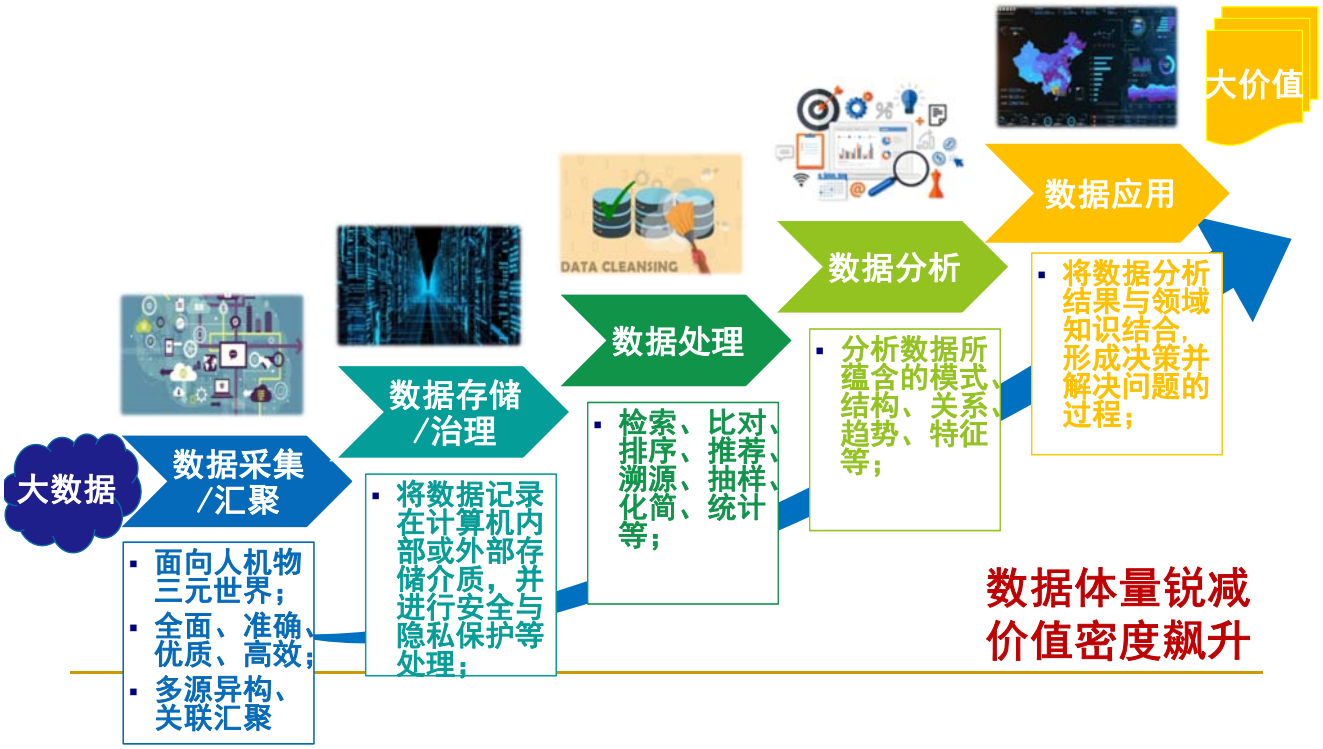
芦山地震十大不实谣言

2013年04月24日17:41　来源：人民网　手机看新闻

8.地震局内部消息成都9.2级地震

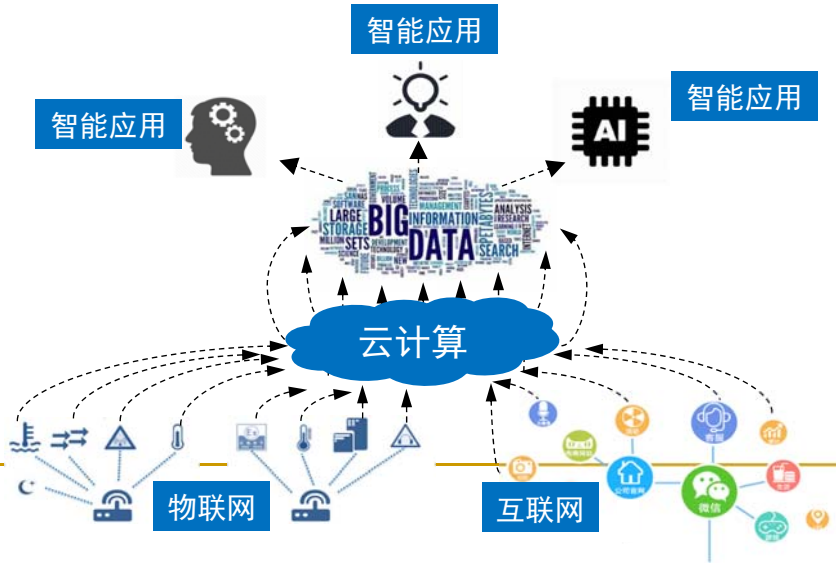
谣言：自称地震局内部人员的网民称发生9.2级地震。”

大数据的大价值（Value）



大数据与物联网、互联网、云计算、人工智能

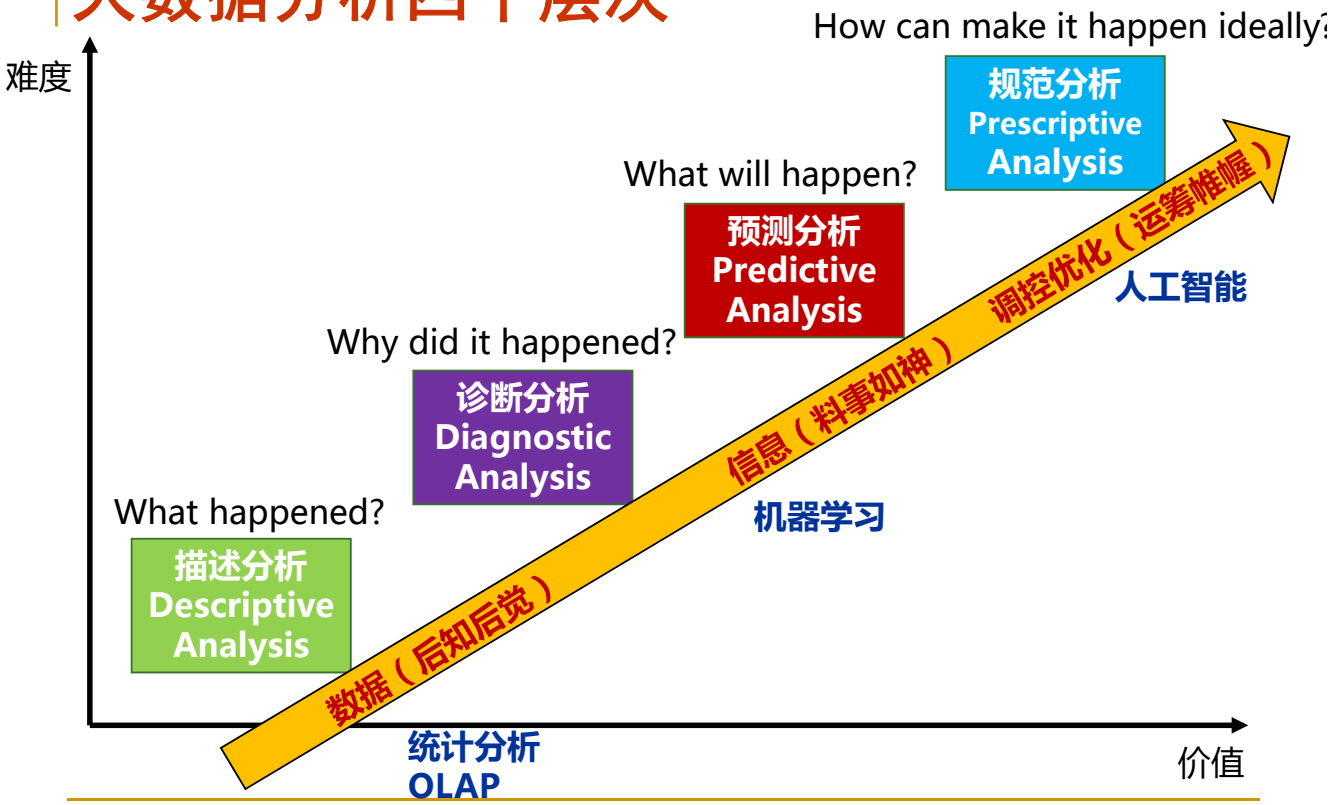
- 物联网和互联网产生的数据成为大数据的主要来源
- 云计算为大数据运算提供核心硬件和软件支撑
- 人工智能是驱动创新应用的重要引擎
- 大数据为人工智能引擎提供必要的燃料，二者缺一不可



什么是大数据分析？

- 大数据分析(Big Data Analysis)：利用各种不同的理论、技术和方法对大数据进行理解，提取其价值的过程。
  - 大数据的价值体现为更紧致的数据/信息/知识、某种现象或规律，或者其他的观察或结论；
  - 大数据分析的终极目标是实现对目标对象的认知，进而提供对策建议；
  - 大数据分析涉及诸多不同领域中的理论、技术和方法，包括应用数学、概率统计、机器学习、数据挖掘、图像识别、人工智能、数据可视化、数据仓库以及高性能计算等。

大数据分析四个层次



Outline

- 大数据与大数据分析简介
- 大数据分析技术与系统
- 大数据统计分析
- 大数据机器学习
- **数据驱动的自然语言处理**
- 文本大数据分析
- 知识图谱与知识计算
- 大图挖掘与分析
- 社会媒体分析

内容

- 关键技术
  - 语法分析
    - 词法分析
    - 句法分析
  - 语义分析
    - 语义表示
  - 内容分析
    - 信息抽取
    - 文本分类
    - 情感分析
  - 机器翻译
  - 问答系统

