基于本体的金融知识图谱自动化构建技术评测

王文广

kdd.wang@gmail.com wangwenguang@datagrand.com

王 文 广

- ✓ 达观数据副总裁
- ✓ CCF会员， CAAI会员，高工
- ✓ 多项国家发明专利
- ✓ 丰富的NLP和知识图谱领域的论文与专著
- ✓ 渊识OCR产品、渊海知识图谱平台和失效模式知识图谱平台三款产品
- ✓ 主导数十个AI领域创新课题研究和产业项目落地

kdd.wang@gmail.com wangwenguang@datagrand.com

达观数据：知识图谱与智能文本领域的知名科技企业

达观数据
DATA GRAND

- **达观数据**是专注于知识图谱和文本智能处理的人工智能准独角兽企业，国家高新技术企业，科技小巨人企业，连续3年入选中国人工智能创业企业50强

- 曾荣获中国人工智能领域最高奖 “吴文俊人工智能奖” ，中国互联网创新大赛全国总冠军，多次摘取ACM国际计算机学会算法竞赛冠军荣誉。唯一同时入选国际知名研究机构IDC和Gartner的中国智能办公软件开发的代表性企业

- 在文本自动化处理领域，拥有100多项技术发明专利和软件著作权，2本技术专著，发表多篇高水平科研论文，和三所中国知名大学建立和联合实验室

- 达观创新性的自动化软件已成功为大量全球500强企业，知名金融机构、大型政府和企业机构等提供了智能办公服务，大幅度降本增效

连续3年入选中国人工智能领域50强

任务背景

达观数据
DATA GRAND

- ✓ 金融研报是金融研究中对宏观经济、行业、产业链以及公司的研究报告。
- ✓ 研报通常是专业人员所撰写，具有数据多、深度好，质量高，内容可靠等特点。
- ✓ 对构建金融行业知识图谱来说，研报是质量非常关键的数据源。
- ✓ 由于研报本身的特点，从研报自动化构建知识图谱困难重重。
- ✓ 具有巨大的技术价值、行业应用价值。

