基于文本向量化方法构建 IT 运维服务台事中处置参考模型①

陈晓伟, 曹逸峰, 尚鸿斌, 付 谦, 包妍苹, 沈 璟

(中国农业银行股份有限公司 数据中心, 上海 200131)

摘 要: 针对传统 IT 运维服务平台事件全文检索准确度不高及大量历史事件单再利用率低的特点, 提出一种基于事件文本向量化处理方法的事中处置参考模型. 通过建立生产运维特征词库, 将事件工单文本表示成特征向量, 并选择合适的算法对不同事件文本特征向量进行相似度计算, 最终通过匹配找到平台历史事件或者知识库中与现有事件相似度高的事件或知识, 供故障处置人员参考. 将此方法应用到企业级服务台事件处置流程中, 能帮助运维人员快速准确地找到类似的历史事件或者相关的知识, 形成事中处置参考, 加快事件处置效率, 同时也进一步提高了运维工具的智能化水平.

关键词: IT 运维; 服务平台; 事件文本; 向量化; 处置参考

Building the Disposal of Reference Model on IT Operational Service Desk Based on Text Vectorization Method

CHEN Xiao-Wei, CAO Yi-Feng, SHANG Hong-Bin, FU Qian, BAO Yan-Ping, SHEN Jing

(Agricultural Bank of China Data Center, Shanghai 200131, China)

Abstract: In view of the traditional IT operational service desk event full-text retrieval accuracy is not high and the low utilization rate of a large number of historical events list, this paper puts forward a disposal reference model based on event text vectorization method. Through the establishment of production operations of key library, the event text is represented into a feature vector, and selects the appropriate algorithm for different event text feature vector similarity calculation, eventually finds the historical events or the knowledge base by calculation with the existing high similarity of events, as a reference disposal for operations staff. Applying this method to the enterprise service desk incident disposal process, can help operations staff quickly and accurately to find the similar events in the history of or related to knowledge, to formin matter disposal of reference, to speed up the treatment efficiency, but also further improves the intelligent and automation level of the operational tools.

Key words: IT operations; service desk; event text; vectorization; the disposal of reference

随着各种运维标准的普及, 企业 IT 部门都已经建立起了面向流程管理的IT服务管理平台(简称服务台). 其中事件管理是服务台功能的重要组成部分, 在 IT 运维支持方面发挥了显著的作用. 服务台事件管理最主要的工作就是生产运维故障及服务请求事件单的流转, 不断重复建单、派单、转单、解决、关闭这一过程, 从而在日常运维中积累了大量事件工单. 这些大量的历史工单中有很多都是重复发生的事件, 很少有企业能将这些历史工单利用起来, 从而使运维人员大部分时间都耽搁在这种重复繁杂的工作中, 导致效率降低和资源浪费.

快速处置是事件管理的重要目标之一, 尤其对一线人员, 当接到故障报警时能够在建单的过程中就能快速地从历史信息中找到对当前事件有益的参考信息, 对事件的快速处置将起到很大的帮助. 因此, 建立一套能够快速准确地匹配历史事件或者知识库、问题库

① 收稿时间:2016-01-31;收到修改稿时间:2016-04-19 [doi: 10.15888/j.cnki.csa.005463]

的事中处置参考模型，将具有重要的实际意义和研究价值.

服务台中流转的事件工单，其本质是文本信息的载体，在大数据处理领域有很多文本数据自动处理的方法，因此，实现现有事件工单与历史信息的自动快速匹配关键就是实现工单数据的文本化并选择合适的文本相似度算法[1]. 目前国内自动化运维还处于起步阶段，将文本处理方法应用到运维领域的案例较少，尤其在事件管理方面，同业基本处于流程优化推广阶段，关于 ITIL 落地实施方面的积极探索屈指可数. 本文基于以上研究和现状，提出了一种利用事件文本向量化方法构建服务台事中处置参考的模型，实现事中处置过程中历史参考信息的自动推送.

1　服务台事中处置参考模型整体结构

传统的服务台中，事件管理流程主要负责事件工单的流转，日常运维工作中常常会遇到重复的事件不断进行建单、流转处置、解决关闭，事件解决关闭后，往往未能继续利用，仅仅限于后期的查阅. 日常繁杂性的工作占用了运维人员大部分时间，大大降低了工作效率. 本文在传统事件流转流程的基础上增加了向量化处理、相似度计算及推送参考环节形成了一种闭环流程结构，如图 1 所示.