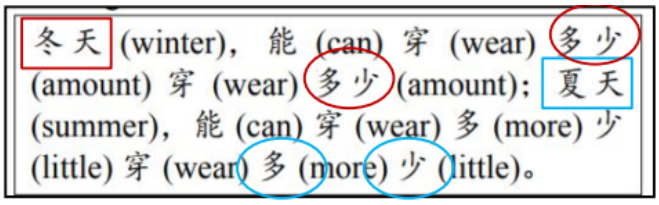
中文分词

- 中文以字为基本书写单位，词语之间没有明显的区分标记
- 中文分词就是要由计算机在文本中词与词之间加上标记
  - 输入：我是中国科学院大学的学生。
  - 输出：我/是/中国科学院大学/的/学生/。
  - 输入：大数据分析是什么？
  - 输出：大数据分析/是/什么/？

- 和中文相比，英语切分问题较为容易
  - 识别出英语文本中的词（tokenization）
    - Let’s go to school.
    - I like programming.
  - 对识别出的词进行词形还原（lemmatization）
    - boys -> boy
    - strongest -> strong

基于长短时记忆网络的中文分词

- 传统的统计方法严重依赖于特征的设计，而手工提取特征非常地费时费力，并且还存在错误传递问题
- 对一个句子正确地切分，需要能够捕捉一些远距离的信息



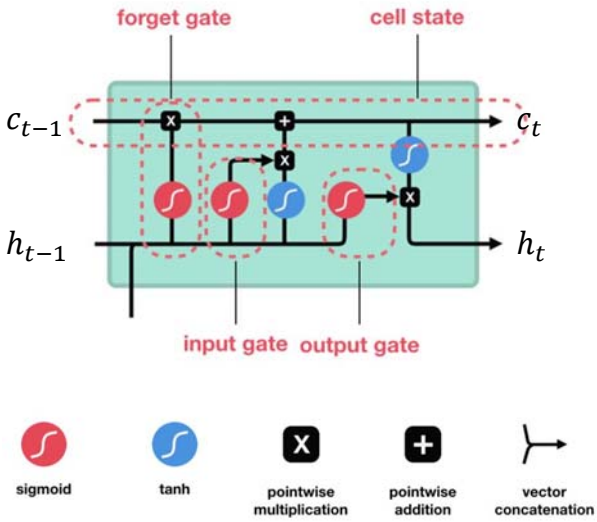
如果不能利用“冬天”和“夏天”，很难对“能穿多少穿多少”进行分词。
LSTM网络可以很好地学习长距离信息。

- Chen等人将长短时记忆网络(Long-Short Term Memory Network, LSTM)用于中文分词

来源：Xinchi Chen et al, Long Short-Term Memory Neural Networks for Chinese Word Segmentation, EMNLP 2015

长短时记忆网络(LSTM)

- LSTM的细胞结构
  - 两个状态向量c和h，c是LSTM的内部状态向量，h是LSTM的输出向量
  - 三个门：遗忘门、输入门、输出门



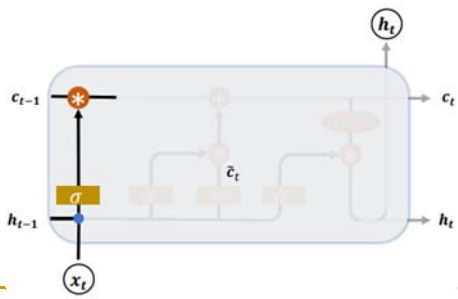
来源：Hochreiter, Sepp, and Jürgen Schmidhuber, Long short-term memory, Neural computation 9.8 (1997)

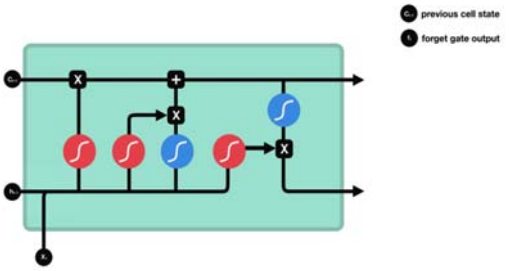
长短时记忆网络(LSTM)

- 遗忘门
  - 遗忘门的功能是决定应丢弃或保留哪些信息。作用于LSTM的状态向量c上，用于控制上一时间戳的记忆c_{t-1}对当前时间戳的影响
  - 具体将前一隐藏状态的信息h_{t-1}和当前输入信息x_t同时传递到sigmoid函数中去，得到遗忘门门控向量：

$$g_f = \sigma(W_f[h_{t-1}, x_t] + b_f)$$

  - 经过遗忘门后，LSTM的状态向量变为：$g_f \odot c_{t-1}$





长短时记忆网络(LSTM)

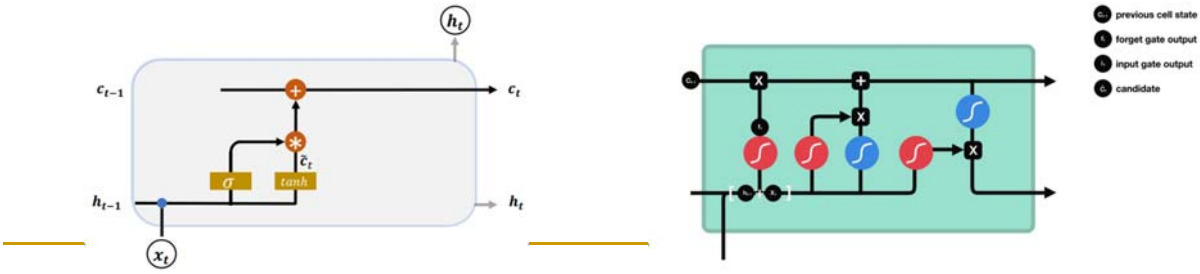
- 输入门
  - 用于更新细胞状态，亦可认为是控制LSTM对输入的接收程度
  - 首先，通过对将前一隐藏状态的信息h_{t-1}和当前输入信息x_t做非线性tanh变换，得到新的输入$\tilde{c}_t$量

$$\tilde{c}_t = tanh(W_c[h_{t-1}, x_t] + b_c)$$

  - 其次，将前一隐藏状态的信息h_{t-1}和当前输入信息x_t输入sigmoid得到输入门门控向量：

$$g_i = \sigma(W_i[h_{t-1}, x_t] + b_i)$$

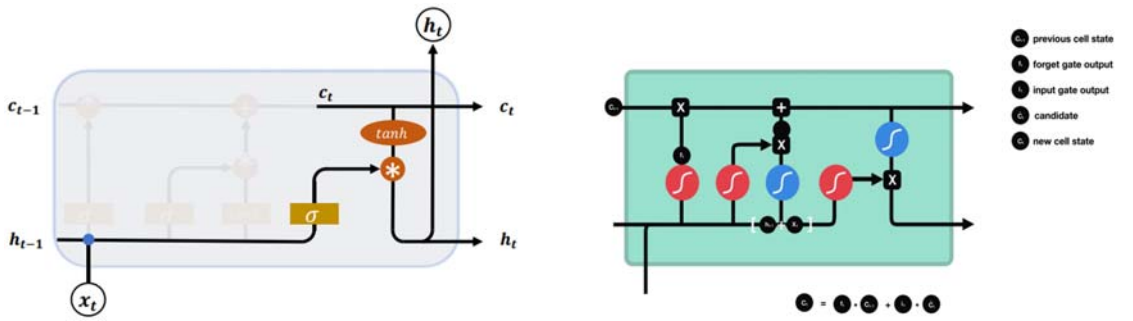
  - 经过输入门后，待写入记忆的向量为：$g_i \odot \tilde{c}_t$



长短时记忆网络(LSTM)

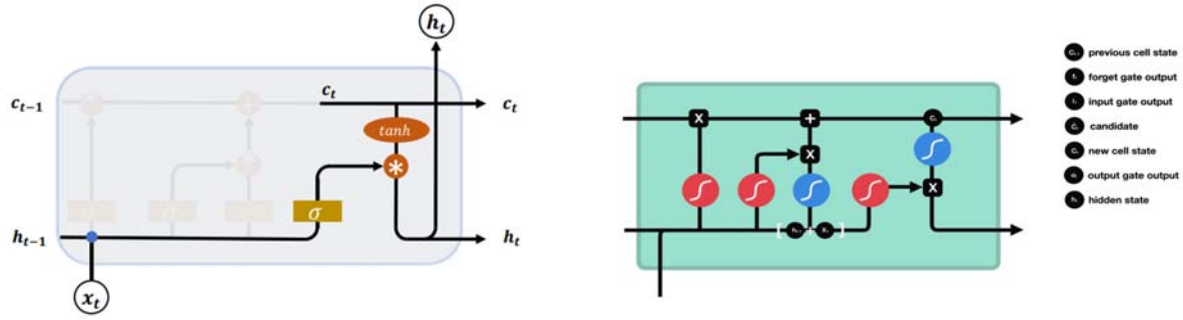
- 更新细胞状态
  - 在遗忘门和输入门的控制下，LSTM接收来自二者的新输入，逐点相加后更新得到当前时间戳的状态向量c_t：

$$c_t = g_i \odot \tilde{c}_t + g_f \odot c_{t-1}$$



长短时记忆网络(LSTM)

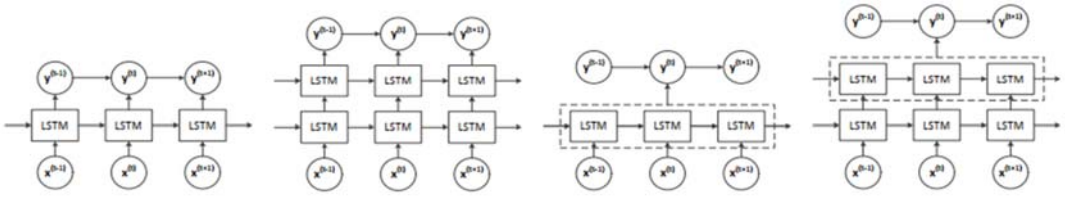
- 输出门
  - 输出门用来确定下个隐藏状态的值，它包含了先前输入的信息
  - 首先，将前一隐藏状态和当前输入传递到sigmoid函数中，得到输出门门控向量：$g_o = \sigma(W_o[h_{t-1}, x_t] + b_o)$
  - 然后，将新得到的细胞状态传递给tanh函数：$\tanh(c_t)$
  - 最后，将tanh的输出与输出门门控向量逐点相乘，确定隐藏状态应携带的信息：$h_t = g_o \odot \tanh(c_t)$