任务介绍

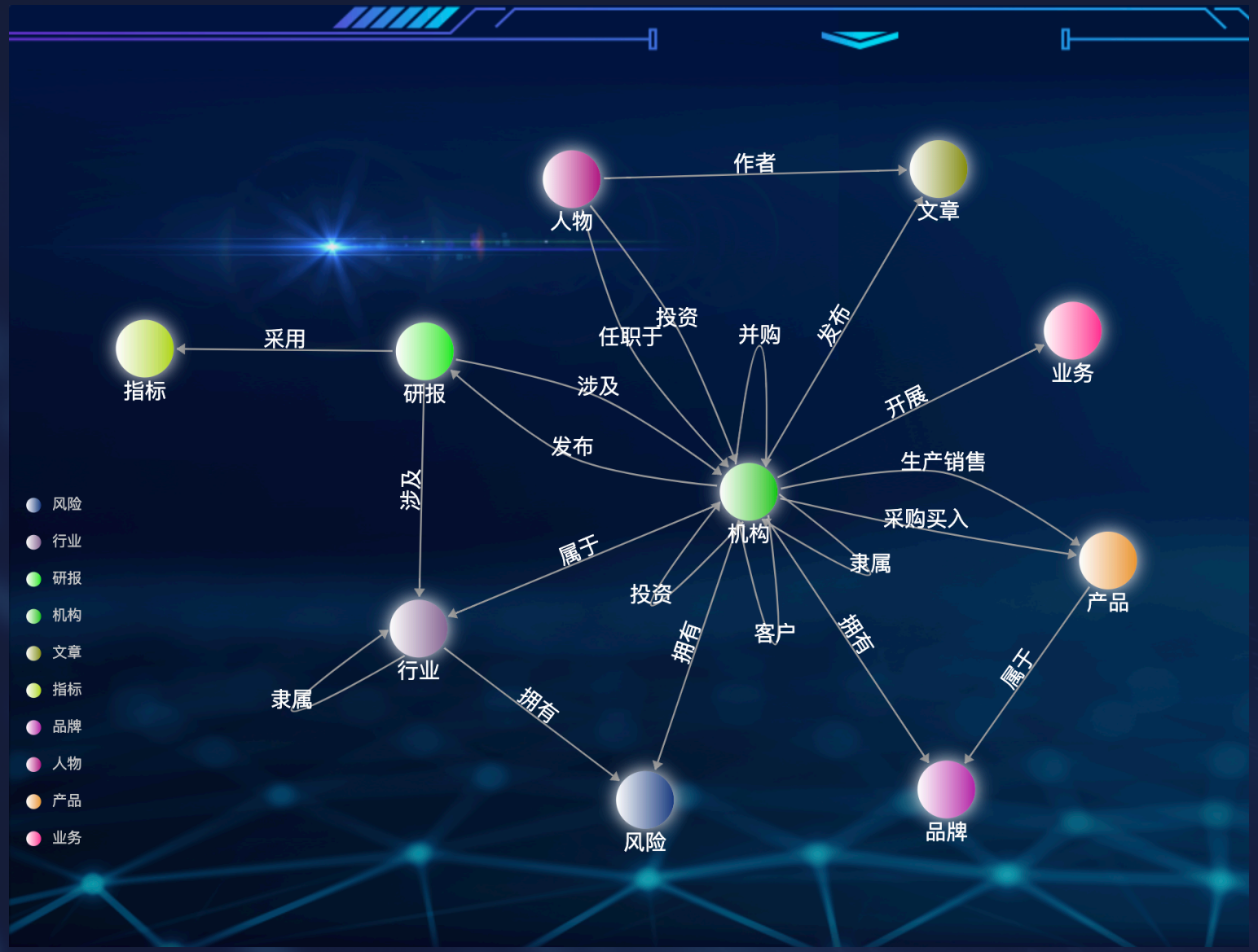
达观数据
DATA GRAND

本评测任务围绕从金融研报自动化构建知识图谱所展开：

- ✓ 给定：预定义图谱模式（Schema，本体）
- ✓ 给定：种子知识图谱开始
- ✓ 给定：金融研报的文本，经过人工处理过的txt格式
- ✓ 要求：选手实现自动化构建图谱的算法、模型和软件
- ✓ 要求包括：实体抽取
- ✓ 要求包括：关系和属性抽取
- ✓ 要求包括：实体合并和对齐
- ✓ 期望：迁移学习、无监督或弱监督、远程监督等
- ✓ 期望：多用算法少用规则

数据集：schema

达观数据
DATA GRAND



属性

```
"研报": {
    "发布时间": "date",
    "评级": "string",
    "上次评级": "string"
},
"机构": {
    "全称": "string",
    "英文名": "string"
},
"文章": {
    "发布时间": "date"
}
```

数据集：

达观数据
DATA GRAND

研报数量

1200份

图谱数据情况

	实体	关系三元组	属性三元组
种子图谱	5131	6091	354
评测图谱	12668	20707	974

本次评测任务采用精确率（Precision, P）、召回率（Recall, R）、F1 值（F1-measure, F1）来评估构建效果。

对于实体、属性三元组（属性，属性名，属性值）和关系三元组（实体，关系，实体），为三种类型的识别内容。

对每一种识别内容，相关的定义如下：

识别内容的精确率 = 识别内容与标注相同的数量 / 识别内容总数量

识别内容的召回率 = 识别内容与标注相同的数量 / 标注内容总数量

识别内容的F1 = 2 * (识别内容的精确率 * 识别内容的召回率)/(识别内容的精确率 + 识别内容的召回率)

最终结果F1定义如下：

$$F_1 = \frac{\text{实体}\,F_1 + 2 \times \text{属性}\,F_1 + 2 \times \text{关系}\,F_1}{5}$$

任务价值

达观数据
DATA GRAND

技术价值

- ✓ 实体抽取
- ✓ 关系和属性抽取
- ✓ 实体对齐
- ✓ 无监督、半监督或弱监督
- ✓ 迁移学习
- ✓ 复杂数据下的知识图谱构建
- ✓ 真实场景下的知识图谱构建
- ✓ 少样本情景下的知识图谱构建

业务价值

- ✓ 金融行业知识图谱的构建
- ✓ 产业链分析
- ✓ 供应链分析
- ✓ 投融资关系图谱及分析
- ✓ 投资分析
- ✓ 金融风险分析
- ✓ 金融监管
- ✓ .....