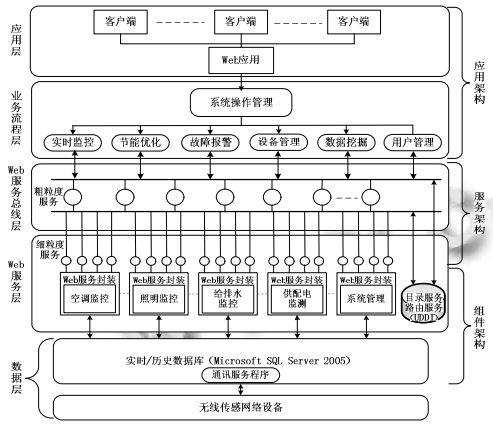

图 1　事中处置参考模型

这个闭环流程结构改变原有事件单一流程结构，将问题流程和知识流程也结合起来，统一形成一个闭环循环结构. 问题库和知识库是历史事件库的提炼和精华，在进行相似度计算时可优先从知识库、问题库抽取文本信息，若样本数量不足再从历史事件库中抽取计算，这样可进一步提高匹配速度和准确度.

当运维人员接到故障报警后，在建单输入故障信息时服务台会自动分析提交事件工单信息，表示成特征向量后与知识库、问题库或者历史事件库中文本进行相似度计算分析，将最可能贴近的处置方案推荐给支持人员，这将大大节省支持人员的时间. 当一个事件解决后，相关的事件单信息就会进入历史事件库，充实现有样本信息，有新的类似事件发生时通过事中处置参考模型匹配计算就可以将之前解决的类似事件的处置方案推送给处置人员. 大量类似重复的事件也可以归纳总结形成知识纳入知识库.

2　事件工单文本向量化

服务台生产运行事件工单中三个最主要的文本标签，分别是故障现象、原因分析和处置过程. 这三段文本信息基本能将一个事件信息概括完整，因此本文所指的文本向量化针对是这三段文本. 文本信息经过分词、过滤表示成特征词集的形式后就可以采用不同算法转换成向量化[2]形式.

2.1 特征词集

收集整理生产运维相关词汇做成特征词库，用此词库过滤事件工单文本中噪音词汇[3]. 下表为部分运维中常见故障的文本信息分词、过滤后形成的特征词集形式.

表 1　特征词集

序号	故障类型	故障现象	原因分析	处置过程
1	EMC 存储连接控制卡故障	EMC 存储 连接控制卡 LCC 指示灯...	报错 连接控制卡 故障 LCC 后端 磁盘...	检测 黄灯 整理 后面 连接线路 拔出...
2	EMC 存储存储硬盘故障	EMC 存储 硬盘黄色 指示灯 存储磁盘柜...	报错 存储 硬盘 故障 硬盘热备盘...	检测 硬盘 黄灯 Service Lan Port 口...
3	EMC 存储存储电源故障	EMC 存储 模块线路 连接处 led 灯...	报错 存储 电源 故障 电源设备...	检测 电源 指示灯 状态 备件...
4	EMC 存储存储电池故障	EMC 存储 模块线路 连接处 led 灯 灭灯	报错 存储 电池 故障 电池 ...	检查 电池 指示灯 状况 电池损坏...
5	P750 小型机 VRM 故障	IBM P750 小型机 HMC 登录界面...	报错 小型机 VRM 故障 更换 VRM...	维护 时间窗口 手动 更换 故障...
6	P750 power supply 故障	IBM P750 小型机 HMC 登录界面...	报错 power supply 故障 小型机...	维护 时间窗口 手动 更换...

7	P750 小型机 I/O 背板故障	IBM P750 小型机 HMC 登录界面 黄色 ...	报错 IBM 小型机 IO 背板故障 ...	维护 时间窗口 手动 更换 故障...
8	P750 小型机内置硬盘故障	IBM P750 小型机 HMC 登录界面 黄色...	报错 IBM power750 小型机 硬盘故障...	维护 时间窗口 手动 硬盘 ROOTVG 镜像...
...	...	...	...	...

2.2 向量化表示

以小型机内置硬盘故障为例演示一个简单文本向量化的过程:

2.2.1 示例事件文本特征词抽取

抽取事件“小型机内置硬盘故障”的故障现象文本, 表示成特征词集形式. 即“IBM P750 小型机 HMC 的登录界面有黄色告警, 故障的 FRU 号是 Hdisk”可以表示成 *dt*={IBM, P750, 小型机, HMC, 登录界面, 黄色, 告警, 故障, FRU 号, Hdisk}

2.2.2 已有故障库词集化表示

将表 1 特征词集中已归类的事件类别的故障现象文本也表示成特征词集形式. 比如表中 EMC 连接控制卡故障的故障现象可表示成 *d1*={EMC, 存储, 连接控制卡, LCC, 指示灯, 黄灯, 警报}. 同理小型机 VRM 故障可表示成 *d*5={IBM, P750, 小型机, HMC, 登录界面, 黄色, 告警, FRU 号, VRM, Parts Number, 46K6300}, 同列的内置硬盘故障可表示成 *d*8={IBM, P750, 小型机, HMC, 登录界面, 黄色, 告警, FRU 号, Hdisk}.