  - 将隐藏状态h_t作为当前细胞的输出，把新的细胞状态c_t和新的隐藏状态h_t传递到下一个时间步长中去

基于长短时记忆网络的中文分词

- LSTM分析模型结构
  - 该模型首先将输入的句子中的每个字都映射成向量表示
  - 然后分别使用四种不同的LSTM+移动窗口的结构提取特征
    - 只使用一层LSTM
    - 只使用双层LSTM
    - 使用一层LSTM，并且联接窗口内LSTM单元输出
    - 使用双层LSTM，并且联接窗口内顶层LSTM单元输出



(a) LSTM-1 (b) LSTM-2 (c) LSTM-3 (d) LSTM-4

  - 通过标签推理层(LSTM层的输出H作为输入)输出每个字对应的标签
    - 进行推断时，输出得分最高的一组句子标注

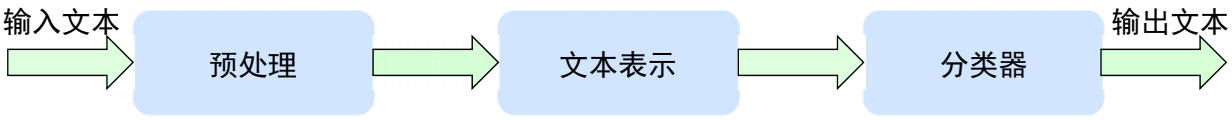
$$s(c^{(1:n)}, y^{(1:n)}, \theta) = \sum_{t=1}^{n} \left(\mathbf{A}_{y^{(t-1)}y^{(t)}} + \mathbf{y}^{(t)}_{y^{(t)}} \right)$$

文本分类

- 文本分类是在自然语言理解中一项基础的任务，是指将一段文本划分到预定义的标签类别中

- 文本分类的效果会直接影响一些下游任务，如：
  - 话题分析
  - 问答系统
  - 自然语言推理

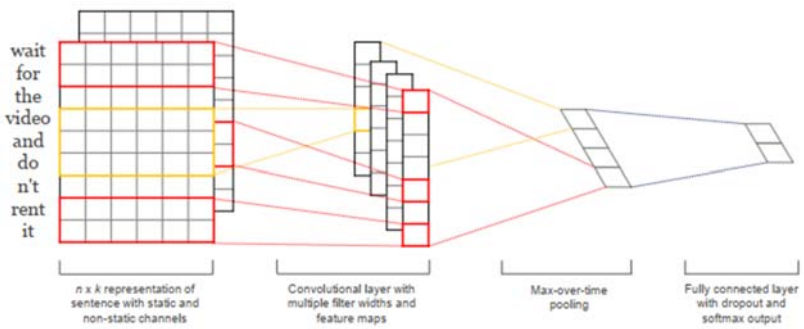
- 文本分类系统



TextCNN模型

- 动机
  - 在由单词的分布式表示组合成的句子表示上，卷积神经网络可以利用过滤器捕捉局部特征
  - 基于卷积神经网络的模型在其他一些传统自然语言处理任务上都取得了不错的效果

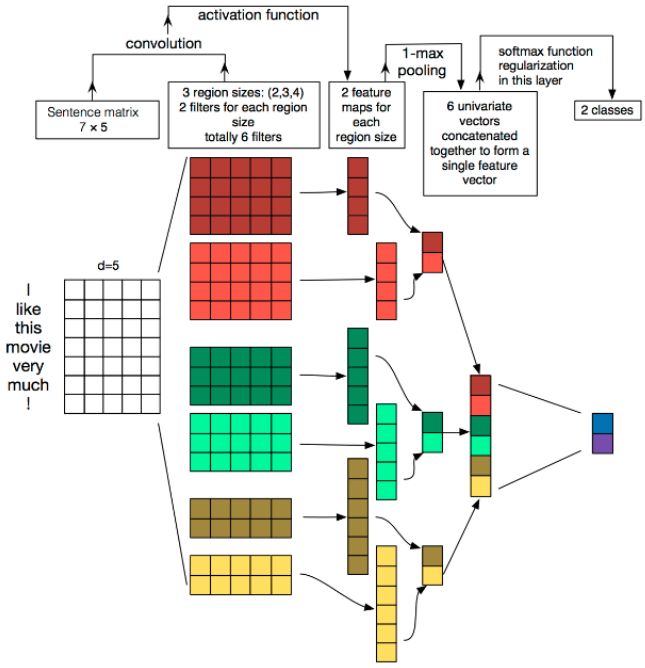
- 模型结构
  - 预训练词向量（Word2Vec, Glove）
  - 卷积层
  - 池化层
  - 全连接层
  - Dropout
  - Softmax



来源：Yoon Kim, Convolutional Neural Networks for Sentence Classification, EMNLP 2014

TextCNN模型

- Embedding：第一层是图中最左边的7乘5的句子矩阵，每行是词向量，维度=5，这个可以类比为图像中的原始像素点
- Convolution：然后经过kernel_sizes=(2,3,4)的一维卷积层，每个kernel_size有两个输出channel
- MaxPolling：第三层是一个1-max pooling层，这样不同长度句子经过pooling层之后都能变成定长的表示
- FullConnection and Softmax：最后接一层全连接的softmax层，输出每个类别的概率



来源：Yoon Kim, Convolutional Neural Networks for Sentence Classification, EMNLP 2014

情感分析

- 情感分析是一种重要的信息组织方式，其研究目标是自动挖掘和分析文本中的立场、观点、看法、情绪和喜恶等主观信息；
- 随着微信、微博、论坛和社交网络等应用的兴起，社交网络上汇聚了海量信息。情感分析在社会管理、商业决策、信息预测等各个方面有着广泛而重要的应用价值。

社交媒体上存在大量包含用户情感的文本

情感词抽取

- 基于启发式规则的方法
  - 利用连词，and、but、either…or…和neither…nor…等
    - Every one wants to be *healthy* **and** *happy*
    - *Delicious* food **but** *bad* service
  - 利用程度副词，very
  - 利用评价对象词

基于启发式规则的方法优点是比较简单，针对性强；缺点在于人工定义的规则具有局限性，可扩展性差。

- 基于统计的方法
  - 话题模型
  - 图模型

基于统计的方法抽取出的情感词召回率高，而基于规则的方法抽取出的情感词准确率高，因此在实际工程任务中，往往采用规则和统计相结合的方法。

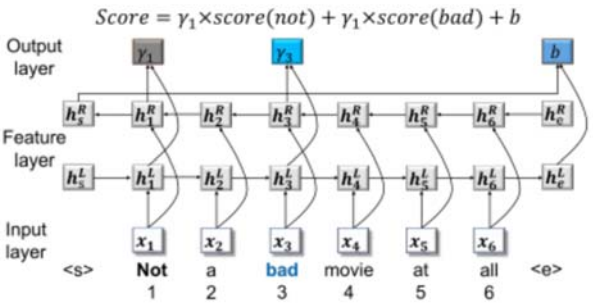
情感词极性判定

- 情感词的极性判定工作大致可以分为三类
  - 基于知识库的方法
    - Wordnet、Hownet，同义词关系、反义词关系、上下位关系等
    - 只能获得一个通用的情感词典，无法获得领域依赖的情感词典
  - 基于语料库的方法
    - 基于假设：具有相同情感倾向性的情感词通常出现在同一句子中
    - 例如，“excellent”和“poor”的互信息差方法
  - 知识库与语料库相结合的方法
    - 半监督学习框架，利用现有知识库作为先验知识，提供精确的种子词集，并结合语料库中的词汇关系，构建更加庞大的情感词典

句子级情感分析

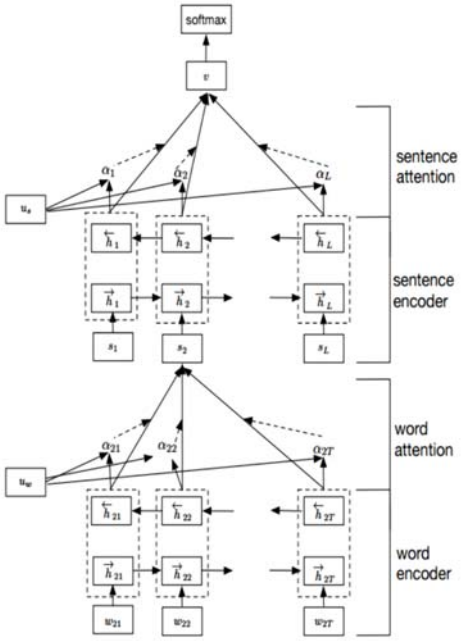
- 句子级情感分析是指对单句的情感极性进行判断的任务
- 基于深度神经网络的方法主要是对句子所包含的语义信息进行表示，进而对其情感极性进行判别
- 常用的方法有基于CNN、RNN、Recursive-NN的方法以及最新的Transformer等

The performances were all **really** It's an **insignificant** [criticism]$_{-1\rightarrow-0.5}$.
Nobody gives a [good]$_{+3\rightarrow-1}$ performance in this movie
She's **not** [terrific]$_{+5\rightarrow+1}$ but **not** [terrible]$_{-5\rightarrow-1}$ either.
It's **not** a **very** [good]$_{+3\rightarrow-0.25}$ movie song!
It **removes** my [doubts]$_{-3\rightarrow+1}$.



