


WebShaper: Agentically Data Synthesizing via Information-Seeking Formalization

Zhengwei Tao*, Jialong Wu*, Wenbiao Yin(✉), Junkai Zhang, Baixuan Li, Haiyang Shen, Kuan Li, Liwen Zhang, Xinyu Wang, Yong Jiang(✉), Pengjun Xie, Fei Huang, Jingren Zhou

Tongyi Lab , Alibaba Group

 <https://github.com/Alibaba-NLP/WebAgent>

 <https://huggingface.co/datasets/Alibaba-NLP/WebShaper>

 <https://modelscope.cn/datasets/iic/WebShaper>

简介

本文提出 **WebShaper**，一种新颖的信息检索任务训练数据构造方法，不同于之前读过的WebDancer、WebSailor “先收集web信息、再围绕这些信息生成问题”的方法，WebShaper采用“先定义检索任务结构，再据此检索信息并生成问题”，确保每个问题背后有清晰的推理逻辑和结构控制。

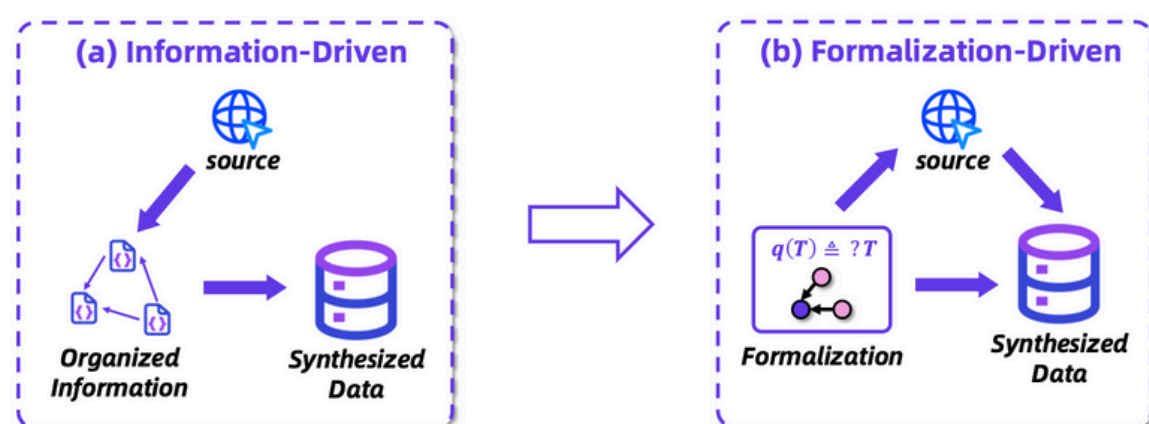
那么如何定义检索任务呢？WebShaper引入了集合论(set theory)、交并集操作，形式化定义信息检索任务的推理路径，在真正创建数据集时用一个具备搜索、总结、验证能力的 **Expander Agent** 进行基于知识图谱的分层扩展策略，逐步构建出结构更复杂、逻辑也清晰的更大的图谱。

注意：本文用集合论形式化定义检索任务，我觉得可以简单理解为三元组、知识图谱，没必要想的很复杂

背景

对于Web Agent任务来说，如何构造高质量高难度的信息检索(Information-Seeking, IS)训练数据，始终是一个关键难题。本文是之前读过的WebDancer、WebSailor续作，作者继续发力如何构建IS训练集，这一次采用了一种形式化(formalization)驱动的方法，先定义好结构化的任务公式，再通过agent去构建问题和answer信息。

