多智能体自驱的微服务故障根因定位

-- 格兰杰因果和陪审团制度的融合实践

Holmes战队 - 安浩
中兴通讯股份有限公司

主办单位：中国计算机学会（CCF）
承办单位：中国计算机学会互联网专委会、中国科学院计算机网络信息中心、中国移动研究院、清华大学
协办单位：华为2012实验室、阿里云、中兴通讯、中国移动九天团队、南开大学、西安电子科技大学、清华大学计算机科学与技术系、神州灵云

目录
CONTENTS

第一章节

团队介绍

团队简介

团队名称

Holmes战队，穿梭于数据洪流中的侦探，见微知著，防患于未然。

团队成员

姜磊、胡锐、冯峻、王文譞、柏骁锐、白晓羽、安浩、周靖鹏、郭华、毛志勇

团队介绍

中兴通讯云原生应用运维引擎团队，致力于通过深化大模型与因果推理等前沿技术的融合，打造更智能、更前瞻的运维大脑，帮助企业构建无人值守的智能化运维体系，实现从“被动救火”到“主动防控”的演进，为业务的稳定与创新提供坚实保障。

第二章节

方案简介

方案简介

数据处理

原始数据

结构化MTL数据

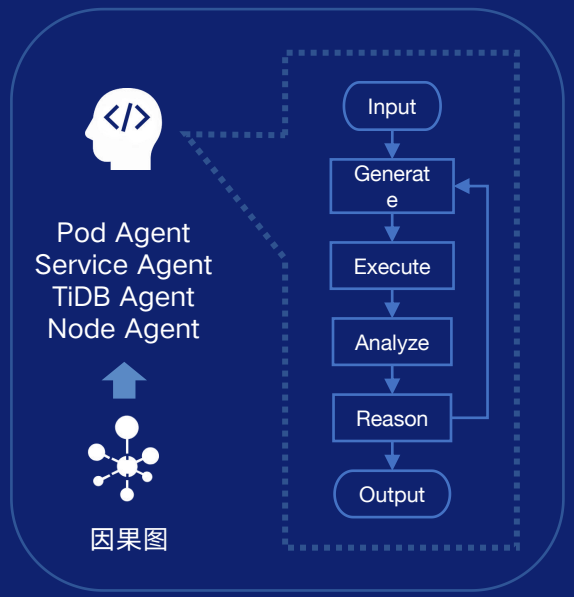
ClickHouse

LLM+格兰杰因果

因果图

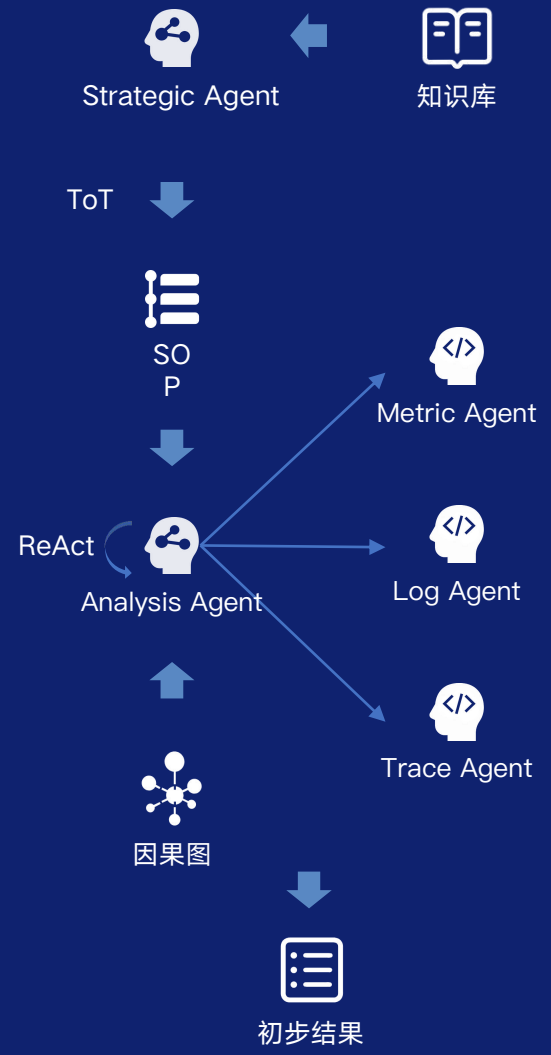
异常检测

故障查询



异常对象

根因分析



评审验证

初步结果

退回异常检测阶段　知识沉淀输出结果

最终结果

第三章节

创新性与实用性

创新点1：大模型结合格兰杰因果-问题

