Splunk 学术发展分析报告

Analysis on Splunk of Academic Development

方建勇[1]（余姚，浙江 315400）

摘要：Splunk 公司成立于 2004 年，2012 年在纳斯达克上市，是大数据时代专注于机器日志的数据分析公司，重点发力数据安全和异常监控领域，友好的可视化界面设计为大数据处理降低了以前只有专业工程师才能介入的门槛。通过 Splunk 大数据分析软件，管理人员可以通过简洁明了的可视化显示结果，对大数据分析结果进行快速解读，并可以据此作出实际管理上的适当反应，为企业级的数据安全和异常处理提供了良好的指引。

关键词：Splunk 大数据 可视化 数据安全 异常发现 机器日志

Abstract: Splunk was founded in 2004 and listed on the Nasdaq in 2012. It is a data analysis company focused on the machine log in the era of large data, focusing on the field of data security and abnormal monitoring, and the friendly visual interface design for large data processing Previously only professional engineers can intervene in the threshold. With Splunk's large data analysis software, managers can quickly interpret the results of large data analysis through a concise, visual display of the results, and can react appropriately to the actual management, providing enterprise-class data security and exception handling Good guidance.

Key words: Splunk; Large data; Visualization; Data security; Anomaly discovery; Machine log

Splunk 公司成立于 2004 年，2012 年在纳斯达克上市，是大数据时代专注于机器日志的数据分析公司，重点发力数据安全和异常监控领域，友好的可视化界面设计为大数据处理降低了以前只有专业工程师才能介入的门槛。通过 Splunk 大数据分析软件，管理人员可以通过简洁明了的可视化显示结果，对大数据分析结果进行快速解读，并可以据此作出实际管理上的适当反应，为企业级的数据安

[1] 方建勇，男，1978 年-，中国工业与应用数学学会会员，中国物流学会会员，资深 IT 项目经理，资深 IT 工程师，资深 DBA（大型数据库管理员），浙江大学历史系硕士研究生学历，浙江大学数学与应用数学专业本科毕业，理学学士学位。

全和异常处理提供了良好的指引。

一、Splunk 学术发展趋势

超星发现系统收录的 Splunk 历年发表的外文学术成果，总量为 1587 篇，其中包括图书 10 册、期刊 915 篇、学位论文 5 篇、会议论文 11 篇、专利 640 项、科技成果 5 篇、信息资讯 1 篇。

按发表的时间段来看，2017 年 48 篇[2]、2016 年 207 篇、2015 年 347 篇、2014 年 342 篇、2013 年 295 篇、2012 年 155 篇、2011 年 53 篇、2010 年 16 篇、2000-2009 年 94 篇、1990-1999 年 1 篇、1989 年以前 29 篇。美国 Splunk 公司成立于 2004 年，那我们重点关注 2004 年以后的统计数字，依次为基准，关于 Splunk 最早的期刊论文是 2005 年发布 6 篇，随后每年都有期刊论文发表，但是数量维持在每年 30 篇以下，2012 年达到 92 篇，2013 年达到单年峰值的 192 篇，随后每年发表的期刊论文一直维持在 130 篇以上，2017 年截至发稿时发表的单年期刊论文已经达到 182 篇，按照这个趋势，应该会超过 2013 年峰值的 192 篇；2006 年开始有专利申请，当年 3 项，2007 年 12 项，2008 年、2009 年、2010 年申请的专利数量都为 0，估计与 2008 年全球经济危机有关系，到了 2010 年，申请的专利数量为 20 项，随后数量逐年上升，2015 年达到单年峰值的 214 项，2016 年 64 项，2017 年截至发稿时为 48 项，有所减少。

Splunk-各类型学术发展趋势曲线							
序号	年份	图书(数量)	期刊(数量)	学位学术成果(数量)	会议学术成果(数量)	专利(数量)	科技成果(数量)
1	1994	0	1	0	0	0	0
2	1995	0	0	0	0	0	0
3	1996	0	0	0	0	0	0
4	1997	0	0	0	0	0	0
5	1998	0	0	0	0	0	0
6	1999	0	0	0	0	0	0
7	2000	0	0	0	0	2	0
8	2001	0	0	0	0	0	0