WebShaper描述



实验设置

典型的TIR工作

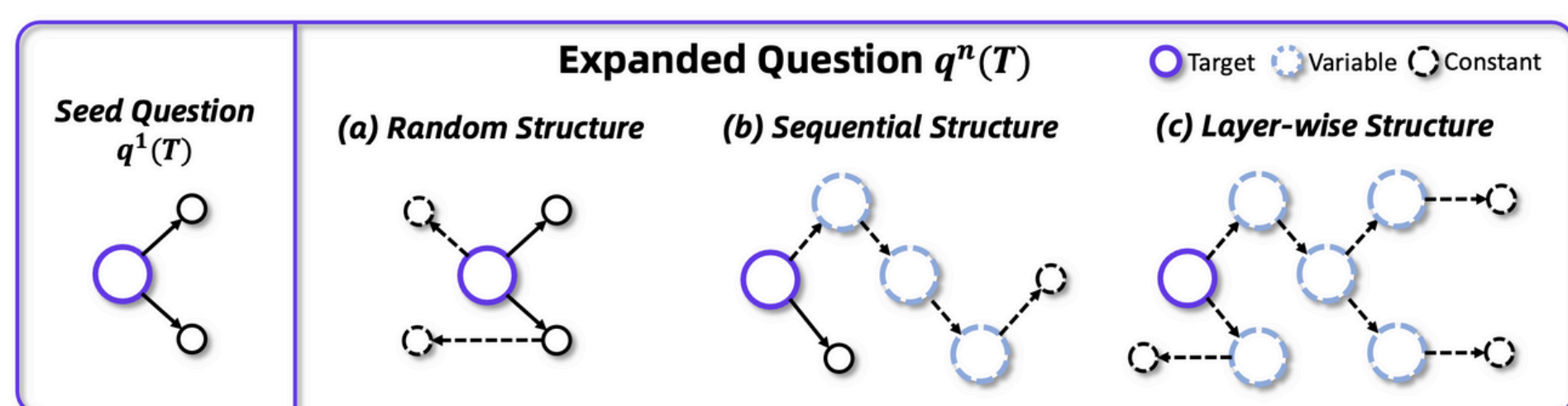
两个tool：搜索引擎和网页访问

- 实验对象：Qwen2.5 系列
- 训练：sft+GRPO两阶段
- 本文重点是构造训练集

Open-sourced Agentic Frameworks									
Qwen-2.5-32B	Search-o1	33.3	25.0	0.0	28.2	-	-	-	-
	WebDancer	46.1	44.2	8.3	40.7	44.3	46.7	29.2	38.4
	WebShaper	61.5	53.8	16.6	52.4	58.1	51.4	47.0	51.4
QwQ-32B	Search-o1	53.8	34.6	16.6	39.8	43.1	35.0	27.1	34.1
	WebThinker-Base	53.8	44.2	16.6	44.7	47.2	41.1	39.2	41.9
	WebThinker-RL	56.4	50.0	16.6	48.5	58.8	44.6	40.4	46.5
	Simple DS	-	-	-	50.5	-	-	-	-
	WebDancer	61.5	50.0	25.0	51.5	52.5	59.6	35.4	47.9
Qwen-2.5-72B	WebShaper	69.2	50.0	16.6	53.3	55.8	49.2	45.4	49.7
	WebSailor	69.2	63.4	16.6	60.1	56.2	52.1	49.5	52.2

$$T_1 = R_{\text{foundIn}}(\{1996\}) \cap R_{\text{isA}}(\{\text{East German football team}\})$$
$$q(T) \triangleq ?T = R_{\text{playAt}}(T_1) \cap R_{\text{playAt}}(\{2004\}) \cup R_{\text{playAt}}(\{2005\}) \cap \bigcup_{1900}^{2999} R_{\text{bornIn}}(\text{year})$$

Question: Which player of a team in the 2004-05 season, who was born in 90s? This team is founded in 1966 and is an East German football team.
Answer: Robert Rudwaleit, Danny Kukulies, ..



思考

我在阅读这篇论文的时候，由于作者用了一种比较新颖的形式化定义：集合、交并集。稍微有些阅读门槛，但是我发现，只要把所谓的知识投影 (knowledge projection) 看作抽象版的三元组 (实体1, 关系, 实体2)就顺畅了，剩下的就是如何构建“知识图谱”。

我们重新组织下简单的语言，看看本文在做什么：对于webagent的训练集，目前普遍采用(query, answer)的方式，其中answer是query的明确检索答案。当然了，你可能会想，不是所有的问题都有明确答案啊？OK恭喜你，你已经思考的很深入了，这个问题目前似乎没有人去尝试解决。切回正题，对于那些有明确答案的query，是不是可以把答案用“实体”描述呢？那么query是不是可以拆解为多个实体之间的关系呢？这似乎就是作者的insight，这里可能有些抽象，比如query是“一个参加过2002年世界杯并且在皇马待过的意大利球员是？”我们可以从中提取实体和关系吧？然后构成一个小图谱，所谓的找answer就是在找一个满足query的实体。这就是本文的KP和KP Representation。