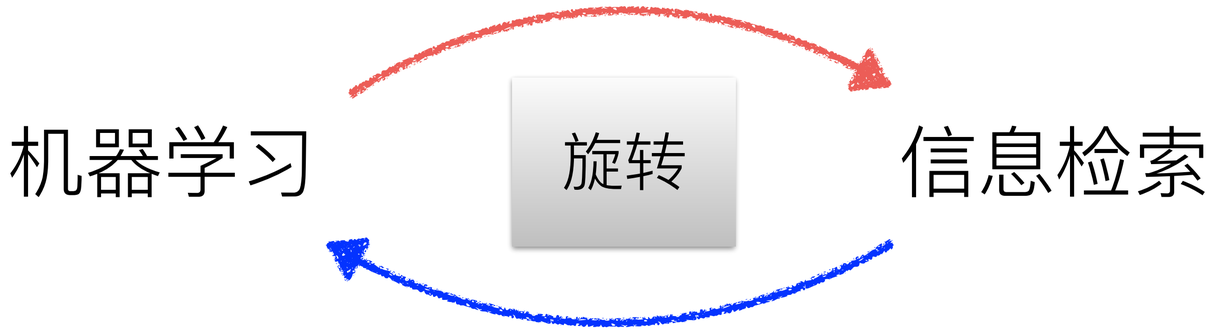
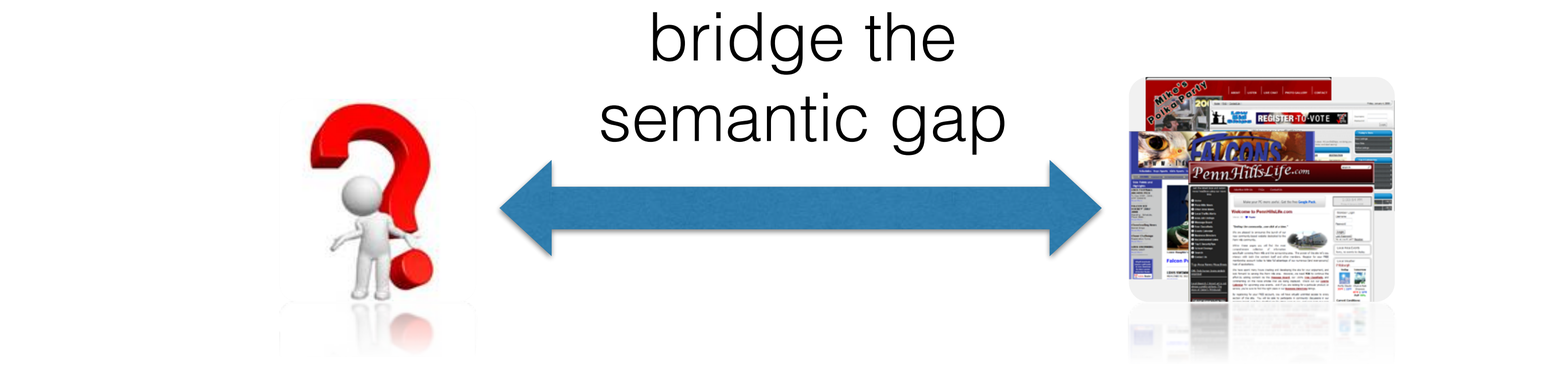
信息检索与机器学习的华尔兹

兰艳艳
副研究员
中国科学院计算技术研究所
lanyanyan@ict.ac.cn
www.bigdatalab.ac.cn/~lanyanyan

华尔兹



信息检索的高潮：现代搜索引擎



华尔兹的开场

华尔兹开场：排序学习

- 向量空间模型
  - TF-IDF
- 链接分析
  - PageRank
- Language Model
- Topic Modeling

检索的领域知识如何共同作用决定一个文档与查询的相关?