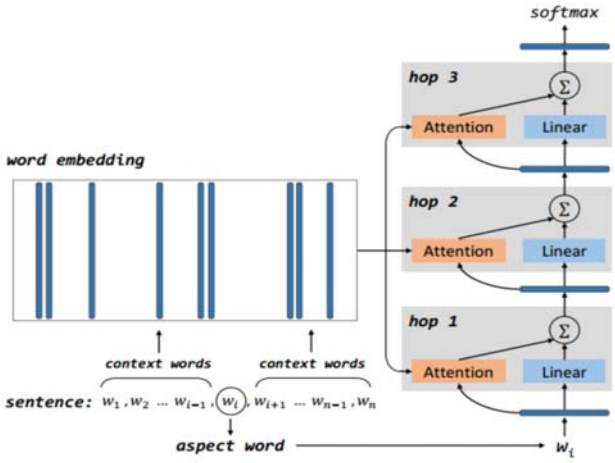
篇章级情感分析

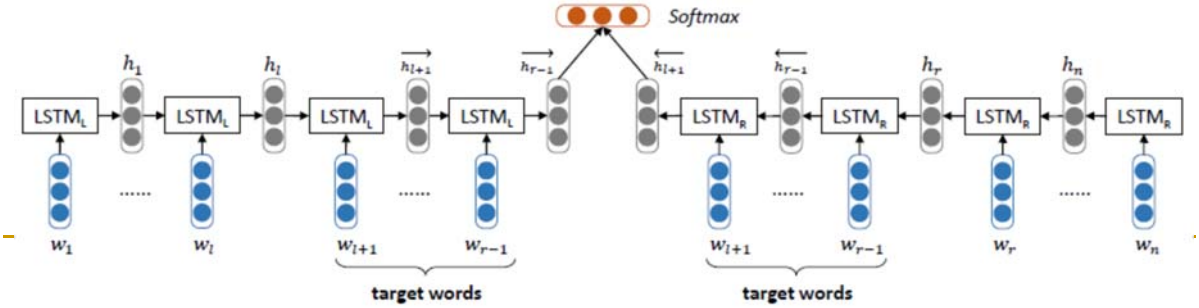
- 篇章级情感分析是指对整篇文档的全局情感极性进行分析判断
- 基本思想是对词->句子->篇章逐层进行语义编码表示，得到篇章的向量表示
- 让机器像人一样对篇章进行阅读：从简单的字词组成句子，句子组成篇章，最后形成思想，这就是自然语言处理中的层级（Hierarchical）概念



属性级情感分析

- 属性级情感分析（Aspect-Level）是细粒度情感分析，指对所描述事物的属性的情感极性进行判断
  - 我非常喜欢这款手机的摄像头
- 基本思想是对属性词和其上下文进行表示，并建立它们之间的关系，进而判断情感极性





Outline

- 大数据与大数据分析简介
- 大数据分析技术与系统
- 大数据统计分析
- 大数据机器学习
- 数据驱动的自然语言处理
- **文本大数据分析**
- 知识图谱与知识计算
- 大图挖掘与分析
- 社会媒体分析

内容

- 文本表达
  - 单词表达方法、句子表达方法
- 文本匹配
  - 基于规则的文本匹配、基于学习的文本匹配
- 文本生成
  - 文本生成任务、方法与评价方式

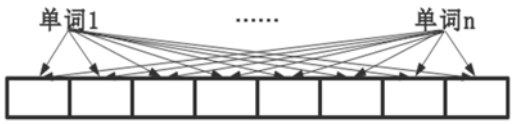
单词的表示方法

- 局部性表示
  - 独热表示
- 分布式表示
  - 横向组合关系
    - 隐性语义索引(Latent Semantic Indexing, LSI)
    - 概率隐性语义索引(Probabilistic Latent Semantic Indexing, PLSI)
    - 隐性狄利克雷分析(Latent Dirichlet Allocation, LDA)
  - 纵向聚合关系
    - 神经网络概率语言模型(Neural Prob. Language Model, NPLM)
    - 排序学习模型(C&W)
    - 上下文预测模型(Word2Vec)
    - 全局上下文模型(GloVe)

局部性表示 vs. 分布式表示

- 局部性表示(Local Representation)：在将单词表示为向量时，每个单词使用向量中独有且相邻的维度。在这种表示下，单词之间是相互独立的
- 分布式表示(Distributed Representation)：将单词映射到特征空间中，每个单词由刻画它的多个特征来高效表示；在形式上使用稠密实数向量（向量多于一个维度非0，通常为低维向量）来表示单词。分布式表示可以编码不同单词之间的语义关联

局部性表示



分布式表示

单词的分布式表示

- 分布式表示方法都基于分布语义假设(Distributional Hypothesis)，即单词的语义来自其上下文(context)。因此，
- 所有的分布式表示模型都利用某种上下文的统计信息来学习单词的分布式表示，使用不同的上下文使得模型建模了单词间的不同关系，可分为横向组合(Syntagmatic)关系和纵向聚合(Paradigmatic)关系

单词的分布式表示

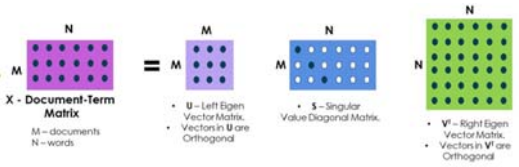
- 横向组合关系指两个单词同时出现在一段文本区域中，强调它们可以进行组合，在句子中往往起到不同的语法作用，如下图中“国科大”和“大学”即存在横向组合关系。对横向组合关系建模的模型通常使用文档作为上下文
- 纵向聚合关系指的是纵向的可替换的关系，如图中的“国科大”和“清华”。纵向聚合关系通常使用当前单词周边的单词作为其上下文

横向组合关系

- 常用的横向组合关系有:
  - 隐性语义索引(Latent Semantic Indexing, LSI)
  - 概率隐性语义索引(Probabilistic Latent Semantic Indexing, PLSI)
  - 隐性狄利克雷分析(Latent Dirichlet Allocation, LDA)
  - 均属于矩阵分解模型

隐性语义索引(LSI)

- LSI是指通过对词项-文档矩阵C (每行代表一个词项，每列代表一篇文档；元素C_{ij}表示第i个单词在第j篇文档中出现的次数)进行矩阵分解(具体采用SVD分解)来找到它的某个低秩逼近，进而利用得到的低秩逼近形成对词项和文档的新的表示
- 给定$M \times N$的词项-文档矩阵C和正整数k，对C进行LSI的过程如下：
  - 1. 将矩阵C分解为$C = U\Sigma V^T$；
  - 2. 保持Σ对角线上前k大个奇异值不变，其余元素置为0，得到Σ_k；
  - 3. 计算$C_k = U\Sigma_k V^T$作为C的低秩逼近，其中矩阵的行表示词项经过LSI后的向量，列表示文档经过LSI后的向量。一般而言，不同于非负整数构成的较为稀疏的矩阵C，C_k是一个实数构成的稠密矩阵



隐性语义索引(LSI)

- LSI仅保留了矩阵C中最大的k个奇异值，相当于将原有的词项-文档矩阵从r维降至k维，每一个奇异值可以理解为对应一个“主题”维度，其值的大小表示与这一“主题”的相关程度，因此LSI也称为主题模型(Topic Model)
- 保持较大的奇异值而将较小的奇异值置为0可以保留文档集中较为重要的信息，并且忽视不重要的细节，从而解决多词一义(synonymy)和语义关联的问题
- LSI得到的不是一个概率模型，缺乏统计基础，结果难以直观解释。此外，很难选择合适的k值，而k的选取对结果的影响非常大

概率隐性语义索引(PLSI)

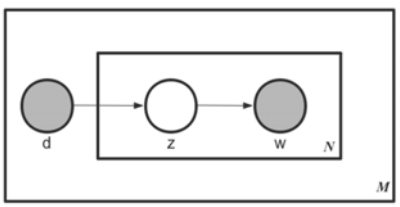
- 鉴于LSI的不可解释性，在其提出后许多学者尝试找到比较严谨的数学方法，一种途径是通过引入概率图模型得到概率化的解释

- Hofmann于1999年提出的PLSI(Probability Latent Semantic Indexing)文本模型便是在这一方向上取得的重要学术成果

- PLSI也是一种主题模型，不同于LSI中启发式的选择秩k进行低秩逼近的做法，PLSI定义了一个概率模型，并对模型中使用的变量及其对应的概率分布和条件概率分布给出了明确的解释

概率隐性语义索引(PLSI)

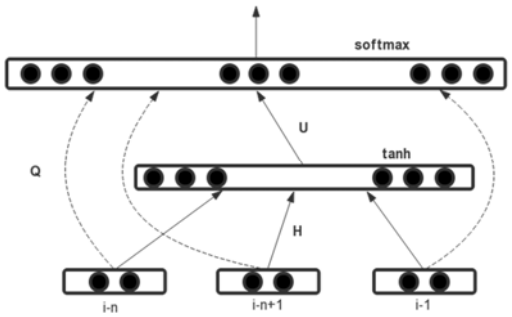
- PLSI假设在M篇文档和N个词项之间存在k个隐藏的主题，但我们无法对其进行观测。按概率$P(d_m)$选择一篇文档$d_m \in D$，$P(z_k|d_m)$表示在d_m下主题$z_k \in Z$的概率分布，$P(w_n|z_k)$表示在主题z_k下词项$w_n \in W$的概率分布。那么，我们可以将文档到词项的过程看做一个有向图，如下图所示



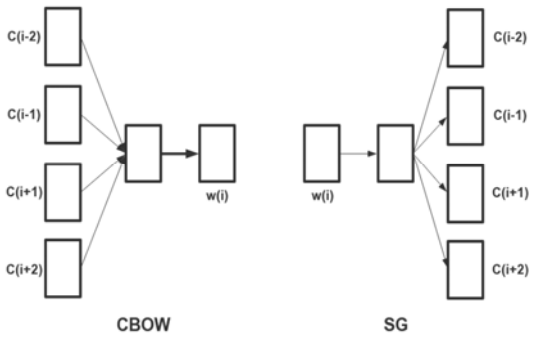
PLSI模型示意图

纵向聚合关系

- 常用的纵向聚合算法有
  - 神经网络概率语言模型(Neural Probabilistic Language Model, NPLM)
  - 排序学习模型(C&W)
  - 上下文预测模型(Word2Vec)
  - 全局上下文模型(GloVe)等



