基于NLP和聚类的服务器异常日志检测

李　岩[1]　龚向宇[2]
1. 深圳证券交易所　广东　深圳　518000
2. 红帽软件(北京)有限公司　北京　100000

摘　要：数据中心服务器多年下来累计了海量的日志，里面有很多的异常模式，如何用自然语言处理（NLP）的方法对这些数据进行管理和组织，并进行必要的分类，从而为后续的运维工作提供参考。我们方法的核心在于为每个日志文件创建一个异常字典，从而生成与该特定服务中异常的语义区域密切相关的特征矩阵。然后使用聚类算法通过模型训练优化模型参数，从而在模型稳定后对新来的日志进行聚类预测，以识别出异常的日志。

关键词：自然语言处理；聚类；日志分析；异常检测

引言

数据中心大量的系统服务产生了海量的日志，用于记录系统、程序运行中发生的各种事件，通过阅读日志，有助于诊断和解决系统故障，日志由服务创建并包含附加到扩展名为.log的文件的半结构化本。日志是增量生成的文本数据，不断反映事件及其对系统的影响，例如syslog事件记录系统活动。日志文件通常不包含相同类型、分类的信息，例如：syslog事件记录系统活动，crond日志保存了定时任务执行记录，以及virtlogd事件记录与虚拟化相关的操作。每个文件都倾向于描述整台服务器的局部视图。存储的信息可以包含：特定事件的时间和日期，以准确记录发生了什么；进程名称和进程标识符；机器主机名及其互联网协议地址。不同的服务可以使用不同的关键字来表达正常或错误的行为。

在本文中，首先清洗掉冗余的数据字段，然后进行数据结构化，以便将数据集组织成可管理的格式，生成的文件可用于识别有利于发现短期和长期数据中心管理的异常趋势和异常活动。我们方法的核心在于为每个日志文件创建一个异常字典，从而形成与该特定服务中异常的语义区域密切相关的特征矩阵。然后使用聚类算法通过模型训练优化模型参数，从而在模型稳定后对新来的日志进行聚类预测。

1 相关工作

一些研究提出了处理从手动操作到自动化操作的异常处理方法[1]，定义了管道，将日志文件转换为分析具可以理解的更易读的格式，例如.csv和.json，根据阈值对观察结果进行分类，并提取异常术语。

日志消息是字符串对象，因此NLP技术适合用以进行日志的预处理，所有与NLP相关的研究都包括从数据中提取相关信息的预处理阶段。霍文君等人[2]使用递归神经网络（长期短期记忆模型，LSTM) 来处理序列数据的能力。

2 数据集

我们使用深圳证券交易所的一些Linux服务器的日志，例如用户切换的日志sudo.log、开源邮件代理 postfix和系统日志syslog。日志文件包含一些数字信息（如返回值等）和描述系统状态和运行时间信息的文本数据，每个日志条目包括包含描述某些事件的自然语本（即单词列表）的消息。

> Jun 19 04:09:02 install-1-nf: logrotated abnormaly exited with [1]

这个是一个日志条目示例，没有先验知识很难理解，包括了文本和分隔符。文本通常由称为动态字段和静态字段，动态字段是运行时生成的记录；静态字段是描述服务的名称和标识等。字段由不同的分隔符所分隔，例如逗号、空格或括号。

3 方法概述

我们用ntlk.TokTokenizer先将文本文档转换为token-count矩阵，先使用 sklearn.CountVectorizer向量化处理以便之后进行的聚类。

3.1 数据预处理

通过特定的ToktokTokenizer将日志文件转换为.csv 文件，为每个服务过滤出一个特质的日志头结构。之后，csv 文件中的条目包含日志消息和一组可变变量，例如：（日期、时间戳）、主机名、（IP地址）、服务名称、进程ID、（组件） 名称。每个文件都与在数据中心的主机名上运行的特定服务相关，它的位置由本地数据库获取并包含在生成的文件中。

