必示科技2020年度
告警数据挖掘方向顶会论文

隋楷心 博士
必示科技 AIOps算法部总监

故障危害

由于大规模软件服务的复杂性，故障总是不可避免的

2020年3月，google云服务由于基础资源问，题造成长时间异常，持续3.5h

2019年12月，微信公众号系统异常，持续30min

2019年12月，支付宝系统因光纤被挖断，造成90min故障

2019年12月，微软Azure系统因软件bug导致故障，持续164min

2019年10月，微信支付系统故障，持续约1h

2019年10月，国庆期间携程酒店系统故障，持续约1天

2018年一小时的宕机时间

损失约1亿美元

智能故障分析

保证软件服务质量和用户体验，减少经济损失，对故障进行智能分析是至关重要的

如何及时、准确的发现故障

如何快速、准确的诊断出故障的根因

如何对故障提前预警，规避故障

告警数据挖掘在故障智能分析中的应用

如何及时、准确的发现故障

告警定级策略不合理，影响故障发现的时效性

自适应的告警动态定级

如何快速、准确的诊断出故障的根因

告警风暴影响故障诊断效率

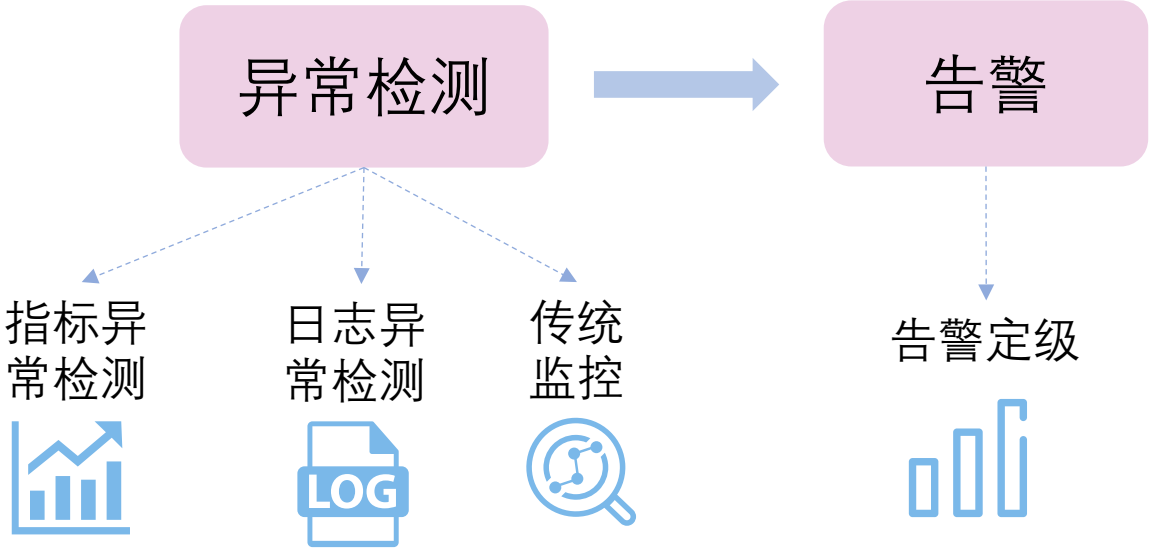
告警风暴摘要

如何对故障提前预警，规避故障

告警中的隐患信号未被有效捕捉，故障预警困难

基于告警的事件预测

问题1：告警定级策略不合理，影响故障发现时效性



Time	Severity	Type
2019-02-20 10:04:32	P2-error	Memory
AppName	**Server**	**Close Time**
E-BANK	IP(*.*.*.*)	2019-02-20 10:19:45
Content		
Current memory utilization is 79% (Threshold is 60%).		
Resolution Record		
Contact the service engineers responsible for E-BANK and get a reply that there is no effect on business, then close the alert.		