[2] 为方便起见，这里单位统一为篇。

9	2002	0	0	0	0	0	0
10	2003	0	0	0	0	0	0
11	2004	0	0	0	0	0	0
12	2005	0	6	0	0	0	0
13	2006	0	21	0	0	3	0
14	2007	0	9	0	0	12	0
15	2008	2	21	0	0	0	0
16	2009	0	18	0	0	0	0
17	2010	1	14	0	1	0	0
18	2011	0	32	0	1	20	0
19	2012	0	92	0	1	61	1
20	2013	5	192	0	0	98	0
21	2014	0	189	1	4	148	0
22	2015	0	130	0	3	214	0
23	2016	1	140	1	0	64	0
24	2017	0	182	0	0	48	0

二、Splunk 学术成果统计[3]

1、关键词

关键词涉及 search engines(4[4])、awards & honors(2)、computer industry(2)、product introduction(2)、new products(2)、it(2)、IBm(2)、google(2)、Microsoft(2)、security(2)、more like this(1)、web sites(1)、congresses & conventions(1)、book reviews(1)、corporate profiles(1)、european union(1)、business growth(1)、children's literature(1)、computer systems(1)、computer peripherals(1)、computer software industry(1)、computer input-output equipment(1)、electronic systems(1)、energy industry(1)、medical equipment(1)、vendors(1)、ford motor co.(1)、technological planning(1)、sales management(1)、electronic information resources(1)、MATLAB(1)、computer centers(1)、NETWORK(1)、apache(1)、ebaY(1)、Instruments(1)、AUTOMATION(1)、CSA(1)、information(1)、APPLE(1)、mit(1)等，最多的关键词 search engines 也仅出现 4 次，大多是出现 2 次与 1 次的关键词，说明有关 Splunk 研究的集中度不高。

[3] 数据来源于超星发现系统。
[4] 括号内数字为出现频次，下同。

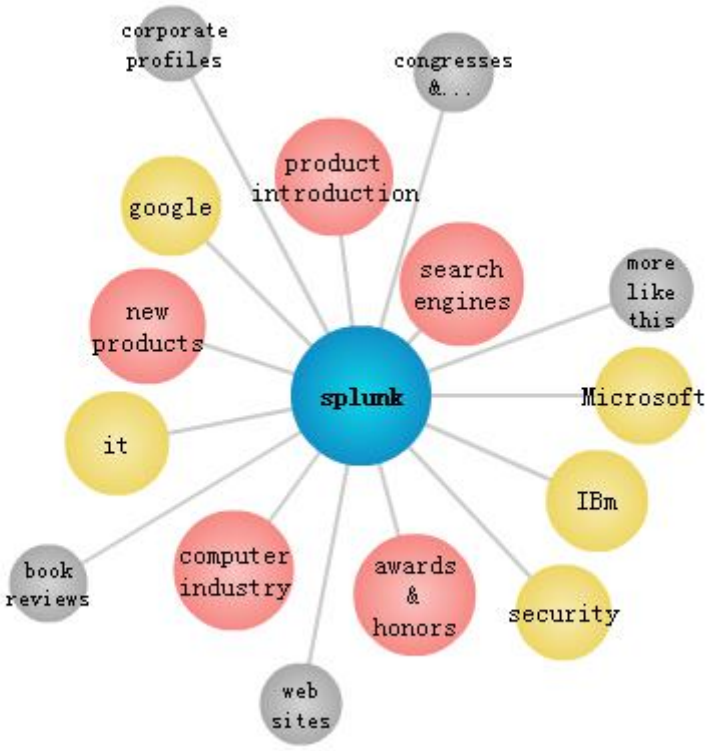

图 1 Splunk 关键词频次泡型图

2、学科、作者、机构和刊种分布

学科分布依次为 Technology(17)、Social sciences(4)、Science(2)、others(844)等，最多的学科分布在 Technology，出现 17 次，最多集中在 2 次或 1 次频率的学科，学科集中度也较分散。

发表作者依次为 John.(1)、Bunge, Jacob(1)、Elisa Bertino(1)、Ovide, Shira(1)、Ben Worthen(1)、Yi-Ming Chen(1)、Hassan, AE(1)、Tarek El-Ghazawi(1)等，都只有 1 篇在手，作者的集中度也是分散的。

作者所在的机构依次为 New Mexico(3)、Sandia National Laboratories(3)、Georgetown University(1)、Dartmouth University(1)、Dartmouth College(1) 等，最多的机构 New Mexico 和 Sandia National Laboratories 都为出现 3 次，与作者的分散性是一致的。

发表的刊物分布依次为 Business Wire (English)(351)、PR Newswire US(110)、Investors Business Daily(96)、Network World(21)、eWeek(17)、Wireless News(16)、Fair Disclosure Wire (Quarterly Earnings Reports)(16)、Wall Street Journal(Eastern