NPLM模型



CBOW和SG模型框架

排序学习模型(C&W)

- 排序学习模型(C&W) 由Collobert & Weston于2008年提出
- 相比NPLM模型，主要有两点改进：
  - 1. C&W同时使用了单词的上下文，这成为了之后学习单词表示的基本做法；
  - 2. C&W对单词序列打分使用了排序损失函数，而不是基于概率的极大似然估计，其损失函数为

$$\max(0,1-s(w,c)+s(w',c))$$

其中，c 表示单词w 的上下文(context)，w'表示将当前上下文中的单词w替换为一个随机采样出的无关单词w'（负样例）；s为打分函数，分数越高表明该段文本越合理

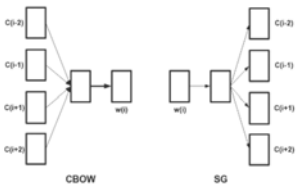
- 显然，在大多数情况下将普通短语中的特定单词随机替换为其他单词时，得到的都是不正确的短语。因此模型的目标是，尽量使正确短语的得分 比 随机替换后的短语的得分高于1

上下文预测模型(Word2Vec)

- 为了更好的利用单词的上下文，Mikolov等人提出了两个简单的神经网络模型CBOW(Continuous Bag of Word)和SG(Skip Gram)进行学习。
- 相比NPLM模型，CBOW模型去除了中间的非线性隐层，将单词w_i上下文的表示经过求和或平均等计算后，用得到的结果h_i直接预测单词w_i；而SG模型则使用单词w_i预测其上下文中的每一个单词。其目标函数分别为：

$$P(w_i|c) = \frac{exp(w_i \cdot h_i)}{\sum_{w'} exp(w' \cdot h_i)} \text{ (CBOW)}$$

$$P(w_i|c) = \prod_{c_j} \frac{exp(w_i \cdot c_j)}{\sum_{w'} exp(w' \cdot c_j)} \text{ (SG)}$$



- 以短语 “国科大 是 位于 北京 的 大学”为例：
  - CBOW基于上下文“国科大 是 位于 的 大学”预测中心词“北京”
  - SG基于中心词“北京”预测上下文“国科大 是 位于 的 大学”

全局上下文模型(GloVe)

- 在Word2Vec算法中，主要使用单词的上下文信息获得单词的表示。GloVe算法额外利用单词的共现信息，将全文的统计信息与句子的信息相结合，以期得到单词在语义和语句上更好的表达
- 令单词w_i出现的次数为X_i，单词w_i与w_k同时出现的次数为X_{ik}，则在单词w_i出现的情况下单词w_k出现的条件概率为$P(w_k|w_i) = \frac{X_{ik}}{X_i}$
- 研究发现，条件概率的比值$ratio_{i,j,k} = \frac{P(w_k|w_i)}{P(w_k|w_j)}$存在如下规律：

$ratio_{i,j,k}$	单词 j, k 相关	单词 j, k 不相关
单词 i, k 相关	趋于1	很大
单词 i, k 不相关	很小	趋于1

全局上下文模型(GloVe)

- 基于这一观察，对每个单词对定义如下的软约束

$$v_i^T v_j + b_i + b_j = log\, X_{ij}$$

其中，v_i和v_j是单词w_i和w_j的向量，b_i和b_j为w_i和w_j的偏差， X_{ij}为权重项，正比于单词w_i和w_j共现的次数

- 进一步，定义如下的目标函数

$$J = \sum_{i,j=1}^{N} f(X_{ij})(v_i^T v_j + b_i + b_j - log\, X_{ij})^2$$

其中，

$$f(X_{ij}) = \begin{cases} \left(\frac{X_{ij}}{XMAX}\right)^{\alpha} & if X_{ij} < XMAX \\ 1 & \text{Otherwise} \end{cases}$$

$f()$防止只从共现率很高的单词对中学习

句子的表示方法

- 传统方法
  - 词集模型
  - 词袋模型
  - TF-IDF表示
- 分布式表示方法
  - 主题模型
  - 基于单词分布式表示组合的表示方法
  - 由原始语料直接学习的表示方法

词袋模型

- 词袋模型(Bag of Words)是在词集模型的基础上，考虑了单词出现的次数，因此，在词袋模型中，句子向量中每个单词对应的位置上记录的是该单词出现的次数，这也体现了各个单词在该句子中的重要程度

- 示例：
  - 句子：“我 来自 中国 科学院 大学，他 在 中国 科学院 计算所 学习”
  - 单词表vocab={我:0, 来自:1, 中国:2, 科学院:3, 大学:4, 他:5, 在:6, 计算所:7, 学习:8}
  - 词袋模型向量表示：(1, 1, 2, 2, 1, 1, 1, 1, 1)

TF-IDF模型

- TF-IDF (Term Frequency - Inverse Document Frequency)是一种用于信息检索与数据挖掘的常用加权技术。TF是词频(Term Frequency)，IDF是逆向文档频率(Inverse Document Frequency)

- TF-IDF的主要思想是：如果某个词或短语在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为该词或者短语具有很好的类别区分能力，适合用来分类

TF-IDF模型

- TF计算公式如下：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

其中，$n_{i,j}$是单词t_i在文档d_j中的出现次数，分母是文档d_j中所有单词的出现次数之和。

- IDF是对一个词语普遍重要性的度量。某一特定词语的IDF，可由总文档数目除以包含该词语之文档的数目，再将得到的商取对数得到。具体如下：

$$idf_i = log\frac{|D|}{|\{j: t_i \in d_j\}|}$$

其中|D|表示语料库中的文档总数

- TF-IDF：

$$\text{tfidf}_{i,j} = \text{tf}_{i,j} \times \text{idf}_i$$

可以看出，某一特定文档内的高词语频率，以及该词语在整个文档集合中的低文档频率，可以产生出高权重的TF-IDF。因此，TF-IDF倾向于过滤掉常见的词语，保留重要的词语

文本中的匹配问题

文本匹配是自然语言理解的一个核心问题，许多文本处理的问题可以抽象成文本匹配的问题

信息检索	查询项 ⟷	文档
问答系统	问题 ⟷	答案
对话问题	前文 ⟷	回复
复述问题	原句 ⟷	改写
机器翻译	中文 ⟷	英文

搜索引擎

问答系统

智能助手

基于启发式规则的文本匹配

- 启发式规则的文本匹配模型直接建模了两段文本共同出现的词的分布。不难理解，在两段文本中同时出现的词，对于度量文本的匹配程度是至关重要的

- 两个经典模型
  - BM25
  - 查询似然模型（Query Likelihood Model）

基于隐语义表达的文本匹配

- 除了两段文本中共同出现的词对计算匹配度有贡献，那些词与词之间的关系（近义词，包含关系等）也应该考虑进匹配度的计算，因此提出了隐语义表达的文本匹配模型

- 这类方法将一段文本映射到一个向量（离散稀疏向量或者连续稠密向量），然后通过计算向量的相似度来表示文本的匹配度。文档表示的各类方法（如TF-IDF和BM25）可以参考上一节的相关内容

基于学习的文本匹配

- 在大数据的背景下，文本匹配除了考虑计算效率外，也要考虑结果的准确性。得益于大量有标注的文本匹配数据，基于学习的文本匹配模型可以通过有监督的方式，得到更为准确的结果

- 基于学习的文本匹配模型分为两类
  - 基于人工特征的排序学习模型
  - 基于表达学习的排序学习模型

基于人工特征的排序学习模型

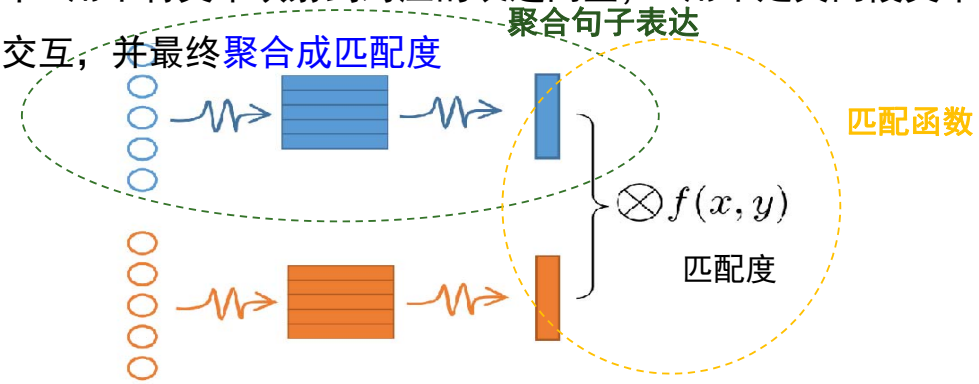
- 人工特征
  - 人工特征是根据实际任务中对数据的理解，设计出的用以抽象数据的特征表示。BM25与TF-IDF都可以看作是一种人工设计的特征
  - 在文本匹配任务中，人工提取的特征可以分为两大类：基于文本内容的特征和基于文本交互的特征。通常情况下，人工提取的特征会拼接成一个特征向量，用来表示相应的文本
  - 基于文本内容的特征
    - 主要包括关键词、文本类型、文本长度等等
    - 还包括当前文本在与其他文本构造的关系图上的PageRank重要度特征，例如通过域名、链接等构造的文档关系图
  - 基于文本间交互的特征
    - 包括关键词匹配的数量、BM25、查询似然模型得到的匹配度得分等
    - 也包括一些精心设计的关于邻近度的特征