告警数据样例

问题1：告警定级策略不合理，影响故障发现时效性

TCP无响应

业务成功率低于95%

CPU利用率超过80%

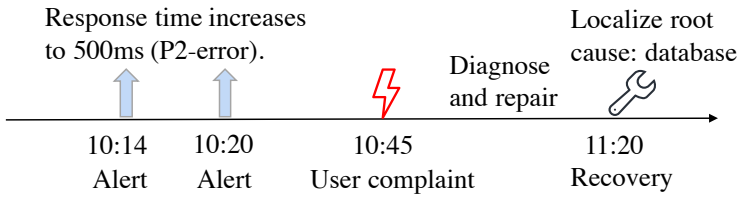
进程数超过限制

PING *** 不通

数据库表空间不足

问题1：告警定级策略不合理，影响故障发现时效性

故障案例

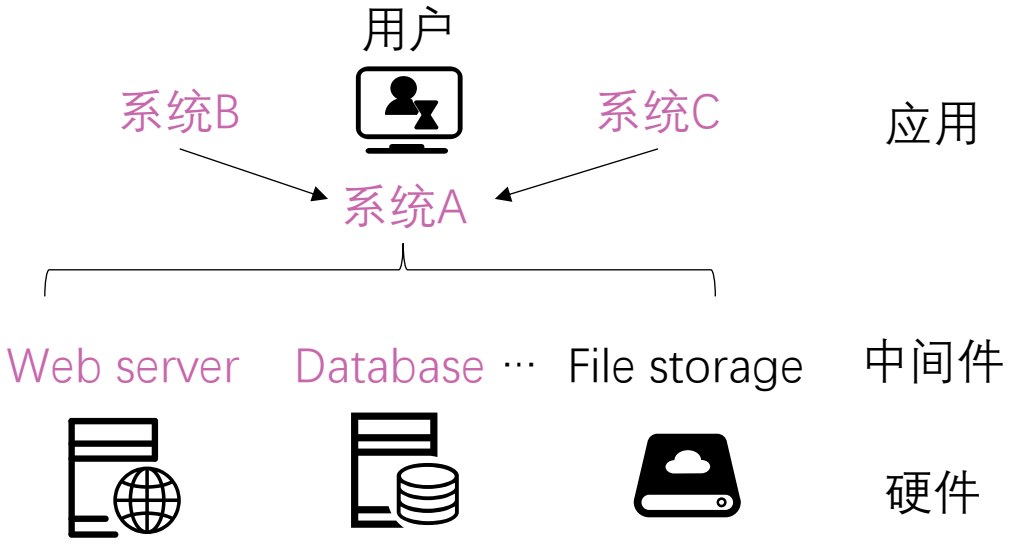


- 故障在10:45通过用户电话报障发现
- 在10：45之前，有一些故障相关的告警（响应时间上升到500ms），但是因为级别不够高，没有引起工程师的注意

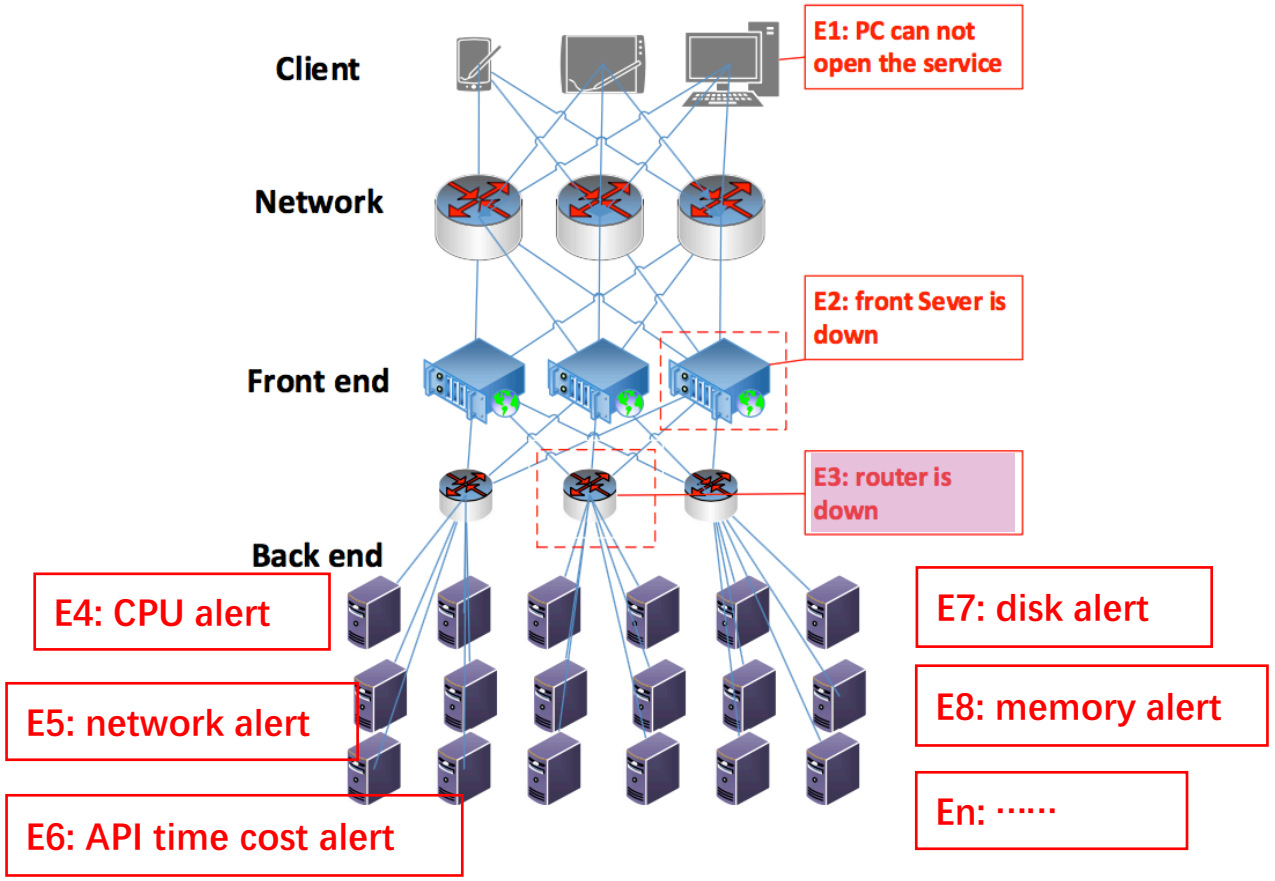
不准确的告警定级方式会导致严重告警的漏报，而工程师却花费了无用的时间去处理一些没有意义的非严重告警

准确的告警定级算法 → 有助于 → 及时的故障发现

问题2：告警风暴影响故障诊断效率



系统A的数据库服务器宕机会引发web server产生告警，调用系统A的相关系统（B和C）产生告警



路由器问题引发的告警风暴

问题2：告警风暴影响故障诊断效率

服务故障 → 告警风暴 → 

1. 到底发生了什么故障
2. 哪些告警是和故障相关的
3. 告警风暴的核心内容是什么

1. 应用架构有很多组件，每个组件下面都有自己的监控数据
2. 复杂的依赖关系：硬件拓扑和软件调用

手工检查所有告警对工程师来说是耗时耗力且易错的

问题3：告警中的隐患信号未被有效捕捉，故障预警困难

学术界现有的故障预测方法主要有以下几个限制：

- 大多数方法专门为某一种故障设计（比如磁盘故障，节点故障，交换机故障），不具备泛化性
- 这些方法大多是利用指标和日志数据来提取有预测作用的征兆特征，训练开销非常大
- 轻量级的告警数据来做事件预测仅仅考虑了每类告警的数量作为特征，实际表现并不理想