预处理步骤：

1）排除了数量太少无意义的服务日志文件；

2）忽略了大小写；

3）拼写错误纠正；

4）进程标识符包含在服务日志文件中；

5）通过正则表达式语法规则删除不必要的文本，如标点符号、非字母数字字符和任何其他不属于语言的字符；

6）停用意义较低的常见通用词（如 of、are、the、it、is）。对于停用词，我们决定保留否定词和其他可能涉及问题的词，例如 up、again、too、ok、 out、yet 等。

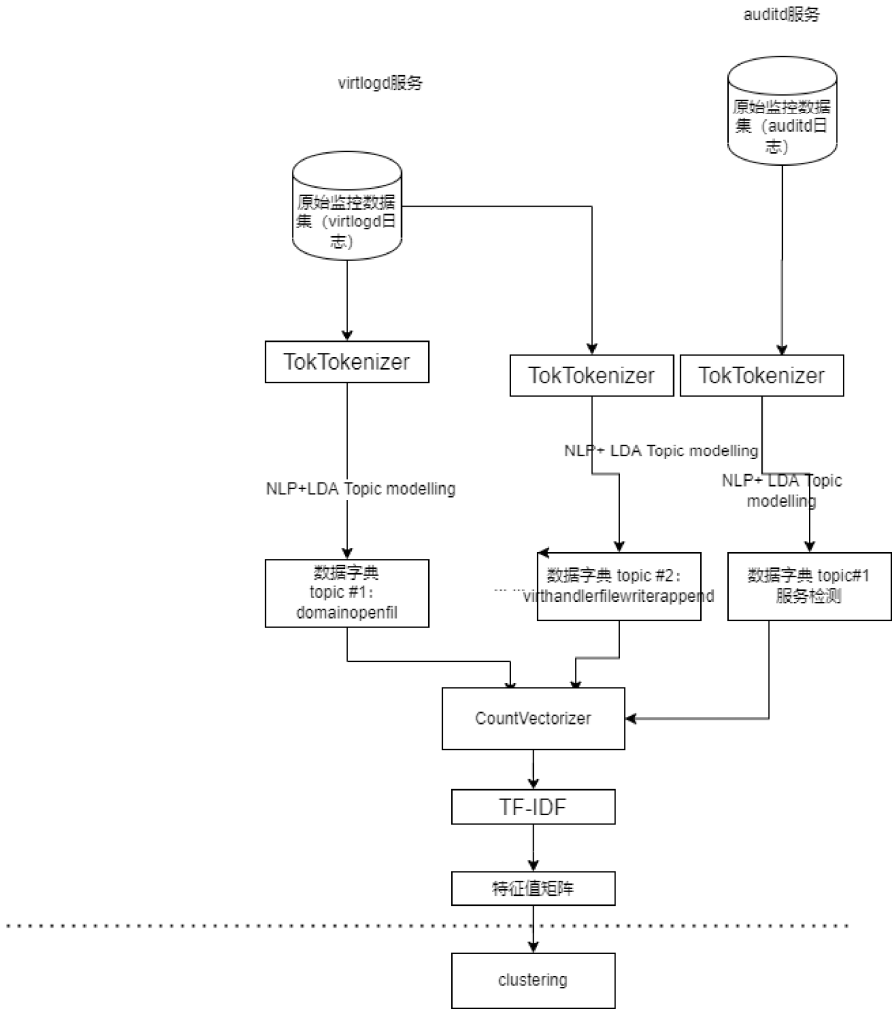

图1 算法总体流程（以服务virtlogd 和auditd为例）

3.2 异常字典的创建

在数据清理步骤之后，采用特征提取技术：Word2Vec 和 TF-IDF。其中，TF-IDF 通常用于在文档集合中查找重要单词，衡量单词与给定文档的关联；而IDF代表逆文档频率，代表一个词的重要性[3]。如果一个词出现在很多日志文件中，这个词的重要性就会降低。单独一种方法不一定能够为异常的识别提供有用的信息，因为对特征的搜索与对异常语义区域的模式搜索没有必然的联系。