基于表达学习的排序学习模型

- 目前已有大量基于表达学习的排序学习模型被提出，可以直接以文本的内容作为输入，以匹配度作为输出，端到端的学习模型。文本匹配任务可以抽象成如下的形式：

$$匹配度 = \mathcal{F}\big(\Phi(S_1), \Phi(S_2)\big).$$

- 其中Φ用来将文本映射到对应的表达向量，$\mathcal{F}$用来定义两段文本表达的交互，并最终聚合成匹配度



文本匹配的评价方法

- 分类准确率（Accuracy）：用于评价分类任务的指标，对于文本匹配任务，只有两类标签，匹配为1，不匹配为0。因此可以把文本匹配看作是一个二分类问题。使用分类准确率可以方便的评价模型对每一对文本的分类是否正确。分类正确的数量占总测试样本数量的比例就是分类准确率

- P@k（Precision at k）：表示前k个文档的排序准确率。假定预测结果排序后，前k个文档中相关文档的数量为Y_k，那么P@k可以定义为：

$$\mathrm{P@k} = \frac{Y_k}{k}$$

- R@k（Recall at k）：表示前k个文档的排序召回率。按照标注的相关度排序后，前k个文档中相关文档的数量为G_k，那么可定义R@k为：

$$\mathrm{R@k} = \frac{G_k}{k}$$

文本匹配的评价方法

- MAP（Mean Average Precision）：该指标综合考虑了所有相关文档的排序状况。将所有相关文档在预测结果排序中的位置定义为$r_1, r_2, \cdots, r_G$，则平均精度均值指标可定义为：

$$\mathrm{MAP} = \frac{\sum_{i=1}^{G} P@r_i}{G}$$

- MRR（Mean Reciprocal Rank）：如果只考虑预测结果排序中第一个出现的相关文档的位置r_1，可以定义MRR指标为：

$$\mathrm{MRR} = \mathrm{P@}r_1 = \frac{1}{r_1}$$

文本匹配的评价方法

- nDCG（normalized Discounted Cumulative Gain）归一化折扣累计收益
  - 有些任务当中标注的相关度本身就有大小之分而不是单纯的匹配和不匹配两个级别，这个时候nDCG这个指标就会更加有效。nDCG让相关度越高的排在越前面
  - 给定按照标注的文档相关度排序后的文档相关度值分别为$\widehat{rel}_1, \widehat{rel}_2, \cdots, \widehat{rel}_N$，若按照预测结果排序后的文档相关度的值分别为$rel_1, rel_2, \cdots, rel_N$。所以nDCG指标的定义如下：

$$IDCG = \hat{rel}_1 + \sum_{i=2}^{n} \frac{\hat{rel}_i}{\log_2 i}$$

$$DCG = rel_1 + \sum_{i=2}^{n} \frac{rel_i}{\log_2 i}$$

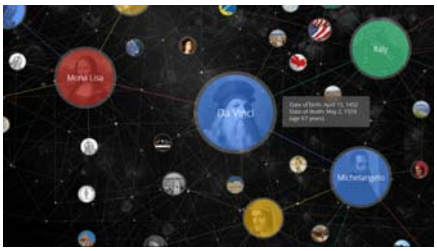
$$nDCG = \frac{DCG}{IDCG}$$

Outline

- 大数据与大数据分析简介
- 大数据分析技术与系统
- 大数据统计分析
- 大数据机器学习
- 数据驱动的自然语言处理
- 文本大数据分析
- **知识图谱与知识计算**
- 大图挖掘与分析
- 社会媒体分析

知识图谱 (Knowledge Graph)

- 知识图谱本质上是一种语义网络(Semantic Net)，其结点代表实体(entity)或概念(concept)，边代表实体/概念之间的各种语义关系；

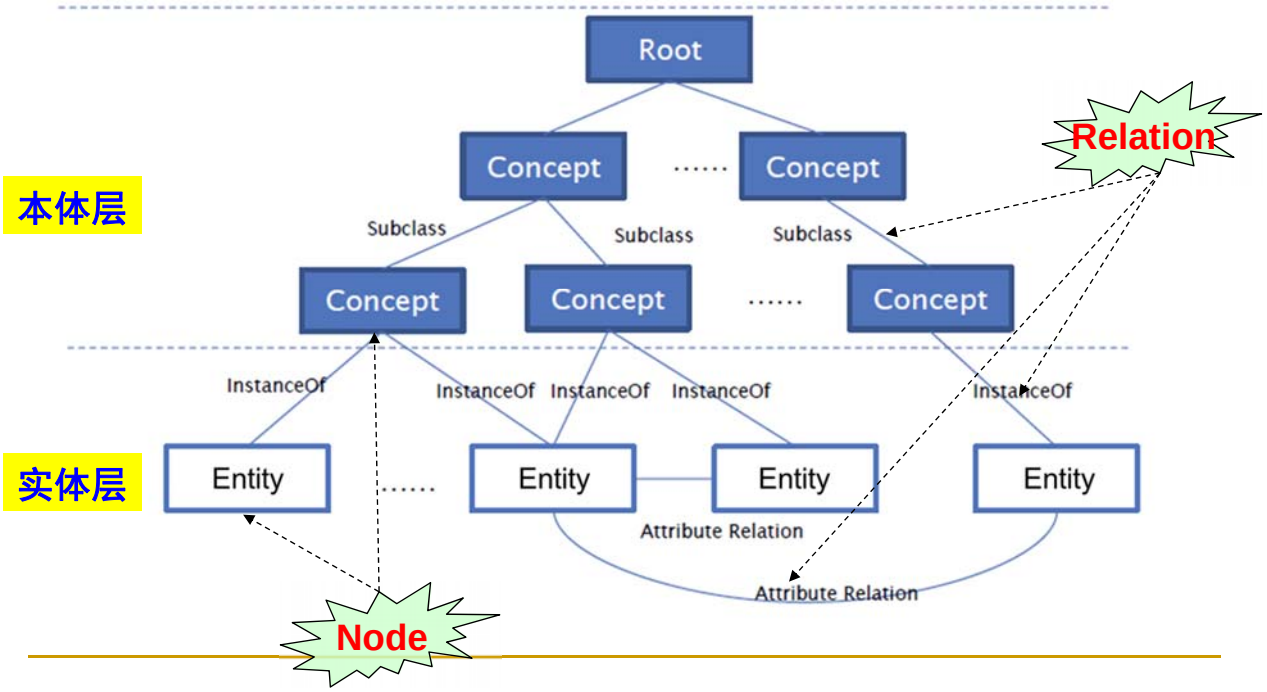


Google, 2012

- A Knowledge Graph (KG) is a system that understands facts about people, places and things and how these entities are all connected;

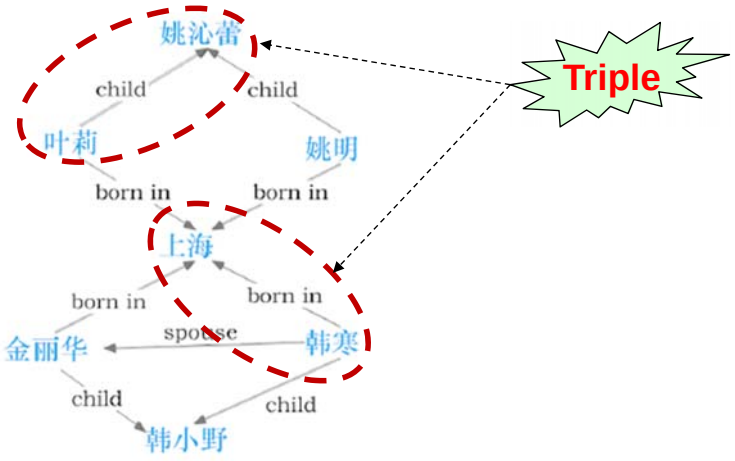
- 知识图谱把不同来源、不同类型的信息连接在一起形成关系网络，提供了从关系的角度去分析问题的能力

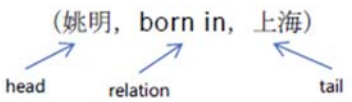
什么是知识图谱？



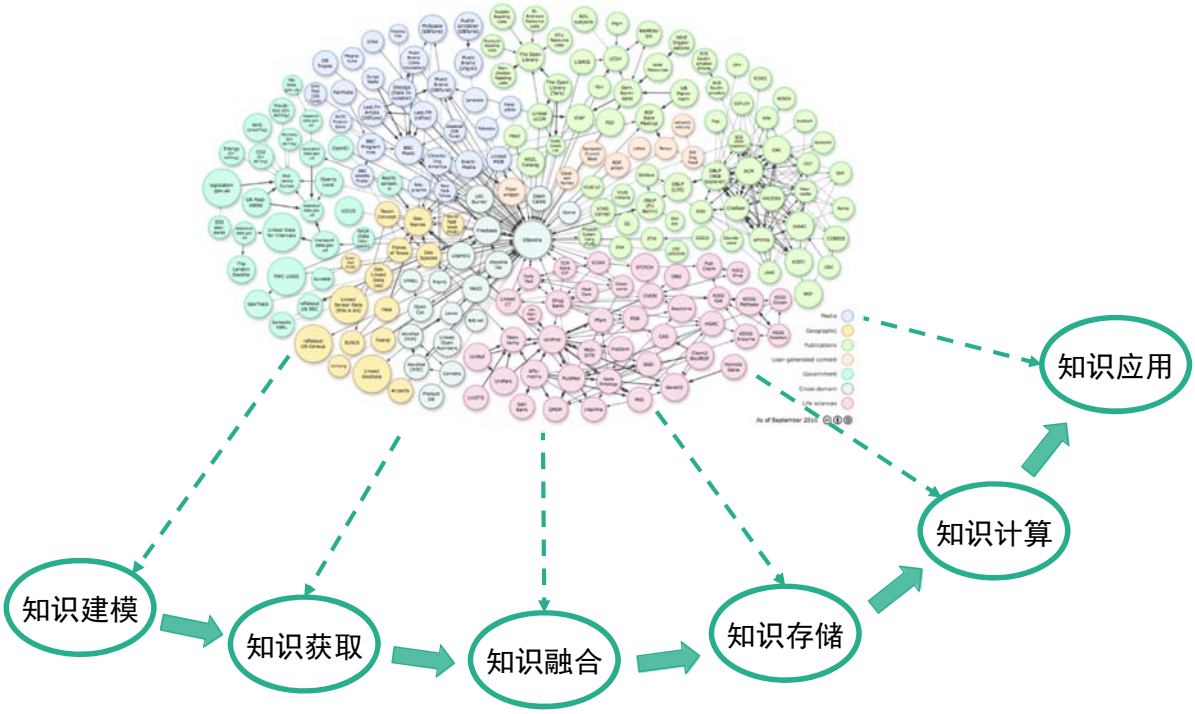
知识图谱中的知识表示：三元组

- 三元组Triple: (head, relation, tail)
  - head：头实体/概念
  - relation：关系/属性
  - tail：尾实体/概念