工业设备日志 → 设备故障预测

节点性能指标 → 节点故障预测

磁盘SMART → 磁盘故障预测

交换机日志 → 交换机故障预测

问题3：告警中的隐患信号未被有效捕捉，故障预警困难

工业界也没有用告警数据做故障预测的有效方案

- 基于专家知识和运维经验，总结事件预测的规则，如果线上告警满足了某一规则，就认为要发生对应的事件。但是基于规则的方法在实际中表现的并不好，经常会出现误报和漏报，原因如下：
  - 维护和制定这些规则需要足够的运维经验，且耗费时间
  - 不同工程师制定规则的偏好是不一样的，很难有统一的标准
  - 服务系统总是会经历不停的变更迭代，固定的规则不能适应动态的环境

- 基于频繁项集挖掘（如FP-growth）的方法。但是基于工程师的反馈，这类方法只能覆盖极小部分的事件，由于告警数据的复杂性和告警内容中混在的参数，大多数事件都没有对应的频繁项告警，因此这类方法在实际中的可用性也不高。

告警数据挖掘研究

故障发现

论文1：自适应的告警动态定级

Automatically and Adaptively Identifying Severe Alerts for Online Service Systems

INFOCOM 2020

故障诊断

论文2：告警风暴摘要

Understanding and Handling Alert Storm for Online Service Systems

ICSE SEIP 2020

故障预测

论文3：基于告警的事件预测

Real-time Incident Prediction for Online Service Systems

ESEC/FSE 2020

告警数据挖掘研究

用历史告警处置记录做标注，利用多特征融合方法训练排序模型，对线上实时到达的告警流做排序

故障发现

论文1：自适应的告警动态定级

Automatically and Adaptively Identifying Severe Alerts for Online Service Systems

INFOCOM 2020

故障诊断

论文2：告警风暴摘要

Understanding and Handling Alert Storm for Online Service Systems

ICSE SEIP 2020

故障预测

论文3：基于告警的事件预测

Real-time Incident Prediction for Online Service Systems

ESEC/FSE 2020

工业界实践

基于手工规则的告警定级

P1-严重 > P2-错误 > P3-警告

指标规则: 固定阈值
- CPU 利用率超过 90% 是P2
- CPU 利用率超过 80% 是P3

日志规则: 关键字匹配
- 日志文件中出现fail关键字是P1
- 日志文件中出现warning关键字是P3

不足：
- 简单的手工规则的定级方式不能满足线上服务系统的动态性与复杂性
- 手工定制和维护大量的规则是非常耗时耗力的

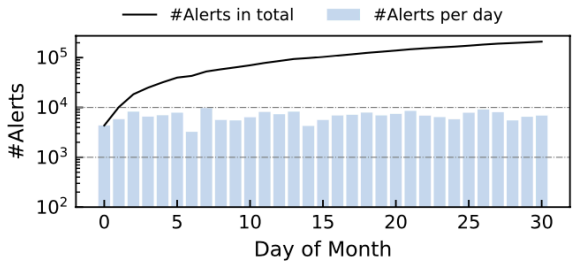
1) 告警的种类很多
2) 由于系统变更会引入新类型的告警
3) 不同的工程师在制定规则时的偏好不一样，很难有统一的规则