问题一

只使用互信息和格兰杰因果的缺陷：

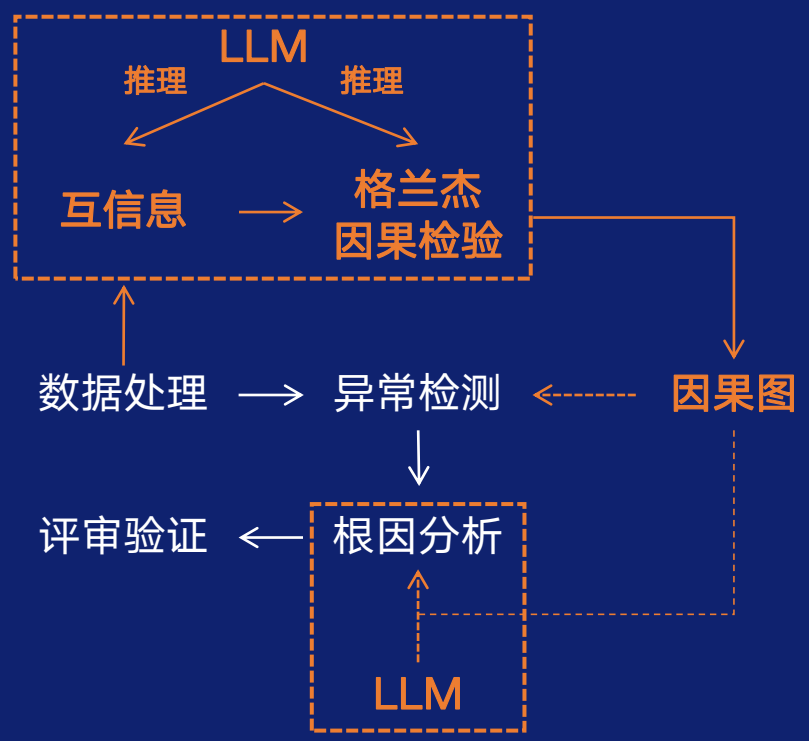
- 固定阈值的互信息可能导致指标聚合的**遗漏或误判**
- 格兰杰因果**缺乏推理和可解释性**

解决方法

使用LLM生成因果图：

- LLM**动态调整互信息阈值**，依据语义上下文与指标分布自适应判定相关性，有效解决指标间关联模糊的问题
- LLM结合语义知识对因果关系**进行逻辑论证**，为格兰杰因果链接提供可解释性

逻辑推理结合统计方法



问题二

只使用LLM完成根因分析的缺陷：

- LLM缺乏根因信息而容易产生**幻觉**
- LLM需要长思维链分析，**效率低下**

解决方法

因果图协助LLM根因分析：

- 因果图**明确变量之间的因果路径**，帮助LLM避免相关性陷阱
- 通过因果图过滤输入数据，**约束推理空间**，减少噪音干扰，提升根因分析效率

创新点1：大模型结合格兰杰因果-实践

因果图

推理

调用

返回

LLM

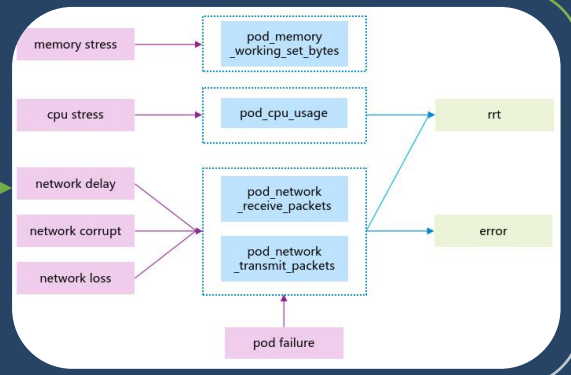
数据预处理 → 平滑性校验

最佳滞后阶数 ← 因果校验

格兰杰因果计算模块

因果图生成

pod_cpu_usage
pod_network
_receive_packets
pod_network
_transmit_packets
error
rrt
pod_memory
_working_set_bytes