知识图谱的生命周期



实体抽取

- 实体抽取定义
  - 从原始语料中自动识别出指定类型的命名实体，主要包括实体名（如人名、地名、机构名、国家名等）、缩略词，以及一些数学表达式（如货币值、百分数、时间表达式等）

- 示例

  5月19日下午，史密斯教授做客北京大学海外名师讲堂。

  时间　　人名　　机构名

基于机器学习的方法

- 序列标注
  - 实体标注一般使用BIO模式

(B-begin，I-inside，O-outside)

输入序列	小明	昨天	晚上	在	公园	遇到	了	小红	。
语块	B-NP	B-NP	I-NP	B-PP	B-NP	B-VP		B-NP	
标注序列	B-Agent	B-Time	I-Time	O	B-Location	B-Predicate	O	B-Patient	O
角色	Agent	Time	Time		Location	Predicate	O	Patient	

  - 还有BIOES标注模式

(B-begin，I-inside，O-outside，E-end，S-single)

基于机器学习的方法

- 隐马尔科夫模型
  - 假定分词后的文档词语序列为$W = (w_1, \dots, w_n)$，$T = (t_1, \dots, t_n)$为词序列的实体标注结果。模型旨在给定词语序列W的情况下，找出概率最大的标注序列T，即，求使$P(T|W)$最大的标注序列

$$T_{max} = arg_T maxP(T|W)$$

根据贝叶斯公式，

$$P(T|W) = P(T)P(W|T)/P(W)$$

其中，$P(W)$可以看成一个常数，则有

$$T_{max} = arg_T maxP(T)P(W|T)$$

其中，$P(T)P(W|T)$是引入隐马尔科夫模型来计算的参数。如果穷举序列W和T的所有可能情况，这个问题是NP难的

基于机器学习的方法

- 隐马尔科夫模型
  - 按照马尔科夫假设，当前状态t_i只和其前一状态t_{i-1}有关，因此有

$$P(T)P(W|T) \approx \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1})$$

其中，$P(w_i|t_i)$表示隐状态为t_i的词语集合中出现w_i的概率，$P(t_i|t_{i-1})$表示上一词语标注为t_{i-1}时，当前词语标注为t_i的转移概率。进一步

$$T_{max} = arg_T max \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1})$$

$$T_{max} = -arg_T min \sum_{i=0}^{n} \{lnP(w_i|t_i) + lnP(t_i|t_{i-1})\}$$

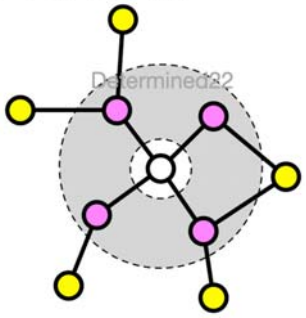
训练时，取$P(w_i|t_i) \approx Count(w_i, t_i)/Count(t_i)$，其中$Count(w_i, t_i)$表示词语$w_i$被标注为$t_i$的次数，$Count(t_i)$表示隐状态$t_i$出现的总次数

基于机器学习的方法

- 条件随机场(Conditional Random Field, CRF)
  - 设$X = (X_1, \dots, X_n)$与$Y = (X_1, \dots, X_n)$是联合随机变量。若在给定随机变量X的条件下，随机变量Y构成一个由无向图$G = (V, E)$表示的马尔科夫模型，则条件概率分布$P(Y|X)$称为条件随机场，即：

$$P(Y_v|X, Y_w, w/= v) = P(Y_v|X, Y_w, w \sim v)$$

其中，$w/= v$表示图$G = (V, E)$中节点v以外的所有节点，$w \sim v$表示与节点v有连边的所有节点



$\{w \sim v\}$
v
$\{w/= v\}$-$\{w \sim v\}$

基于机器学习的方法

- 线性链条件随机场模型
  - 与隐马尔科夫模型相同，将CRF用于命名实体识别，其目标也是求$T_{max} = arg_T max P(T|W)$，但是这里

$$P(T|W) = \frac{1}{Z(W)} exp\left(\sum_{i,k} \lambda_k \psi_k(t_{i-1}, t_i, W, i) + \sum_{i,l} \delta_l \phi_l(t_i, W, i)\right)$$

$$Z(W) = \sum_t exp\left(\sum_{i,k} \lambda_k \psi_k(t_{i-1}, t_i, W, i) + \sum_{i,l} \delta_l \phi_l(t_i, W, i)\right)$$

其中$Z(W)$是归一化因子，在所有可能的输出序列上求和；λ_k和δ_l为权重因子

基于机器学习的方法

- 线性链条件随机场模型

$$P(T|W) = \frac{1}{Z(W)} exp\left(\sum_{i,k} \lambda_k \psi_k(t_{i-1}, t_i, W, i) + \sum_{i,l} \delta_l \phi_l(t_i, W, i)\right)$$

  - $\psi_k(t_{i-1}, t_i, W, i)$是转移函数，依赖于当前和前一位置，表示从标注序列中位置$i-1$的标记t_{i-1}转移到位置i上的标记为t_i的概率

$$\varphi_k(t_{i-1}, t_i, W, i) = \begin{cases} 1, & \text{满足条件} \\ 0, & \text{其他} \end{cases}$$

  - $\phi_l(t_i, W, i)$是状态函数，表示标记序列在位置i上标记为t_i的概率

$$\phi_l(t_i, W, i) == \begin{cases} 1, & \text{满足条件} \\ 0, & \text{其他} \end{cases}$$

关系抽取

- 关系抽取示例

- 抽取方法分类
  - 有监督关系抽取
  - 半监督关系抽取
  - 远程监督关系抽取
  - 无监督关系抽取

远程监督关系抽取

- 初始动机：通过外部知识库代替人对语料进行标注，从而低成本地获取大量有标注数据 [Mintz et al., 2009]

- 核心思想：如果知识库中存在三元组$\langle e_1, R, e_2 \rangle$，那么语料中所有出现实体对$\langle e_1, e_2 \rangle$的语句，都标注为表达了关系R

- 根据这一假设，对每个三元组$\langle e_1, R, e_2 \rangle$，将所有$\langle e_1, e_2 \rangle$共现的句子都标注标签$R$，用分类方法解决关系抽取问题

远程监督关系抽取

- Riedel等[Riedel et al., 2010]认为Mintz的假设过强，可能引入噪声模式，因而提出“at-least-once”假设：
  - 如果存在三元组$\langle e_1, R, e_2 \rangle$，那么所有$\langle e_1, e_2 \rangle$实体对共现的语句中，至少有一句体现了关系R在这两个实体上成立的事实

- 引入了多实例学习机制，将所有$\langle e_1, e_2 \rangle$共现的句子聚成一个句袋，并将任务由对句子分类变为对句袋分类

实体对齐

- 实体对齐的定义
  - 实体对齐也称为实体匹配或实体解析，指的是判断相同或不同的知识库中的两个实体是否指向同一个对象的过程
    - 如：“诗仙”和“李太白” 两个指称词都应指向同一个实体“李白”

- 实体对齐方法分类
  - 成对实体对齐
  - 集体实体对齐

实体链接

- 实体链接的定义
  - 实体链接指的是利用知识库中的实体对知识抽取阶段所获得的实体指称词进行消歧的过程
- NIL实体
  - 如果实体指称在知识库中找不到对应的实体，则称其为“NIL实体”，实体链接还需要对NIL实体进行预测
- 实体链接为每一个实体指称词在知识库中找到对应的映射或者给出NIL实体的标签

什么是知识推理？

- 人类视角
  - 人们从已知的事实出发，通过运用已掌握的知识，找出其中蕴含的事实或归纳出新的事实的过程
  - 按照某种策略由已知判断推出新的判断的思维过程
  - 基于特定的规则和约束，从存在的知识获得新的知识
- 计算机视角
  - 在计算机或智能系统中，模拟人类的智能推理方式，依据一定的推理控制策略，利用形式化的知识进行机器思维和求解问题的过程