自适应的告警动态定级算法 → 有助于 → 及时的故障发现

挑战

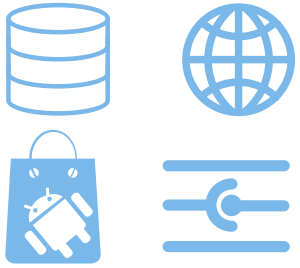
1. 标注开销大

大量的告警数据



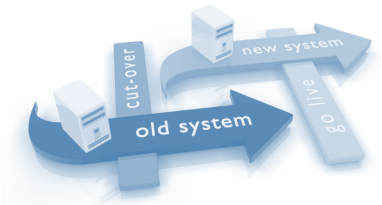
2. 告警类型复杂多样

应用，网络，数据库，中间件，存储



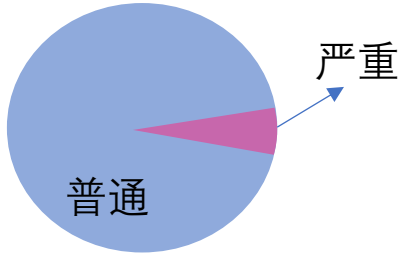
3. 复杂动态的线上环境

适应动态的线上环境，如系统变更引入的新告警

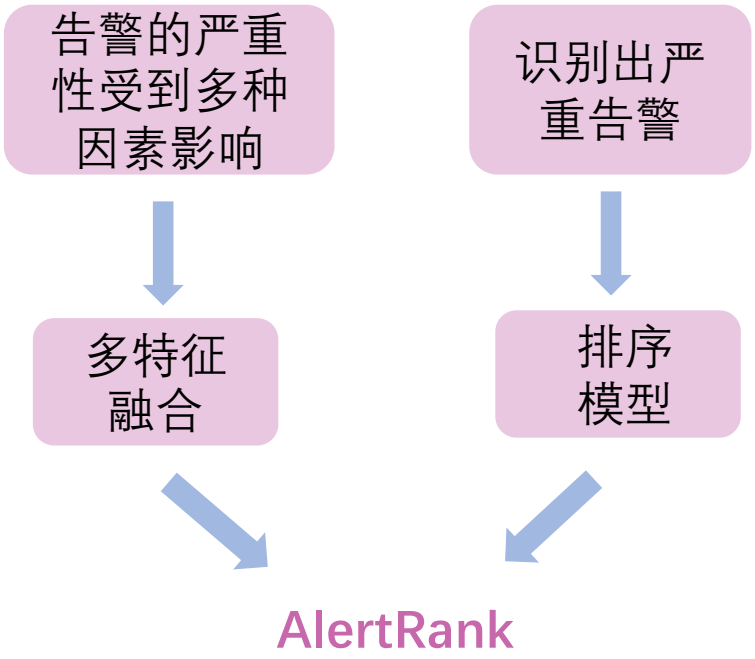


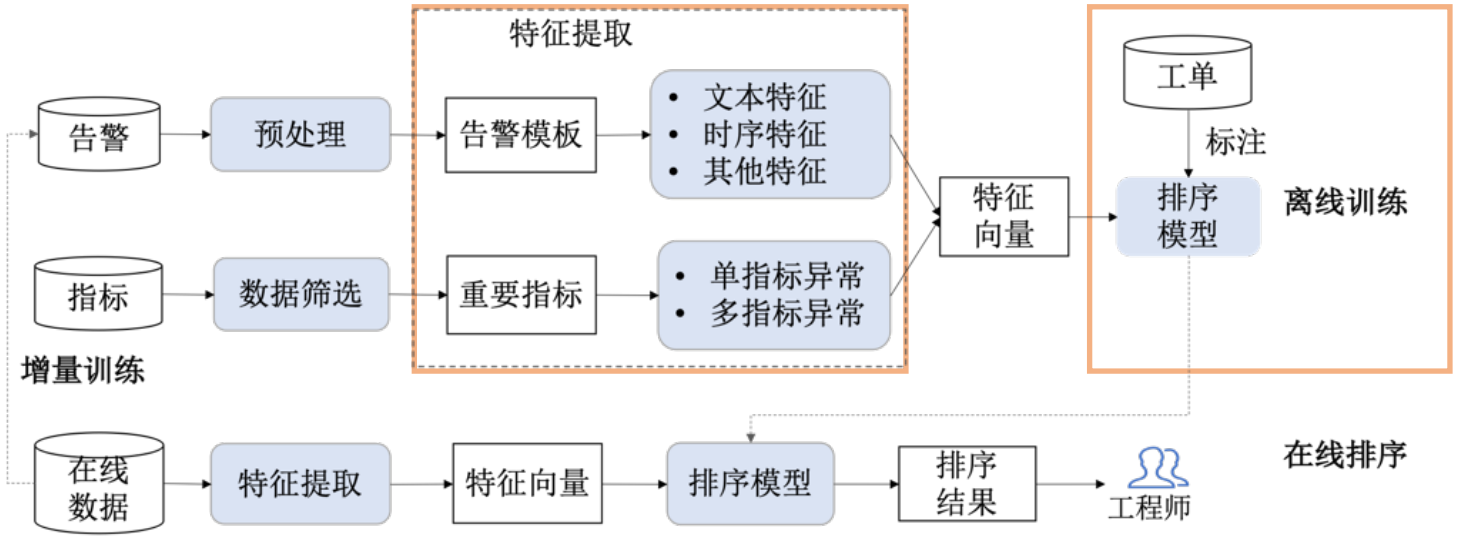
4. 数据不平衡

严重告警的比例很小



方法





核心思想:

- 多特征融合: 告警特征和指标特征
- 建模成ranking model

数据描述

- 告警
- KPI
- 工单

TABLE I: An example alert.

Time	Severity	Type
2019-02-20 10:04:32	P2-error	Memory
AppName	**Server**	**Close Time**
E-BANK	IP(*.*.*.*)	2019-02-20 10:19:45
Content		
Current memory utilization is 79% (Threshold is 60%).		
Resolution Record		
Contact the service engineers responsible for E-BANK and get a reply that there is no effect on business, then close the alert.		

用于给每一个告警的严重性打分

1. None. (0; 65.1%)
2. This alert is in white list. (0.1; 4.2%)
3. This alert has been recovered automatically. (0.2; 7.8%)
4. Contact the service engineers and there is no effect on business. (0.4; 10.6%)
5. Known reasons. This alert has been resolved. (0.6; 6.4%)
6. Contact the service engineers and there is an effect on business. Already resolved. (0.8; 3.8%)
7. Create a ticket. (1; 2.1%)

不同类型的处置记录

特征工程

告警特征

多特征融合

指标特征

1

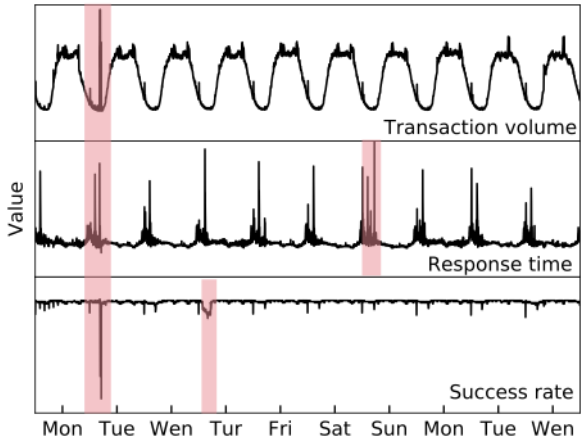
文本特征:
- 告警语义：主题模型
- 文本熵: Inverse document frequency