对信息检索的影响：领域知识作为数据表示，采用数据驱动的方式来自动决定如何共同作用

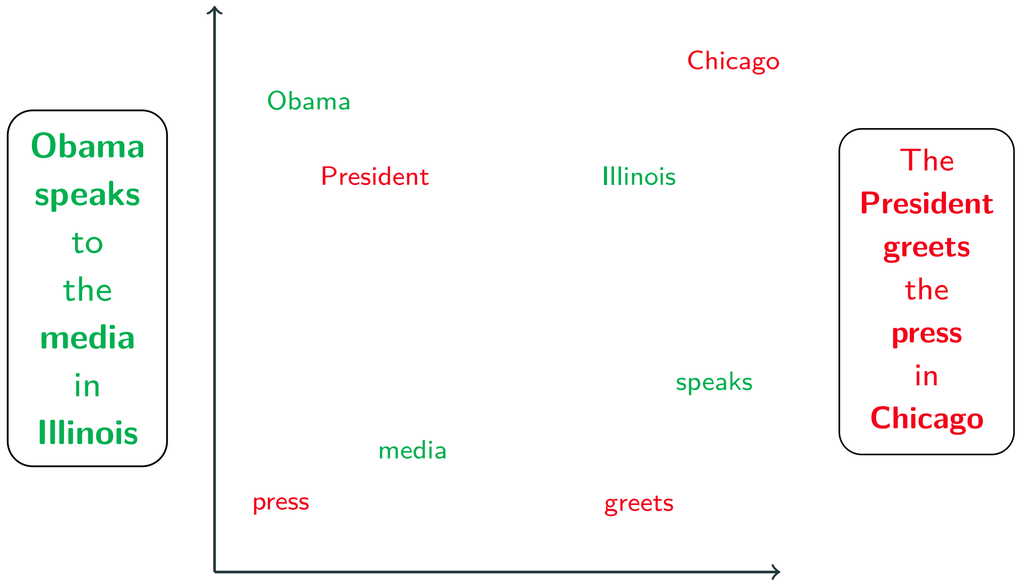
排序学习

对机器学习的影响：排序是一个并列于回归和分类的另一个问题，机器学习理论和算法都得到新的发展

华尔兹的中场

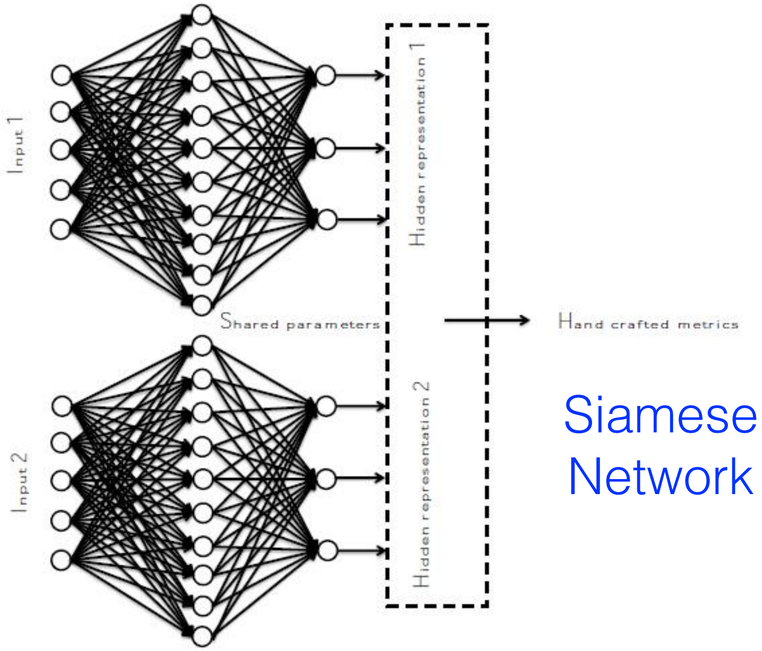
华尔兹的中场：深度学习的交融

- 文本表达技术
  - 从one-hot到semantic word embedding
  - 从数据底层表达开始刻画语义关系

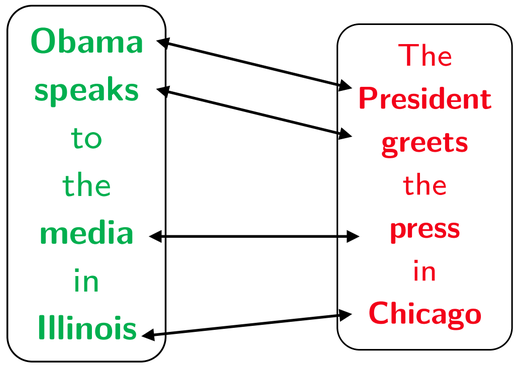


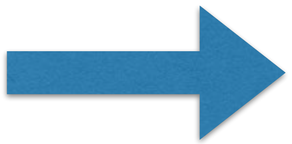
华尔兹的中场：深度学习的交融

- 深度文本匹配技术
  - 使用高度非线性函数表达复杂的文本匹配模式
  - 刻画抽象的语义关联规则

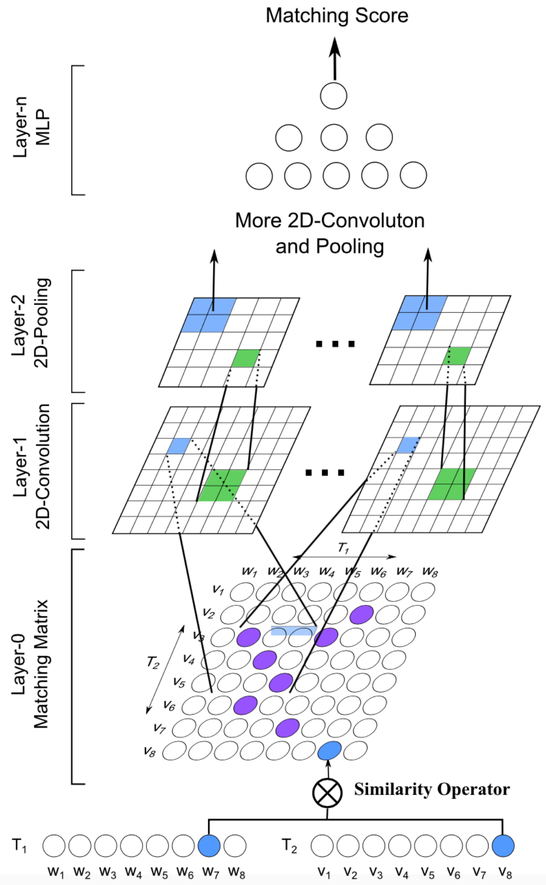


Siamese Network





Interaction Network



华尔兹的终场

信息检索的未来

Search

Information Need

Relevant Documents

NLU

Digestion
Analysis