生成层

指标分类

推理

调用

返回

LLM

原始数据

数据预处理 → 概率分布计算

过滤合并指标 ← 互信息值计算

互信息计算模块

pod_cpu_usage
pod_processes

rrt
rrt_max
pod_network_receive_packets
pod_network_receive_bytes
pod_network_transmit_packets
error_ratio
error
client_error

指标分类与过滤

pod_cpu_usage
rrt
pod_network_receive_packets
pod_network_transmit_packets
error

聚合层

平滑性校验
业务指标 性能指标

Service
业务指标 性能指标

业务指标知识

Node
性能指标

Database
性能指标

性能指标知识

原始数据

pod_cpu_usage
pod_processes
pod_fs_writes_bytes

rrt_max
server_error
client_error_ratio

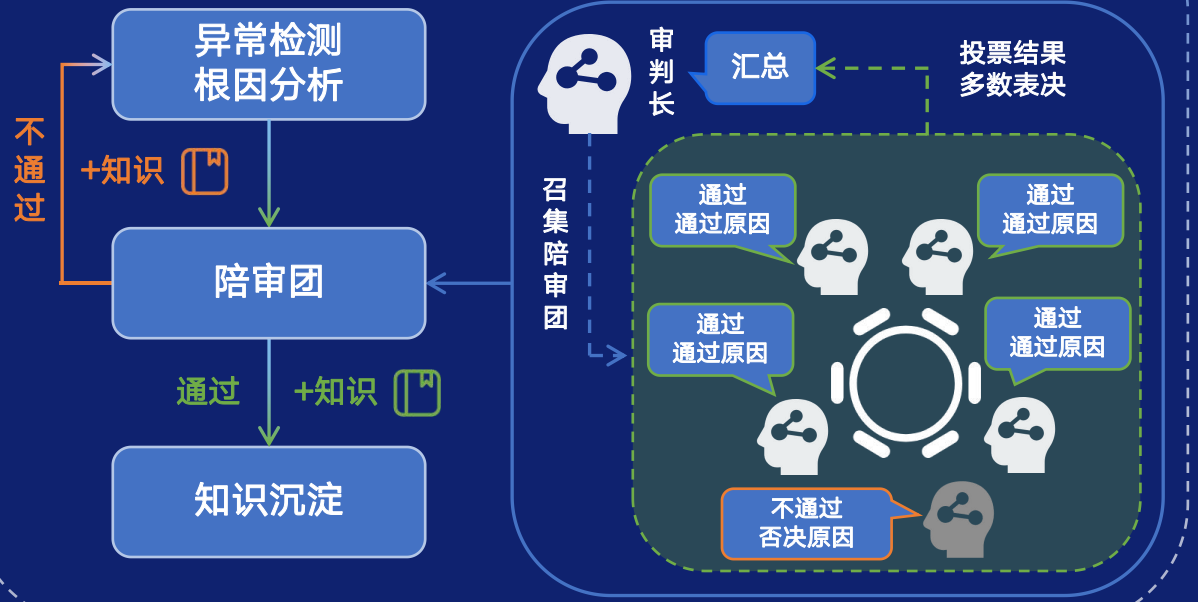
node_cpu_usage_rate
node_filesystem_usage_rate
node_disk_written_bgytes_total

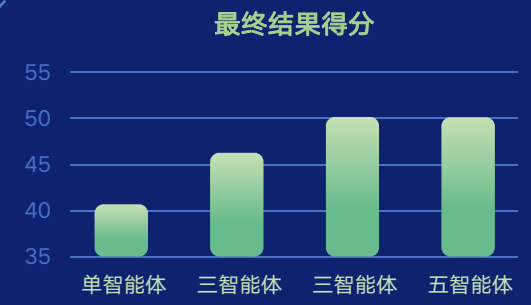
connection_count
failed_query_ops
store_slow_count