2

时间特征:
- 频率
- 周期性
- 告警数量
- 间隔时间

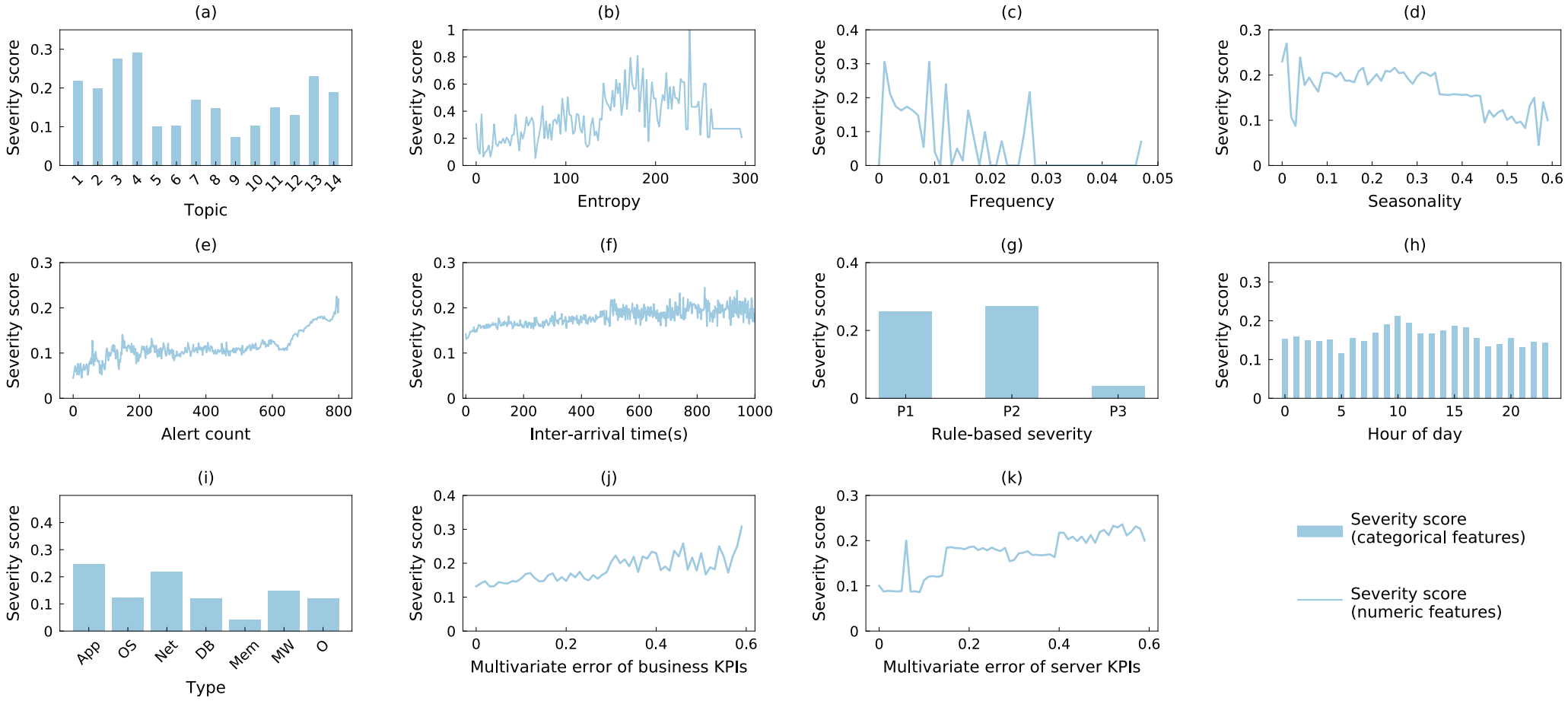
3

其他特征：
- 告警类型
- 发出时间…



- 所在应用系统的关键业务指标：如交易量，响应时间，成功率
- 异常检测的异常分数作为特征

特征分析



一些典型的特征和告警严重性之间的关系

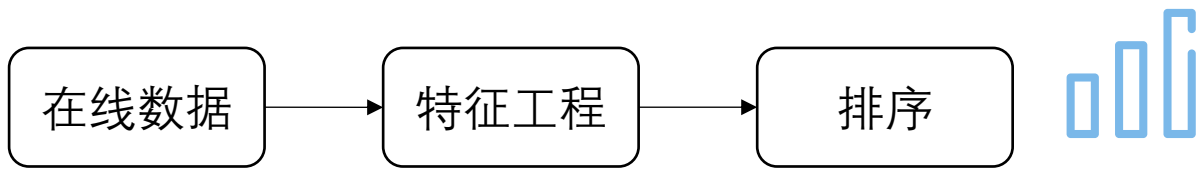
排序

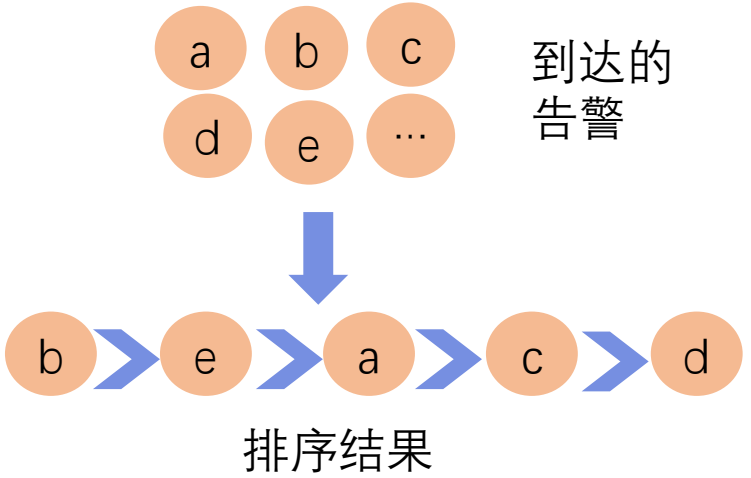
特征工程 —训练→ 排序模型

Feature Type	Feature Name	#Feature
Alert Textural	BTM Topics (14), Entropy (1)	15
Alert Temporal	Frequency (1), Seasonality (1), Alert count (4), Inter-arrival time (1)	7
Alert Attributes	Original severity (1), Alert time (3), Type (1)	5
KPI Anomaly	Univariate anomaly (11), Multivariate anomaly (2)	13