2.2.3 向量化计算

dt 与 *d*1、*d*2...*dn* 进行向量化表示, 下面以 *dt* 与 *d*8 进行计算为例介绍其过程:

① *dt*={IBM, P750, 小型机, HMC, 登录界面, 黄色, 告警, 故障, FRU 号, Hdisk}

*d*8={IBM, P750, 小型机, HMC, 登录界面, 黄色, 告警, 故障, 备件, FRU 号, Hdisk}

② 将特征词转换成数值. 数值可用概率来表示, 概率计算包括两方面, 特征单词在文本中出现的频率 *p*(该词在所属文本中出现的词频除以全部文本的特征词数), 还有该词的反文本频率 *q*[4](表示该词出现在多少个文本中的频率, 如果一个单词在很多文本中出现的频率都很高, 那么这个单词就太普遍了, 不足以用来表征一篇文档). 那么某一维的表征概率值就是 p 和 *q* 的一个因式乘积, 当然根据需要还有可能乘上另外一些影响因子.

③ 计算 *p* 和 *q*:

首先统计 *dt* 和 *d*8 文本中每个词出现的频率如下表:

表 2 词频统计

	IBM	P750	小型机	HMC	登录界面	黄色	告警	故障	备件	FRU 号	Hdisk	总词
dt	1	1	1	1	1	1	1	1	0	1	1	10
d8	1	1	1	1	1	1	1	1	1	1	1	11
总词	2	2	2	2	2	2	2	2	1	2	2	21

计算 *p* 值: 以 IBM 为例其在 *dt* 中对应的 *p* 值其在 dt 中出现的次数除以总词数即 1/21=0.048.

表 3 *p* 值计算

P 值	IBM	P750	小型机	HMC	登录界面	黄色	告警	故障	备件	FRU 号	Hdisk
dt	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0	0.048	0.048
d8	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048

q 的计算可利用公式 $ln((1+|D|)/|Dt|)$[5](众多算法中的一种), 其中其中$|D|$表示文本总数, $|Dt|$表示包含特征词 t 的文本数量. 以 IBM 为例其 *q* 值为 $ln(3/2)$=0.4.

表 4 *q* 值计算

q 值	IBM	P750	小型机	HMC	登录界面	黄色	告警	故障	备件	FRU 号	Hdisk
ln	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	1.1	0.4	0.4

计算 p*q 值: 即 p 值和 q 值相乘.

表 5 *p***q* 值计算

*p***q* 值	IBM	P750	小型机	HMC	登录界面	黄色	告警	故障	备件	FRU 号	Hdisk
dt	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0	0.019	0.019
d8	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.053	0.019	0.019

④ 表示成向量. 最终就可以将 dt 和 d8 表示成如下的空间向量:

dt={0.019,0.019,0.019,0.019,0.019,0.019,0.019,0.019,0.000,0.019,0.019}

*d*8={0.019,0.019,0.019,0.019,0.019,0.019,0.019,0.019,0.053,0.019,0.019}

其中向量对应的维度特征变量[6]为: IBM, P750, 小型机, HMC, 登录界面, 黄色, 告警, 故障, 备件, FRU 号, Hdisk.

3 事件文本相似度计算及参考信息推送

事件工单文本向量化后就可以采用合适的算法进行向量的相似度计算, 从而计算出新工单文本与历史库、问题库或者知识库中已有数据的接近程度, 最终将最可能的故障原因或最佳处置结果推送给运维处置人员.

3.1 事件工单文本相似度计算

向量间相似度计算. 这里可以采用数据挖掘中的分类和聚类算法, 常见有 KNN(最近邻)[7]、贝叶斯[8]、VSM(向量空间模型)[9]等. 为了方便计算我们采用最简单的余弦相似性公式[10]进行计算:

$$Sim=\cos(\theta)\frac{\sum_{i=1}^{n}Ai\times Bi}{\sqrt{(\sum_{i=1}^{n}A_i^2)(\sum_{i=1}^{n}B_i^2)}} \quad (1)$$

计算得到 *dt* 与 *d*8 的夹角余弦值为 0.75, 同样方法计算 *dt* 与 *d*5 的夹角余弦值为 0.42, *dt* 与 *d*1 的夹角余弦值为 0, 这个余弦值越接近 1 表明夹角越小也就越相似. 同理文本较多时可算出 *dt* 与其余文本夹角余弦值, 将值按大小排列就可以分出相似度高低.

3.2 事中处置参考信息推送

事中处置参考信息包括故障原因和处置过程两方面. 通过模型在实际生产环境中的部署, 实现实时动态的数据分析和推送.