数据层

创新点2：陪审团制度

陪审团制度流程



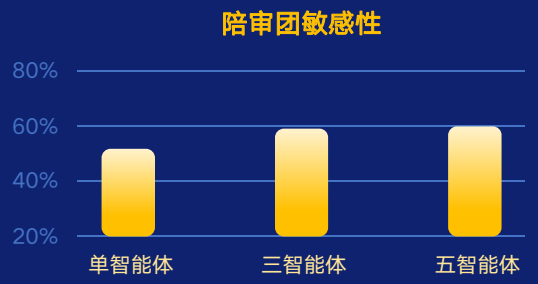


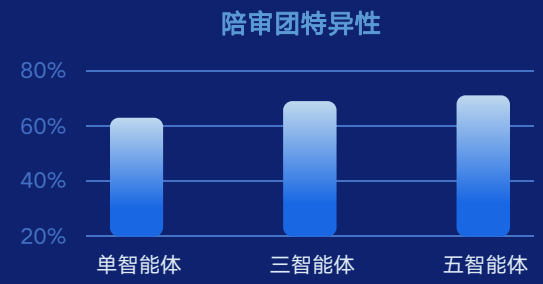
敏感性：

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

特异性：

$$\text{Specificity} = \frac{TN}{TN + FP}$$





陪审团制度数据结果

3个智能体陪审团

陪审团显著提高分数：41.18 -> 50.42

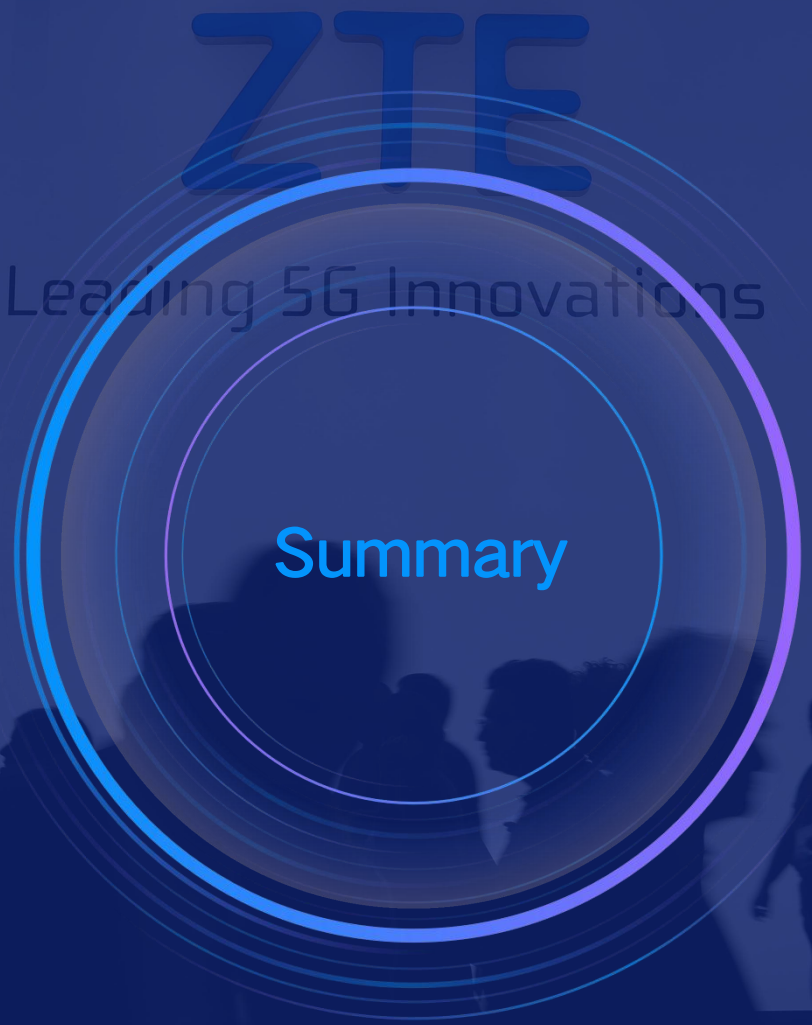
使用3个智能体陪审团效率高

陪审团能使错误案例重新进行根因分析

第四章节

总结展望

总结



● 数据处理

对指标、日志、调用链等原始遥测数据进行清洗、转换、标准化处理，建立高效检索库和因果图，提供高质量、结构化的数据底座。

● 异常检测

采用多Agent并行计算模式，在各数据维度上协同检测，精准识别异常Pod、Service等集合。将全局问题收敛至具体范围，为根因分析提供明确目标，极大提升分析效率。