算法中用到的特征（40个）

- 标签: 来自历史告警数据的处置记录
- 模型: XGBoost ranking





评估

数据集：来自真实工业界的三组告警数据集

Datasets	A			B			C		
Methods	P	R	F1	P	R	F1	P	R	F1
AlertRank	**0.85**	**0.93**	**0.89**	**0.82**	**0.90**	**0.86**	**0.93**	**0.92**	**0.93**
Rule-based	0.43	0.68	0.53	0.47	0.70	0.56	0.41	0.74	0.53
Bug-KNN	0.72	0.76	0.74	0.79	0.62	0.70	0.80	0.53	0.64

- 对比方法：
  - Rule-based：传统基于规则的方法
  - Bug-KNN: bug report定级算法
- AlertRank可以达到平均0.9的F1-score

Datasets	A			B			C		
Methods	P	R	F1	P	R	F1	P	R	F1
AlertRank	**0.85**	**0.93**	**0.89**	**0.82**	**0.90**	**0.86**	**0.93**	**0.92**	**0.93**
Alert Only	0.82	0.79	0.80	0.75	0.80	0.77	0.67	0.77	0.72
KPI Only	0.42	0.40	0.41	0.32	0.39	0.35	0.36	0.31	0.33

告警特征和指标特征两部分对结果都是有贡献的

告警数据挖掘研究

故障发现

论文1：自适应的告警动态定级

Automatically and Adaptively Identifying Severe Alerts for Online Service Systems

INFOCOM 2020

故障诊断

论文2：告警风暴摘要

Understanding and Handling Alert Storm for Online Service Systems

ICSE SEIP 2020

对告警风暴做实时的检测和摘要（降噪和聚类），减少工程师需要查看的告警的数量，快速定位问题

故障预测

论文3：基于告警的事件预测

Real-time Incident Prediction for Online Service Systems

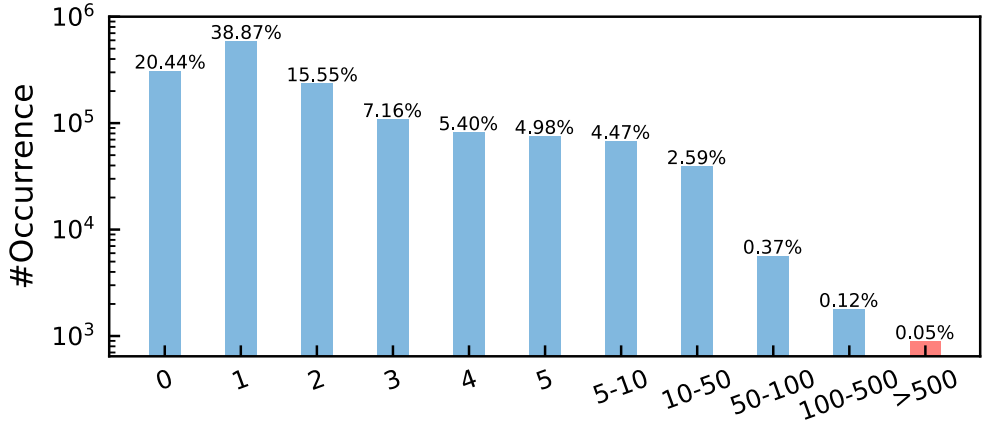
ESEC/FSE 2020

问题定义

服务故障 → 告警风暴

1. 到底发生了什么故障
2. 哪些告警是和故障相关的
3. 告警风暴的核心内容是什么

- 告警风暴：一般情况下，当一个应用服务出现故障时会伴随着大量的告警产生
- 由于告警风暴中告警的数量巨大，运维工程师手工检查每条告警进而对系统排障的过程会变得非常费时费力