3.2.1 最可能故障原因参考

当只知道故障的异常现象, 通过向量化表示计算找出与此异常现象最接近的一个或几个已知故障类别, 已知故障的原因甚至处置过程就可供故障处置人员参考.

待计算文本 *dt* 与现有类别文本 *d*1、*d*2...*dn* 进行向量化表示并依次进行向量的相似度计算, 相似度值按大小顺序排列, 当计算文本为故障现象时, 如上述小型机内置硬盘故障一样 *dt* 与 *d*8 的相似度值最高, 因此可把 *d*8 的原因分析当做 *dt* 最可能的故障原因.

3.2.2 最佳处置过程参考

当同时知道故障的异常现象和原因时, 将异常现象和原因分析都进行向量化表示并计算, 能更精确地确认故障类型, 从而从已知故障中找到最佳处置参考.

以 XX 型号小型机宕机切换异常导致 XX 系统交易受影响的故障事件为例, 通过故障现象、原因分析文本计算得到与此故障最接近的已知故障是 YY 小型机 HA 切换失败处置案例, 因此可以将 YY 小型机 HA 切换失败处置案例的处置过程作为本次故障的处置参考. 若只通过故障现象来计算, 由于两次影响的业务系统名称及设备型号都不一样, 计算结果并不一定精确, 但从原因分析上看都是由小型机切换异常导致, 因此加上原因分析的相似度计算就可以确定这两个故障最相似.

3.2.3 模型部署及实际效果

事中处置参考模型可以作为企业运维服务平台的一个挂载模块, 数据输出集中在事件(或故障)处理流程部分, 数据采集主要集中在事件、问题和知识流程部分. 当故障异常发生时, 运维人员在运维服务平台建立事件工单, 输入初步的异常现象信息, 系统模型根据已填信息计算推送此次故障可能原因, 运维人员结合推送信息和经验快速判断并准确转单给合适的处置人员, 处置人员在处理工单时也能获得系统模型推送的处置参考信息, 从而促进故障异常的快速定位跟及时解决.

事中处置参考模型将事件管理流程与现有的知识管理流程和问题管理流程进行更紧密结合, 大幅提高知识库、问题库的利用率, 有效提升企业运维服务平台事件处置效率及企业自动化运维水平.

4 结语

本文将数据处理相关技术运用到运维自动化领域, 通过引入文本特征向量表示及文本相似度计算技术, 提出了一种事中处置参考模型, 并阐述了具体的实现过程. 其中文本词集向量化和相似度算法的选择是工作的重点和难点. 事件工单文本向量化的前提就是过滤无用噪音词汇然后把文本表示成词集形式, 但在实际中可能存在工作人员工单填写不规范, 企业的特征词库数据不完整等现象, 这样就会造成词集信息不能有效代表文本含义导致最终计算的偏差, 因此正确地进行词集化表示是所有工作的前提. 另外需要选择合适的相似度算法, 很多算法最开始计算的相似度偏差很大, 不能满足实际需要, 这就要通过不断调整参数校正改进来提高计算的准确度, 最终才能实现推送的高准确率.

事中处置参考模型为事件管理提供有力的支撑, 将大部分运维人员从繁琐的日常工作中解放出来, 减少事件处置时间. 随着运维自动化水平的不断提高,

这种自动化的处置参考模型将有广阔的应用前景，尤其是在突发事件的事中控制方面，能提供很好的决策支持，提高事件处置效率.

参考文献

1 曹逸峰,陈晓伟.基于知识分层提取模型的服务台知识库建设.计算机系统应用,2015,22(2):261–265.

2 董奥根,刘茂福,黄革新,等.基于向量空间模型的知识点与试题自动关联方法.计算机与现代化,2015,(10):6–9.

3 陈海燕.基于搜索引擎的词汇语义相似度计算方法. 计算机科学,2015,(1):261–267.

4 周戈.一种基于反向文本频率互信息的文本挖掘算法研究.计算机应用研究,2012,29(2):487–489.

5 刘千仞.文档分类技术的研究与应用[硕士学位论文].北京:北京邮电大学,2014.

6 胡建鹏,陈强,黄容.逐步贝叶斯判别分析中的变量优化方法研究.计算机工程与应用,2014,(21):63–67.

7 戴上平,冯鹏,刘盛英杰,等.基于余弦距离的局部敏感哈希的 KNN 算法在中文文本上的快速分类.计算机工程与科学,2015,37(10):1971–1976.

8 张春,郭明亮.大数据环境下朴素贝叶斯分类算法的改进与实现.北京交通大学学报:自然科学版,2015,39(2):35–41.

9 叶施仁,严水歌,杨长春.基于 VSM 和 LSA 的微博搜索排序方法研究.情报科学,2015,(7).

10 谢翠萍,陈家益,白金山.基于全文索引与余弦公式医学文本相似性分析.微型电脑应用,2014,30(1):25–27.