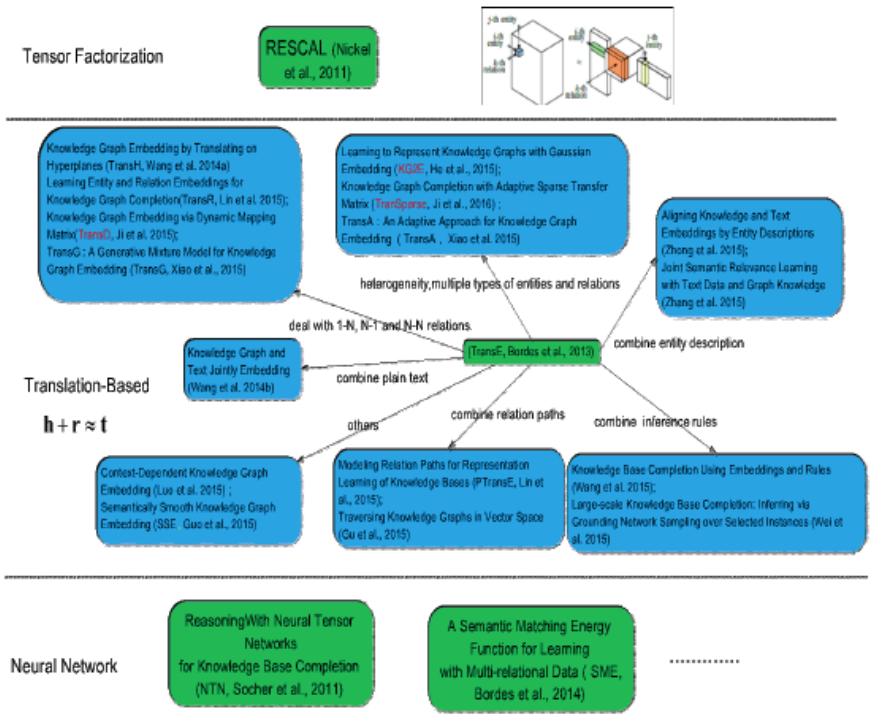
利用已知的知识推出新知识的过程

基于分布式表达的知识计算

张量分解方法

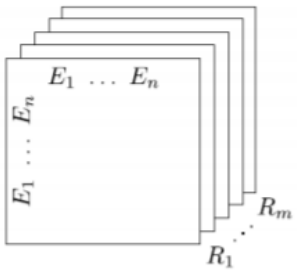
基于翻译的方法

神经网络方法



用张量表示知识图谱

知识图谱中三元组的结构是（头部实体h，关系r，尾部实体t），其中r连接头尾实体。以E_1，E_2，… E_n 表示知识图谱中的实体，以R_1，R_2，… R_N 表示知识图谱中的关系，则可以使用一个三维矩阵（张量）表示知识图谱



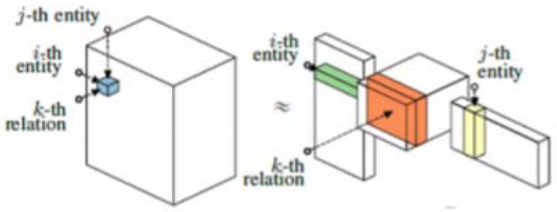
Nickel et al. (2011). A three-way model for collective learning on multi-relational data. In Proceedings of the 28th international conference on machine learning (ICML-11).

张量分解的目标函数

- 表示知识图谱的张量记为$\mathcal{Y}$，其第k个矩阵记为Y_k，是第k种关系的矩阵，表示该种关系在向量空间中与头尾部实体相互作用
- 对Y_k可以进行如下的低秩分解：

$$\mathrm{Y_k} = \mathrm{AR_kA^T} \qquad k = 1,2,\cdots m$$

其中，$Y_k \in R^{n*n}$，$A \in R^{n*r}$，$R_k \in R^{r*r}$，r表示矩阵A的秩；A是实体向量矩阵，每一行表示一个实体的向量，转置后其每一列表示一个实体的向量



张量分解的目标函数

- 由上述内容可知，A 和R_k 均是待求解的变量。因此目标函数是：

$$\min_{A,R_k}(f(\boldsymbol{A},\boldsymbol{R}_k)+g(\boldsymbol{A},\boldsymbol{R}_k))$$

其中 $f(\boldsymbol{A},\boldsymbol{R}_k)$ 是目标函数

$$f(\boldsymbol{A},\boldsymbol{R}_k)=\frac{1}{2}\left(\sum_k\left\|\boldsymbol{Y}_k-\boldsymbol{A}\boldsymbol{R}_k\boldsymbol{A}^{\mathrm{T}}\right\|_F^2\right)$$

$g(\boldsymbol{A},\boldsymbol{R}_k)$ 是正则化项

$$g(\boldsymbol{A},\boldsymbol{R}_k)=\frac{1}{2}\gamma\left(\|\boldsymbol{A}\|_F^2+\sum_k\|\boldsymbol{R}_k\|_F^2\right)$$

张量分解的目标函数

- 将目标函数写成分量形式

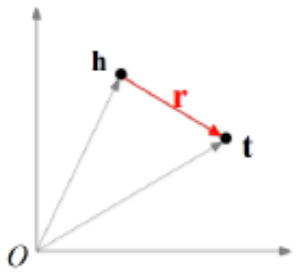
$$f(\boldsymbol{A},\boldsymbol{R}_k)=\frac{1}{2}\left(\sum_k\left\|\boldsymbol{Y}_k-\boldsymbol{A}\boldsymbol{R}_k\boldsymbol{A}^{\mathrm{T}}\right\|_F^2\right)\Rightarrow f(\boldsymbol{A},\boldsymbol{R}_k)=\frac{1}{2}\sum_{i,j,k}\left(y_{ijk}-\boldsymbol{a}_i^T\boldsymbol{R}_k\boldsymbol{a}_j\right)^2$$

其中，y_{ijk}是张量中的一个元素，$\boldsymbol{a}_i$表示$\boldsymbol{A}$的第i行，即

$$[\boldsymbol{a}_1,\boldsymbol{a}_2,\ldots,\boldsymbol{a}_n]=\boldsymbol{A}$$

基于分布式表达的推理：TransE

- 关系事实=(head, relation, tail)，其对应的向量表示为$(\mathbf{h},\mathbf{r},\mathbf{t})$
- 基本思想：把关系看作是头尾实体之间的平移(翻译)操作



$(\mathbf{h},\mathbf{r},\mathbf{t})$

$\mathbf{h}+\mathbf{r}=\mathbf{t}$

向量加法的三角形法则：

中国 + 首都 = 北京

法国 + 首都 = 巴黎

俄罗斯 + 首都 = 莫斯科

Bordes, et al. Translating embeddings for modeling multi-relational data. In Advances in Neural Information Processing Systems, 2013 (pp. 2787-2795).

基于分布式表达的推理：TransE

- 势能函数
  - 对于真实事实的三元组(h, r, t)，要求$\mathbf{h}+\mathbf{r}=\mathbf{t}$；而对于错误的三元组则不满足该条件

$$f(h,r,t)=\|\mathbf{h}+\mathbf{r}-\mathbf{t}\|_2$$

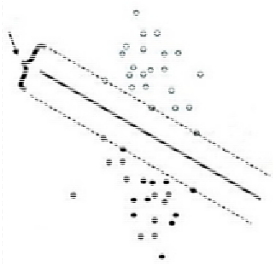
$$f(\text{姚明},\text{出生于},\text{北京})>f(\text{姚明},\text{出生于},\text{上海})$$

基于分布式表达的推理：TransE

- 损失函数：

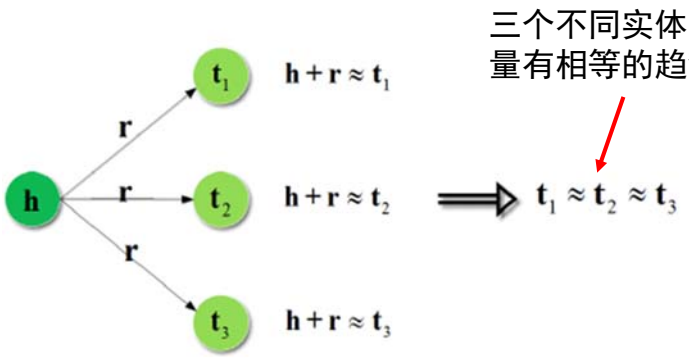
$$L=\sum_{(h,r,t)\in\triangle}\sum_{(h',r,t')\in\triangle'}\max\left(0,f_r(h,t)+M_{opt}-f_r(h',t')\right)$$

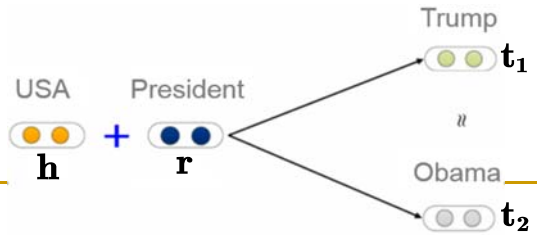
正例三元组集　　负例三元组集　　最优Margin超参



关系多样性问题

- 知识图谱中关系有1-1、1-N、N-1、N-M多种类型

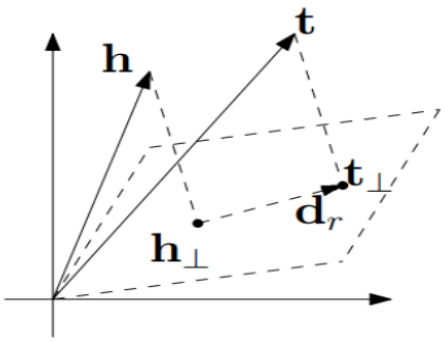




Trans系列

- TransH

$$f_r(h,t) = -\|\mathbf{h}_\perp + \mathbf{r} - \mathbf{t}_\perp\|_{L_1/L_2}$$

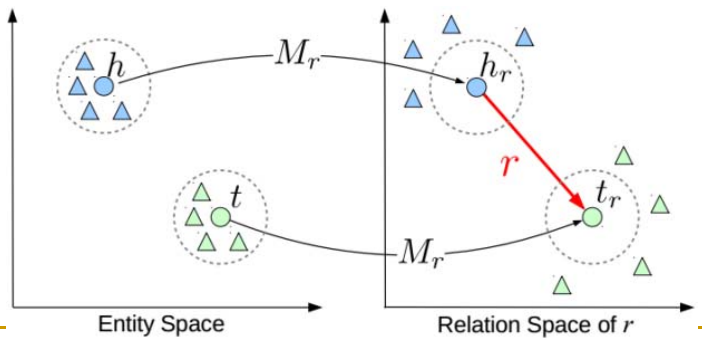


TransH

- TransR

$$f_r(h,t) = -\|\mathbf{h}_\mathbf{r} + \mathbf{r} - \mathbf{t}_\mathbf{r}\|_{L_1/L_2}$$

- ...



TransR

谢谢聆听！

预祝大家考试取得好成绩！