X：系统每分钟产生的告警数量 Y：不同#alert/min对应的次数

因此，为了保证服务质量，对告警风暴的处置至关重要

现状和挑战

1. 告警风暴出现的频率很高（大约每周一次）并且在现实中会给工程师带来巨大困难

2. 当前对告警风暴的识别方法只是人工设置一个固定的阈值，但是这种方法并不能适应动态的在线服务场景

3. 告警风暴中包含一部分于服务故障无关的常规告警，并且许多与故障相关的告警中也存在一定的关联关系，例如文本相关性和拓扑相关性

1. 软件服务频繁变更引发告警数量动态变化 → 固定阈值的告警风暴检测方法表现不好

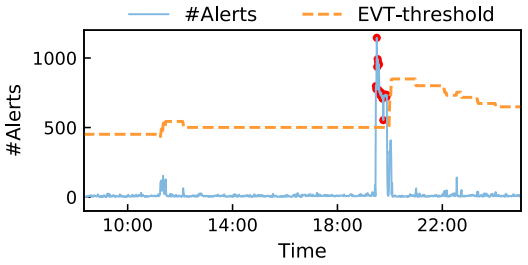
2. 故障不相关的告警混在在告警风暴中，影响排障 → 告警降噪：过滤掉噪声告警

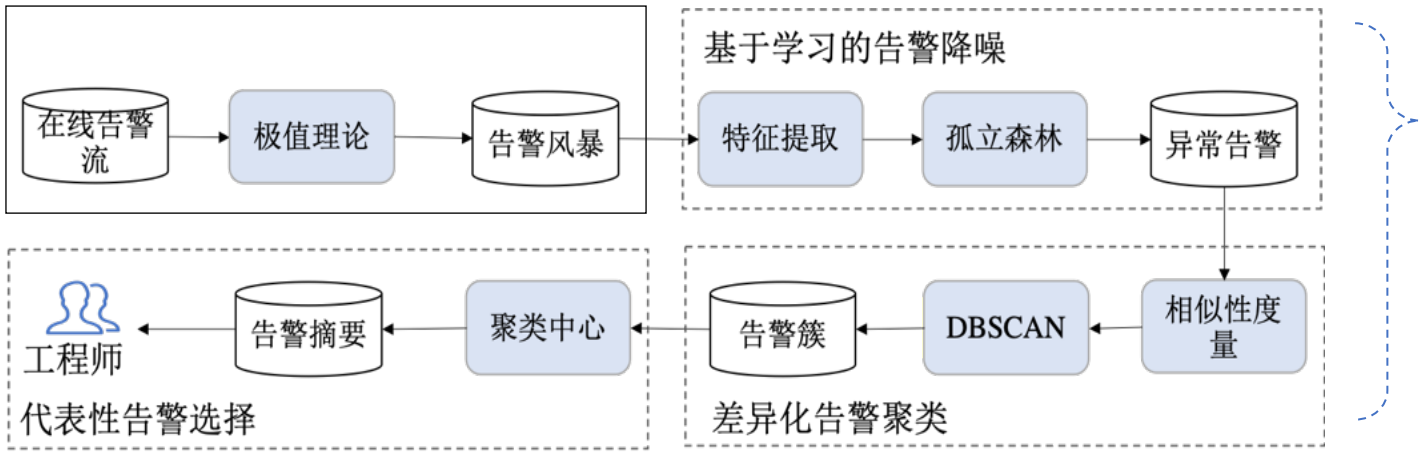
3. 告警之间复杂的关联关系 → 拓扑关联；内容相似

方法

① 告警风暴检测

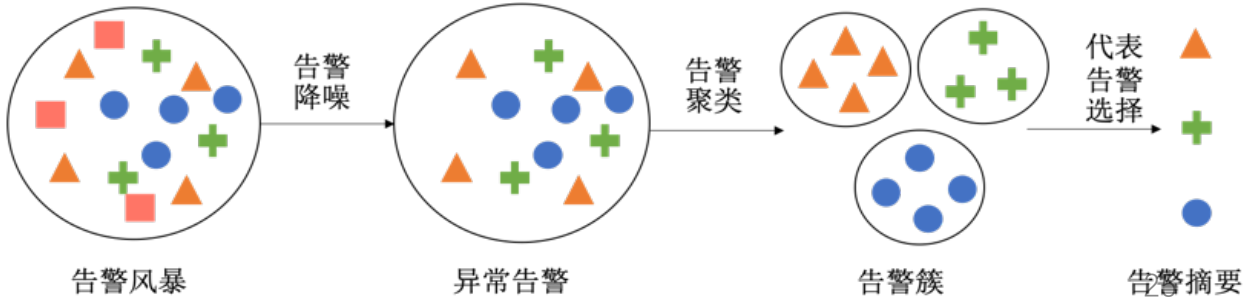
告警风暴的检测可以转化为突变点检测问题，使用极值理论（EVT）能够自适应地、准确地检测出告警风暴





② 告警风暴摘要

告警风暴摘要的目的是选取出与服务故障相关的告警集合，并且这些告警能够从多方面反映故障



方法

告警风暴摘要

0. Oracle database connection detection alerts, TNS: no listener program
1. Weblogic JDBC pool status is not running. The alert item is NBANK: EntServer
2. Syslog alert, lan24 on system SDII11P1 has gone down due to lost connection with the link partner
3. Memory check timeout. Description/Type table : No response from remote host ipaddr
4. CPU usage: Can't get necessary data
5. Disk load current value: 100.00% exceeding configured threshold 95.00%
6. Oracle log alerts. ORA-01034: Oracle not available
7. Database transfer status is abnormal. Alert item is NBANKDB_STB
8. Node ping alert. Host check timed out after 10 seconds
9. Source ipaddr ping destination ipaddr, response timeout. Average RTT is 106ms, threshold is 100ms
10. System VCS cluster alert. SDII13P1_vcs has faulted in cluster CL_NBANK_DBSRAC_SHD

1 基于学习的告警降噪

- 定义成异常检测问题
- 特征: 告警的属性和统计特征
- Isolation forest

2 告警聚类

- 相似性度量
  - 文本相似性: Jaccard distance
  - 拓扑相关性: graph path
- DBSCAN聚类

3 代表性告警选择

- 选择聚类中心

$$\text{centroid} = \arg\min_{i \in \text{cluster}} \frac{1}{n} \sum_{j=1}^{n} \text{similarity}(i, j)$$

评估

数据集：来自真实生产环境的166个告警风暴案例

检测

Datasets	A			B			C		
Methods	P	R	F1	P	R	F1	P	R	F1
EVT	0.92	0.96	0.94	0.90	0.97	0.93	0.95	0.96	0.95
Threshold	0.82	0.99	0.90	0.75	0.92	0.83	0.59	0.91	0.72

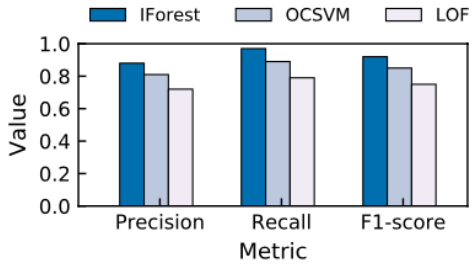
基于极值理论的告警风暴检测方法比固定阈值方法好

摘要

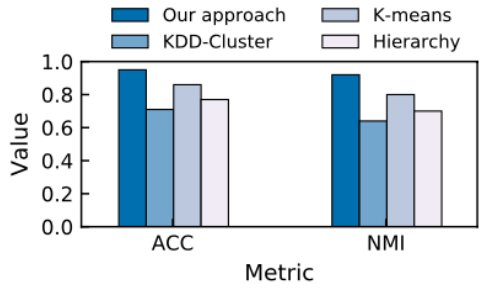
Datasets	Raw	Severity	Denoising	Summary
A	0%	88.7%	6.9%	98.8%
B	0%	85.6%	5.1%	98.2%
C	0%	84.1%	8.4%	99.1%

