基于网络拓扑及告警的故障根因定位系统实现及算法研究

时间：2020-03-06        来 源：中国软件杯

**第九届中国软件杯大赛--A组赛题**

|  |  |
| --- | --- |
| **赛题名称:基于网络拓扑及告警的故障根因定位系统实现及算法研究**  **组类： A 本科及以上** | |
| 赛题简介：介绍整个赛题的实现目标、实用价值、涉及技术和整体要求 | 大型电商平台内部涉及上百个系统间的相互调用，每天会产生上万条告警数据。如何利用网络拓扑信息及告警数据，及时有效的对告警进行过滤、分析，最终给出有效告警及疑似根因是网络运维面临的主要挑战。  本赛题的目标是分析实际生产中的网络拓扑及告警数据，设计并实现一种故障根因定位算法，准确的定位出故障根因节点。  赛题的价值在于准确、快速的定位网络故障，提升一线网络运维的效率，降低网络故障产生的损失。  该赛题主要涉及数据分析、图论、信息论、自然语言处理、神经网络等技术，要求参赛队员掌握相关的基础知识，具有python数据分析、处理及功能开发能力。 |
| 赛题业务场景：描述赛题相关的真实企业业务背景。从真实场景中，适当简化或者提炼出适合比赛的赛题场景 | 赛题业务背景：网络拓扑中一个节点出现故障，往往会导致与其相连的其他节点也发生异常，进而产生大量告警，将真正根因淹没掉。当出现大量告警时，为保证电商业务的稳定运行，我们需要对这些告警进行分析处理，过滤掉无效的告警，准确定位出可能的疑似根因节点，缩短故障定位时间。  在实际中，同一故障触发的告警信息往往相似的。网络故障定位的一般过程为：先根据历史告警数据及标记进行离线分析，总结因果关系的规律；然后在线分析阶段，首先对告警数据进行去噪和聚类，然后对类内的告警进行因果关系匹配，得到当前告警的可能根因。  赛题场景：提供实际生产中的网络拓扑数据，100组不同时间段的告警数据，每组中可能有根因，也可能没有根因，根据拓扑数据及告警数据，设计并实现故障根因定位算法，最终在测试数据上评估告警根因定位的F1值，可视化展现根因信息及根因关联的局部拓扑图。 |
| 基本功能要求 | 1、基于提供的训练数据，设计并实现故障根因定位算法。 2、故障根因定位算法中要包含告警数据预处理和故障根因定位两个模块；  3、可视化展现根因信息及根因关联的局部拓扑图。 |
| 非功能性要求 | 系统运行稳定、可靠；  能满足故障根因定位实时性的要求； |
| 实现条件：开发环境、实验平台、开发语言、数据库、编译器、涉及硬件等实现条件 | 操作系统：Linux  开发语言：python |
| 测试数据或平台：提供给参赛者的测试环境和测试数据。 | 训练数据：  1、实际生产中的部分网络拓扑数据，json格式的dict，key为源节点，value为目标节点list  2、和网络拓扑相关的告警数据，csv格式，包括告警节点，告警信息，告警时间，告警级别，是否为根因  3、提供100组不同时间段的告警数据，每组中可能有根因，也可能没有根因，如果有根因，只有1个准确的根因。  测试数据：  1、提供20组训练数据，格式同训练数据，不提供根因标注  数据样例：  1、拓扑图样例：  key为源节点，value为目标节点list，类似下面这样：AAA表示系统名，NG表示是Nginx。  {"AAA-10.10.0.1": ["AAA-NG-10.10.0.2", "AAA-NG-10.10.0.3", "AAA-NG-10.10.0.4"], "AAA-NG-10.10.0.2": ["AAA-JBOSS-10.10.0.5", "AAA-JBOSS-10.10.0.6", "AAA-JBOSS-10.10.0.7"]}  2、告警数据样例：  告警数据为csv格式，如下面所示，其中sysEname为系统名称, software为软件类型,time告警时间,triggername告警信息，里面包含ip信息，isrootcause表示是否为根因，0-否，1-是。  ,sysEname,software,time,triggername,isrootcause  0,AAA,Elasticsearch 1.7.0 ,2019-06-01 00:01:04,"eth0 网卡接受数据流量,net.if.in[eth0],主机10.10.0.8(VM)网卡eth0接收流量持续10分钟平均值大于180M",0  1,BBB,WildFly standalone ,2019-06-01 00:01:05,主机10.10.0.9日志中有READONLY信息,0  2,BBB,WildFly standalone ,2019-06-01 00:01:06,主机10.10.0.10日志中有READONLY信息,0 |
| 开发所需设备及设备指标需求说明 | 个人笔记本电脑、台式机、服务器、公有云均可。 |
| 文档及其他要求 | 算法原理说明文档：详细说明算法的实现原理，以及处理的各个步骤。  系统详细设计报告：介绍系统的整体结构，以及接口定义，模块设计，并对各个模块提供全面、准确的功能描述。  测试报告：要求包含测试环境，运行时间，运行结果等。要给出测试集的定位结果，并对根因信息及根因关联的局部拓扑图进行可视化展现。特别注意：测试报告中要包含测试集预测结果，以进行F1值计算。 |
| 评审要点 | 测试集F1值：F1=2\*准确率\*召回率/(准确率+召回率)，其中准确率=准确定位根因的组数/定位根因的组数,召回率=准确定位根因的组数/实际存在根因的组数  系统功能：要求能可视化展现根因信息及根因关联的局部拓扑图；要求系统运行稳定、可靠、流畅；  系统设计：要求目录结构设计合理；功能模块划分合理。代码要求整洁，易读；符合程序编码规范，命名规范；注释清楚；  文档：要求提供《算法原理说明文档》，《系统详细设计报告》，《测试报告》等。要求提供的文档要完整（缺少会扣分）。文档结构要合理，表达要清晰。 |
| 初赛作品提交要求 | 算法的源代码：要求包含整个项目源代码。如使用开源代码，需进行说明。  相关说明文档：包括《算法原理说明文档》，《系统详细设计报告》，《测试报告》，详细要求见上文。  演示ppt。 |

数据示例下载  http://www.cnsoftbei.com/uploads/allimg/200320/1623024240-0.gif[数据示例.zip](http://www.cnsoftbei.com/ueditor/php/upload/file/20200306/1583466305271970.zip)

训练和测试数据下载  http://cnsoftbei.com/include/ueditor/dialogs/attachment/fileTypeImages/icon_rar.gif[data\_release.rar](http://www.cnsoftbei.com/ueditor/php/upload/file/20200320/1584692570230885.rar)

**出题企业：南京苏宁软件技术有限公司**

答疑联系人：施云涛、夏丹

答疑邮箱：13901580054@163.com、shiyuntao@suning.com

答疑微信：13901580054