Edition)(15)、 Forbes.com(9)、 CIO: Chief information officer(7)、 InfoWorld(7)、Computer Weekly News(7)、International Business Times(6)、The Journal of Allergy and Clinical Immunology: In Practice(5)、Information Week(4)、M2PressWIRE(4)、Computer Weekly(3)、Network Computing(3)、Productivity Software(2)、Computer Security Update(2)、Revue Francaise d'Ornithologie Scintifique et Pratique(2)、PR Newswire(2)、Baseline(2)、Physical Review Special Topics: Accelerators and Beams(2)、American Banker(2)、Windows IT Pro(2)、Hill(2)、The New York Times(2)、Australian, The(2)、Scholastic Parent and Child(1)、Horn Book Magazine(1)、In the Black(1)等，排名前三的刊物 Business Wire (English)(351)、 PR Newswire US(110)、 Investors Business Daily(96)占了 35.1%，相对来说很集中。

三、参考文献

[1] 超星发现系统[EB/OL].http://www.chaoxing.com/

[2] 美国知识产权局[EB/OL].https://www.uspto.gov/

四、附录

Splunk 公司专利

序号 专利号 标题

1 9,715,329 提供服务的云网络

2 9,699,205 网络安全系统

3 9,667,641 计算机网络数据的复杂事件处理

4 9,667,640 根据在查询处理系统中从搜索结果获取的信息自动生成警报

5 9,660,930 动态数据服务器节点

6 9,648,037 使用访问模式和域名注册的安全威胁检测

7 9,646,398 最小化模糊操作，为图像创建模糊效果

8 9,645,975 通用数据中实数的近似订单统计

9 9,614,736 使用从机器数据派生的关键性能指标，沿着基于时间的图形通道定义图形可视化

10 9,609,011 具有可选择的交互视图的界面，用于评估潜在的网络妥协

11 9,609,009 用户/用户实体行为分析的网络安全威胁检测

12 9,607,414 三维点对多边形操作，以方便显示三维结构

13 9,596,254 机床数据处理平台数据采集阶段的小型图

捕获捕获网络数据的触发器 14 9,596,253

15 9,596,252 使用事件组摘要识别可能的安全威胁

16 9,596,146 将从机器数据导出的关键性能指标映射到仪表板模板

17 9,594,828 对非结构化数据的文本记录执行结构化查询

时间序列搜索主要和次要记忆

19 9,594,545 用于显示组件实例之间的通知依赖关系的系统

20 9,591,010 双路径分布式架构，用于网络安全分析

21 9,590,877 服务监控界面

22 9,589,012 应用于对象查询的数据模型生成

23 9,584,374 使用从机器数据导出的聚合关键性能指标监控整体服务级别性能

24 9,582,585 发现字段以过滤响应搜索返回的数据

25 9,582,557 用过程选择进行规则创建的抽样事件

26 9,521,047 具有每个实体状态的机器数据导出的关键性能指标

27 9,516,053 用户/用户实体行为分析的网络安全威胁检测

28 9,516,052 网络安全调查事件的时间线显示

29 9,516,046 分析从机器数据事件中提取的一组相对于这些值的人口统计量的值

30 9,516,029 根据用户角色搜索索引数据

31 9,514,189 处理包括外部数据源的系统搜索请求

32 9,514,175 标准化事件数据的时间戳

33 9,514,021 移动应用性能测量系统

34 9,509,765 来自多个搜索对等体的消息的异步处理

35 9,497,199 存储在云数据存储中的事件数据的访问控制

36 9,495,187 从虚拟机管理程序环境的架构和性能的自上而下的介绍

37 9,491,059 IT 服务拓扑导航仪

38 9,471,362 将虚拟机的虚拟机管理程序数据与关联的操作系统数据相关联

39 9,442,981 使用图形用户界面预览解析的原始数据

40 9,442,789 识别机床数据的异常情况

41 9,437,022 基于时间的可视化场景的各种值的事件数

42 9,432,396 使用域名注册的安全威胁检测

43 9,430,574 显示事件字段的许多唯一值

44 9,430,488 文件更新跟踪

45 9,426,172 使用域名访问的安全威胁检测

46 9,426,045 具有严重性状态排序的主动监视树

47 9,419,870 具有状态分配环的主动监视树