我们生成的特征矩阵可能是高维度的，日志文件数量巨大的情况下，这个计算复杂度是很大的。每条消息都是由各种动态字段组成；因此，我们尽量最小化n-gram的维度，比如n <= 5来重建异常消息，省略动态字段， n值选择取决于具体的服务和topic。

3.3 字典的创建

字典的创建可能是基于先验知识来完成的，此外，特征矩阵可能具有高维度，因为所有unique n-gram消息将构成此类矩阵的一列。基于先验知识，通以下下的关键词都表示异常： abort，alert， cannot， can't，deprecate， disabled， error， exception， fail， fatal，unable，suspend， stop，suppress，invalid，impossible，huped， problem， unable， unsupported， warn。在这个阶段我们已经开始跟踪日志事件的类型，并识别异常关键术语，可用于分类服务中的问题原因。

3.4 主题建模LDA分析

主题建模是一种在大多数非结构化日志消息集合中找到潜在主题，并识别属于特定主题的语义结构和异常。我们应用了潜在狄利克雷分配（LDA），假设文档是主题的混合，而主题是单词的分布，通过最大化整体主题连贯性和平均主题重叠之间的差异来确定最佳主题数量，计算为主题之间“Jaccard 相似度”值的平均值。

3.5 特征矩阵

每个日志文件都可以根据异常字典生成一个特征矩阵。

Log 事件消息	Error count	Failure	Timed out
error count since last fsck	32	0	0
rsyslog.service main process exited	0	12	0
mount.nfs portmap query failed RPC Timed out	0	6	6

通过在 TF-IDF 函数的参数中更简单地在中指定这样的字典，可以生成一个与异常的语义严格连接的特征矩阵，并且能够针对新来的日志条目进行聚类，以分辨新日志是否包含这些异常模式。

4 聚类结果

表2 日志聚类举例

日期	时间	主机名	进程名	日志	聚类结果
2021-06-18	17:58：01	Ompweb-1-1-ft	kernel	page allocation failure. order:5， mode:0x0	15
2019-05-22	16:54:48	Backup-1-1-ft	Sshd	[12784]: fatal: monitor_read: unpermitted request 104	13
2020-05-13	23:47:02	Backup-1-2-nf	sshd	sshd[21215]: fatal: Access denied for user testpam by PAM account configuration	14
2020-09-06	11:20:48	Db-1-nf	sshd	sshd[7802]: error: PAM: Authentication failure for root from 192.168.181.21	14
2020-09-06	11:20:48	Db-1-nf	sshd	sshd[7802]: Failed keyboard-interactive/pam for root from 192.168.181.21	14

无监督聚类是深度学习中一种建模框架，无监督聚类只能够聚类成指定数量的类。sklearn中聚类算法实现有很多种，基于本问题的复杂特性，我们采用了相对比较合适的Mean-Shift算法进行聚类，结果如表2所示。

5 讨论

在本文中，我们讨论了数据提取，对服务器日志的自然语言处理解决方案和 Mean-shift聚类方法，在数据集中包括了对物理和虚拟资源的服务器日志的探索。我们之前也进行过启发式规则的研究[4]，更多的基于领域先验知识并结合服务器监控数据从而形成更准确的对异常的判断和分类是未来的研究方面。

参考文献

[1]程云观等，一种云环境下的高效异常检测策略研究，计算机应用与软件37（1），327-333

[2]霍文君，一种基于聚类的系统日志解析算法计算机科学与应用，2020，10(1)

[3]黄承慧等，一种利用TF-IDF方法结合词汇语义信息的文本相似度量方法研究，计算机学报，2011（5），856—864

[4]李岩， 吴智铭，一种混合 GA， SA 和启发式规则的 FMS 调度方法，上海交通大学学报 2001，33 (11)，1329-1332