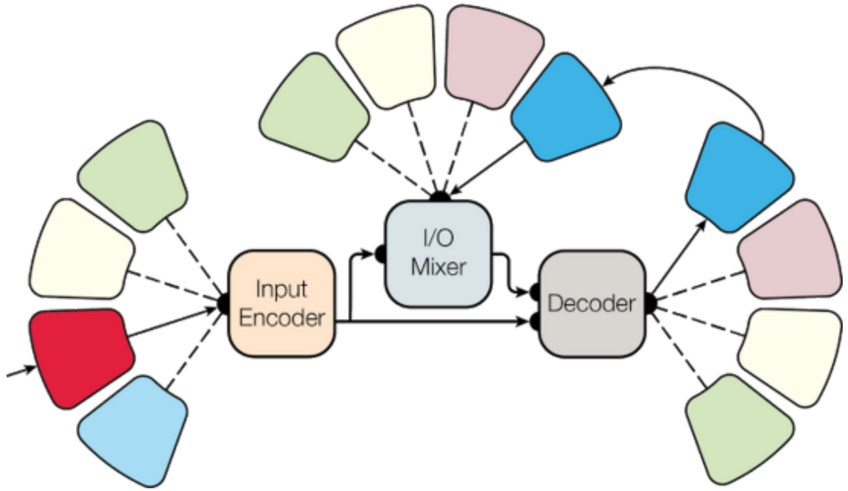
Response

Human-Computer
Interaction

智能搜索，智能检索，智能问答，智能对话

华尔兹的终场：局限与挑战

- 大数据vs小数据
  - zero shot learning；迁移学习；单任务到多任务；异构数据的融合
- 数据的表达问题：连续or离散
- 学习范式问题
  - 平均或极大似然的缺陷
  - Worst Case为目标
  - 知识与统计的融合



We demonstrate that MultiModel is capable of learning eight different tasks simultaneously: it can detect objects in images, provide captions, recognize speech, translate between four pairs of languages, and do grammatical constituency parsing at the same time. The input is given to the model together with a very simple signal that determines which output we are requesting.

Thanks!