告警摘要算法能显著减少工程师需要查看的告警数量；且推荐的告警和故障相关的准确度可以达到75%

Method	Raw	Severity	W/o denoising	Summary
Precision	0.08	0.42	0.64	0.75



(a) Alert denoising



(b) Alert clustering

告警降噪和告警聚类模块表现得比现有方法都好

告警数据挖掘研究

故障发现

论文1：自适应的告警动态定级

Automatically and Adaptively Identifying Severe Alerts for Online Service Systems

INFOCOM 2020

故障诊断

论文2：告警风暴摘要

Understanding and Handling Alert Storm for Online Service Systems

ICSE SEIP 2020

故障预测

论文3：基于告警的事件预测

Real-time Incident Prediction for Online Service Systems

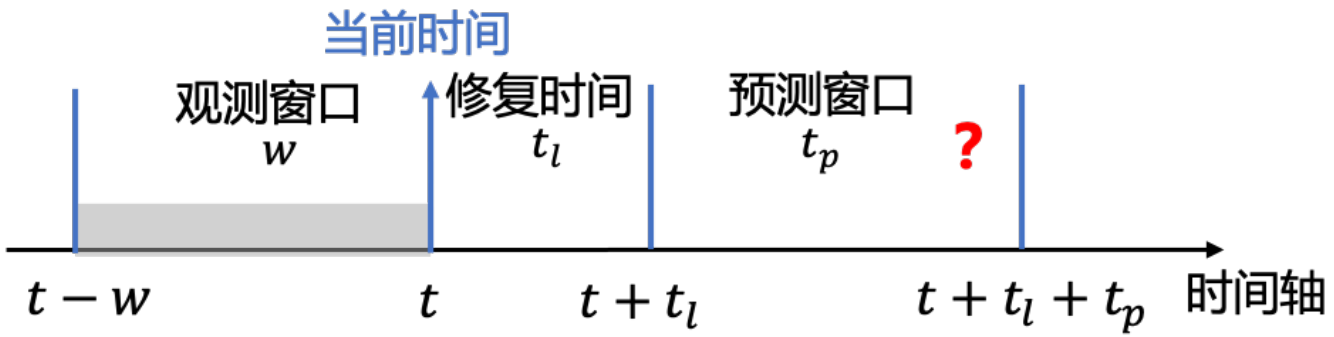
ESEC/FSE 2020

用历史低级别告警预测和故障相关的高级别告警，并提供可解释的预测分析

问题定义

事件预测问题定义为一个窗口分类问题

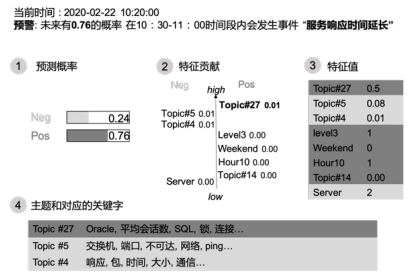
- 即在当前时刻t，我们拿到过去一段历史告警数据（观测窗口），基于这个窗口内的告警数据，预测在未来一段时间（预测窗口）内是否会出现某个事件
- 修复时间是留给工程师采取措施避免事件发生的时间



方法

- **特征提取：**文本特征（通过LDA主题模型提出文本的潜在语义特征）+统计特征（这个窗口内的告警数量，窗口的时间，告警间隔时间等其他的统计特征）

- **特征聚合：**在实际场景中，并不是所有的告警对事件预测是有帮助的，有征兆的告警可能被淹没在大量的噪声告警中，导致预测结果不好。为了避免噪声告警带来的影响，采用了多示例学习的思想，在聚合的过程中弱化噪音告警的权重，增强征兆告警的权重，以此避免噪声告警对预测产生的干扰作用。



特征解释：采用LIME对每次预测结果给出可解释分析，告诉工程师哪些特征对事件的发生起到的作用最大，那么这些特征极有可能和事件的根因有关，从而帮助工程师主动采取措施来规避故障。

告警数据挖掘研究总结

故障发现

论文1：自适应的告警动态定级

- 提出了基于排序模型的告警动态定级算法
- 基于多特征融合，设计并从多源数据中提取了多种特征，用于表征告警的严重性

故障诊断

论文2：告警风暴摘要

- 聚焦在告警风暴这一运维问题中，提出了用动态剧变点检测的方法准确检测告警风暴
- 提出了一套有效的告警风暴摘要算法框架：包括噪声告警过滤、告警聚类和典型告警选择

故障预测

论文3：基于告警的事件预测

- 采用多示例学习的思想解决噪声告警对预测的干扰，基于有监督的分类算法构建了一个事件预测模型，提供一个可解释的结果方便工程师主动采取措施去规避故障

必示科技诚邀您的加盟

- **国际顶尖研究成果的发表机会：**接触海量的真实运维数据（如机器监控数据、服务运行日志、业务监控指标等），设计和研究数据挖掘算法，来解决大型数据中心和在线业务场景下富有挑战的实际运维问题（如异常检测、根因定位、故障预测等）、产生业务价值、发表顶尖研究成果。
- **工业界真实应用的影响力：**指导实力强大的工程化团队将前沿的自研算法产品化，融入必示智能运维平台体系，在工业界产生巨大的真实影响力。
- **实力强劲的算法团队：**大牛带队，在学霸密度超高的公司与业内优秀团队共事，与来自纽约大学石溪分校、匹兹堡大学、清华大学的多位博士思维碰撞，迅速进步。
- 充分尊重优秀人才的个人价值，我们为你提供有**绝对争力的薪酬**、弹性工作、免费午餐&晚餐、无限零食&饮料、健身房、顶级办公设备。

招聘岗位：算法工程师、算法研究员 、算法研究实习生

简历投递：suikaixin@bizseer.com