● 根因分析

运用故障分析思维树（ToT） 拆解复杂任务，生成分析SOP，并结合格兰杰因果算法量化指标间因果关系。

● 评审验证

引入多智能体“陪审团”制度，模拟专家会审，对结果进行集体裁决，评审不通过则回流重析。确保结果高可靠性，并将每次分析内容转化为知识沉淀，实现系统的自我进化。

展望

大模型结合因果阶梯

突破传统关联分析，赋予智能体理解“为什么”的能力，实现从诊断、预测到干预的因果决策闭环。

赋予陪审团成员个性化认知Profile

为每个智能体注入自适应认知模型，使其能基于反馈动态调整决策偏好与专业领域，持续进化；

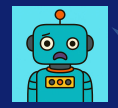
设计基于共识、辩论与证据加权的交互协议，使群体决策过程更加透明、可靠且逼近最优解。

附件：工作流程实例化

1 

在6月3日上午6点的10分到30分时间段内发现故障，请分析根因。

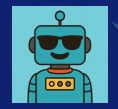
异常检测阶段

2 

通过1-sigma检测到frontend-[0-2]、adservice-[0-2]的rrt指标存在异常，通过P90检测到cartservice-[0,1]的timeout指标异常。

3 

通过2-sigma检测到frontend、adservice的rrt指标存在异常。

4 

检测到node-0的memory使用率超过70%。

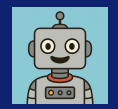
5 

TiDB没有检测到异常。

根因分析阶段

6 

1. 使用Trace Agent，检测frontend调用adservice的span是否异常。
2. 使用Log Agent，检查frontend-[0-2]、adservice-[0-2]、cartservice-[0,1]的日志信息。
3. 使用Metric Agent，检测adservice-[0-2]的cpu和io的指标是否异常。
4. 使用Metric Agent，检测frontend-[0-2]的cpu和io的指标是否异常。
5. 使用Trace Agent，检测调用cartservice的span是否异常。
6. 使用Node Agent，检测发生异常的pod是否在node-0上。

7 

请Trace Agent分析6月3日上午6点的10分到30分时间段内frontend调用adservice的span是否异常。

8 

该时间段内frontend调用adservice时间异常上升，但返回值和返回消息无异常。

9 

请Log Agent分析6月3日上午6点的10分到30分时间段内frontend-[0-2]、adservice-[0-2]的日志信息。

10 

adservice-[0-2]的日志中检测到'stress'关键字，frontend-[0-2]的日志信息无异常。

11 

请Metric Agent分析6月3日上午6点的10分到30分时间段内adservice-[0-2]的cpu和io的指标是否异常。

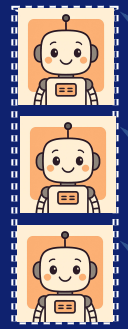
12 

检测到adservice-[0-2]pod_cpu_usage异常，其他指标无异常。

13 

此次故障是服务级别，造成故障的服务是adservice，故障原因是cpu stress。

评审验证阶段

14 

Agent运行过程正常，答案形式化检查通过，Agent分析逻辑检查通过。

Agent运行过程正常，答案形式化检查通过，Agent分析逻辑检查通过。

不通过，原因：Agent运行轮次过长。

15 

根因分析结果正确，将此次分析加入案例库。

THANKS

主办单位：中国计算机学会（CCF）

承办单位：中国计算机学会互联网专委会、中国科学院计算机网络信息中心、中国移动研究院、清华大学

协办单位：华为2012实验室、阿里云、中兴通讯、中国移动九天团队、南开大学、西安电子科技大学、清华大学计算机科学与技术系、神州灵云