# 数据科学导论
## Introduction to Data Science

# Task1: 文献调研报告

陈恩红，黄振亚

Email: cheneh@ustc.edu.cn, huangzhy@ustc.edu.cn

**课程主页：**
http://staff.ustc.edu.cn/~huangzhy/Course/DS2024.html

**助教：肖桐**
**ds_intro2024@163.com**

9/10/2024

# 课程要求与考核方式

- 课程目标：用科学的方法研究和应用数据
- 课程要求
  - 文献调研报告 1份
    - <span style="color:red">独立完成，每人一份</span>
    - <span style="color:blue">提交节点：第10周教学周（2024年11月05日）</span>
    - 用给的**LaTex**模板写报告，篇幅至少**4页**

# 任务一：文献调研

□ 结合上课的内容，调研相关的文献并撰写报告
   □ 文献可以从**模型、应用或大数据平台等**任一**主题**进行调研，调研的内容尽可能广，并将学到的内容和心得以报告形式提交
   □ **建议：大数据（数据分析）＋ 本人学科**
   □ 语言要求：中英文不限（包括调研的内容）
   □ 格式要求：专业、美观
□ 报告内容：
   □ 题目
   □ 调研结果综述
   □ 学习心得与思考
   □ 参考文献

# 任务一：文献调研

□ 什么是文献调研

1

# The Gap of Semantic Parsing: A Survey on Automatic Math Word Problem Solvers

Dongxiang Zhang, Lei Wang, Luming Zhang, Bing Tian Dai and Heng Tao Shen

**Abstract**—Solving mathematical word problems (MWPs) automatically is challenging, primarily due to the semantic gap between human-readable words and machine-understandable logics. Despite the long history dated back to the 1960s, MWPs have regained intensive attention in the past few years with the advancement of Artificial Intelligence (AI). Solving MWPs successfully is considered as a milestone towards general AI. Many systems have claimed promising results in self-crafted and small-scale datasets. However, when applied on large and diverse datasets, none of the proposed methods in the literature achieves high precision, revealing that current MWP solvers still have much room for improvement. This motivated us to present a comprehensive survey to deliver a clear and complete picture of automatic math problem solvers. In this survey, we emphasize on algebraic word problems, summarize their extracted features and proposed techniques to bridge the semantic gap, and compare their performance in the publicly accessible datasets. We also cover automatic solvers for other types of math problems such as geometric problems that require the understanding of diagrams. Finally, we identify several emerging research directions for the readers with interests in MWPs.

**Index Terms**—math word problem, semantic parser, reasoning, survey, natural language processing, machine learning

◆

# 任务一：文献调研

- 什么是文献调研

# A Survey of Knowledge Tracing

Qi Liu, Shuanghong Shen, Zhenya Huang, Enhong Chen, *Senior Member, IEEE* and Yonghe Zheng

**Abstract**—High-quality education is one of the keys to achieving a more sustainable world. In contrast to traditional face-to-face classroom education, online education enables us to record and research a large amount of learning data for offering intelligent educational services. Knowledge Tracing (KT), which aims to monitor students' evolving knowledge state in learning, is the fundamental task to support these intelligent services. In recent years, an increasing amount of research is focused on this emerging field and considerable progress has been made. In this survey, we categorize existing KT models from a technical perspective and investigate these models in a systematic manner. Subsequently, we review abundant variants of KT models that consider more strict learning assumptions from three phases: before, during, and after learning. To better facilitate researchers and practitioners working on this field, we open source two algorithm libraries: EduData for downloading and preprocessing KT-related datasets, and EduKTM with extensible and unified implementation of existing mainstream KT models. Moreover, the development of KT cannot be separated from its applications, therefore we further present typical KT applications in different scenarios. Finally, we discuss some potential directions for future research in this fast-growing field.

**Index Terms**—Knowledge Tracing; Intelligent Education; Educational Data Mining; Adaptive Learning; User Modeling

# 任务一：文献调研

- 可以通过以下途径了解更多更前沿的研究领域
  - 期刊及其子刊：
    - Nature, Science, IEEE/ACM Transactions等
  - 国际/国内会议
    - 《中国计算机学会推荐国际学术会议和期刊目录》
    - 《清华大学计算机学科推荐学术会议和期刊列表》
    - 《中国科大计算机学院学位分委会认定的顶级会议与期刊》
    - 《信息与智能学部学位分委员会认定的A档/B档会议与期刊》
    - 例如，<u>KDD, WSDM, ICDM</u>, <u>AAAI, IJCAI</u>, <u>SIGIR, CIKM</u>, <u>NIPS, ICML</u>, <u>VLDB, SIGMOD, ICDE</u>, <u>ACL,EMNLP</u>, <u>WWW</u>
  - 各种科技论坛、公众号
    - 机器之心，新智元，paperweekly等
  - 在线课程

# 任务一：文献调研

- 例如，从模型角度出发：
  - 分类
    - 决策树
    - 支持向量机(SVM)
  - 聚类
    - K均值算法(K-means)
    - 基于密度的聚类(DBScan)
    - 层次聚类
  - 概率主题模型
    - 隐马尔科夫模型(HMM)
    - 隐狄利克雷分配模型(LDA)

- 深度学习
  - 卷积神经网络(CNN)
  - 循环神经网络(RNN)
  - 图神经网络
  - Transformer
- 强化学习
  - 马尔科夫决策过程(MDP)
  - 深度Q网络(DQN)
- 大模型
  - GPT模型结构
  - 提示工程，COT等

9/10/2024

# 任务一：文献调研

- 例如，从应用角度出发
  - 如，实验室主页上可下载 http://staff.ustc.edu.cn/~cheneh/
  - **教育**
    - EKT: Exercise-aware Knowledge Tracing for Student Performance Prediction
    - DisenQNet: Disentangled Representation Learning for Educational Questions
  - **推荐系统**
    - Relevance meets Coverage: A Unified Framework to Generate Diversified Recommendations
    - Personalized Travel Package Recommendation
  - **社交网络**
    - An Influence Propagation View of PageRank
    - From Footprint to Evidence: An Exploratory Study of Mining Social Data for Credit Scoring
  - AlphaX系列
  - 大模型在不同场景中的应用

# 文献调研-报告要求

- **多文献**调研报告
  - 不允许抄袭，不允许使用ChatGPT等生成式AI模型写作
  - 调研报告需有明确主题、且主题需要与课程内容相符
  - 需要有背景与现状综述，**对现状综述的总结需要有框架、有逻辑**
  - 需要有对各文献清晰具体的单独表述
  - **需要对各文献涉及模型、方法以及实验做有意义的比较分析**
  - 需逻辑通顺
  - 需要有参考文献
  - 需要**使用助教提供的LaTeX模版编写**，并最终提交PDF文档，格式要求详见LaTeX模版
  - **及时提交**

# 文献调研-报告要求

□ **多文献**调研报告，LaTeX模版编写
- □ 学会使用科学工具
- □ 掌握科学写作