报名参赛情况

达观数据
DATA GRAND

总共参赛队伍数目	740
提交结果队伍数目	57
参与评奖的队伍数目	18
获奖队伍数目	5

- 有部分队伍因不愿公开方法而放弃评奖

获奖情况

达观数据
DATA GRAND

排名	参赛队名	单位	得分
1	UPSIDE-DOWN	国网信息通信产业集团有限公司	0.49704
2	Solaris99	北京大学	0.48340
3	BOOMBOOM	上海交通大学 北京元年科技有限公司 The University of Edinburgh	0.46455
4	SGIT	福建亿榕信息技术有限公司	0.45376
5	iceburg	北京邮电大学	0.41169

✓高校和企业 各一半，刚刚好！

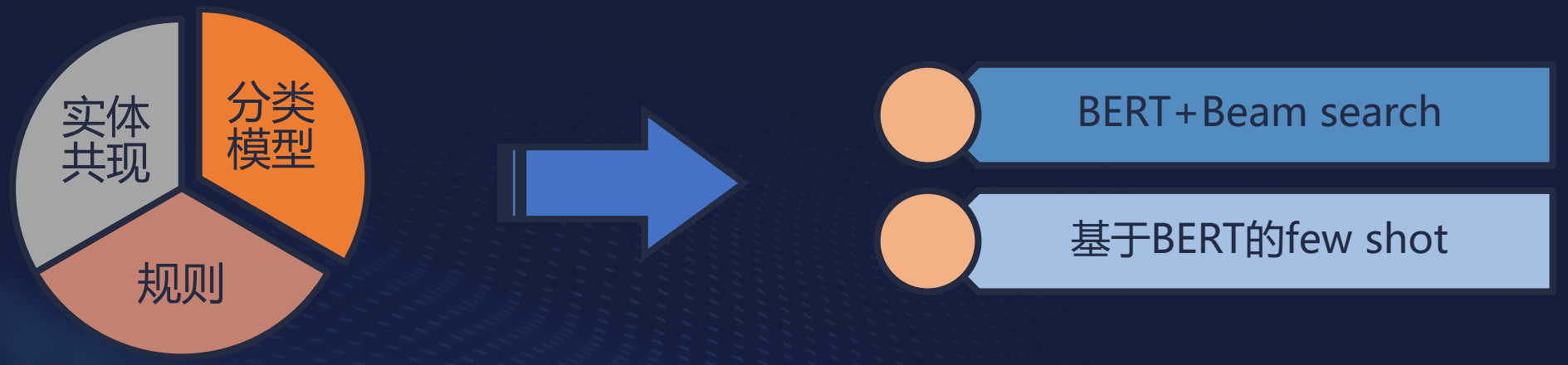
获奖技术方案总结：实体抽取

达观数据
DATA GRAND

- ✓ 在少样本情况下，直接使用BERT的效果是最好的
- ✓ BERT或其变体是当前进行实体抽取的主流
- ✓ 规则在真实场景中的价值依然很大

获奖技术方案总结：关系和属性抽取

达观数据
DATA GRAND



- ✓ 真实场景下的规则，是关系和属性识别的主要方法
- ✓ 实体间的共现是最关键的假设，这也是远程监督的基本假设
- ✓ 真实的复杂场景下的纯算法进行实体和关系的识别，还有待创新

获奖技术方案总结：杂项

达观数据
DATA GRAND

总结

达观数据
DATA GRAND

- 评测本身是完成从显示场景中抽象和简化的，有很强的现实意义
- 评测对选手的要求是综合性的，无法靠堆巨大的单一预训练模型来获得高的排名
- 评测的最终F1值为0.5，效果还是略低，与大规模应用落地的要求还有些距离
- 从选手提交的方案来看，使用了大量的规则，弱监督和远程监督等算法用的较少

- ✓ 完全自动化的实现知识图谱的构建，依赖于少样本下的关系和属性抽取技术的进一步发展
- ✓ 未来在做类似的评测，会提供一个平台，让选手专注于算法的开发
- ✓ 知识图谱的大规模应用落地，期待着更优秀的关系和属性抽取算法和模型

千层网络 • 万卷诗书
纵横交错 • 其用不穷


达观 智能办公机器人专家
达观数据
DATA GRAND