48 9,417,774 具有节点固定用于并发节点比较的主动监视树

49 9,384,261 自动创建用于识别机器数据中的事件边界的规则

50 9,363,149 网络安全调查的管理控制台

51 9,361,357 使用字段和关键字标准搜索从机器数据导出的事件

52 9,356,934 存储索引数据的数据量缩放

53 9,355,006 测量在移动设备上运行的应用程序的用户满意度

54 9,323,557 在一段时间内基于相关子组件的性能状态确定虚拟机环境中父组件的性能状态

55 9,317,582 识别与机床数据相匹配的与机床数据特定部分匹配的事件

56 9,298,805 使用提取来搜索从机器数据导出的事件

57 9,294,361 使用关键性能指标（KPI）相关搜索来监控服务级别的性能

58 9,292,590 根据第一个事件的提取部分识别从机器数据导出的事件

59 9,286,413 使用从机器数据派生的关键性能指标来呈现服务监控仪表板

60 9,280,594 从不同来源的机器数据导出的事件的统一存储和搜索

61 9,276,946 将安全相关事件列入黑名单

62 9,275,338 机器数据事件模式的预测分析

63 9,256,501 高可用性调度程序，用于调度 map-reduce 搜索

64 9,251,221 根据搜索查询结果为对象分配分数

65 9,248,068 新注册域名的安全威胁检测

66 9,245,057 使用从机器数据派生的关键性能指标，沿着基于时间的图形通道呈现图形可视化

67 9,245,039 跨多个搜索会话的事件记录跟踪

68 9,229,985 使用动态指针绑定特征的中央注册表

69 9,225,724 弹性资源缩放

70 9,215,240 从大数据中的事件调查和动态检测潜在的安全威胁指标

71 9,210,056 业务监控接口

72 9,208,463 从机床数据得出的主要性能指标的阈值

73 9,208,206 选择基于数据分析的解析规则

74 9,208,000 根据应用事件计算计算机应用的质量指标

75 9,185,007 具有严重性状态排序的主动监控树

76 9,177,002 使用分布式索引器系统中的中间结果报告加速度，用于搜索事件

77 9,173,801 基于访问新注册域的指示，图形显示安全威胁

78 9,164,786 根据一段时间内相关子组件的性能状态确定虚拟机环境中父组件的性能状态

79 9,160,798 高可用性和灾难恢复的集群

80 9,158,811 事件审查界面

81 9,152,929 实时显示所选正则表达式的统计信息和值

82 9,152,682 作为一系列命令对表中的列进行跟踪元数据在表上运行

83 9,146,962 使用信息字段识别事件

84 9,146,954 从搜索结果集创建实体定义

85 9,142,049 主动监控树提供分支叠加的分布流图

86 9,130 ,971 基于站点的搜索关联

87 9,130 ,860 使用从机床数据导出的关键性能指标监控服务级别性能

88 9,130 ,832 从文件创建实体定义

89 9,129,041 更新有助于评估定性搜索词的上下文的技术

90 9,129,028 事件字段分布式搜索显示

91 9,128,995 使用从机器数据派生的关键性能指标，沿着基于时间的图形通道定义图形可视化

92 9,128,985 补充高性能分析商店，评估各种事件以响应事件查询

93 9,128,980 应用于查询的数据模型生成

94 9,128,916 机床数据网

95 9,128,779 用于检索补充工作信息的分布式任务

96 9,124,612 多站点聚类

97 9,122,746 对非结构化数据执行结构化查询

98 9,087,090 促进包含定性搜索术语的概念查询的执行

99 9,055,075 项目资源访问控制

100 9,052,938 虚拟机数据和存储性能数据的相关和相关显示

101 9,047,246 高可用性调度程序

102 9,047,181 集群数据的可视化

103 9,043,717 机器数据事件的多通道时间同步可视化

104 9,043,332 集群性能监控

105 9,037,562 数据量的弹性缩放

106 9,036,979 根据名称信息确定媒体内容的位置

107 9,031,955 用于开发用于事件搜索的字段的字段提取规则的事件抽样

108 9,015,716 具有节点固定用于并发节点比较的主动监视树

109 9,009,539 识别和分组程序运行时错误

110 9,002,854 内插时间戳的时间序列搜索

111 8,990,637 计算和访问计算机应用的质量指标

112 8,990,245 确定和显示为分布式数据存储中的事件定义的字段的唯一值的数量

113 8,990,184 时间序列搜索引擎

114 8,983,994 生成用于搜索机器数据的数据模型

115 8,978,036 从外部来源收集和处理数据的任务的动态调度

116 8,977,638 文件识别管理和跟踪

117 8,972,992 具有状态分配环的主动监视树

118 8,943,056 机床数据网

119 8,909,642 根据样本事件中的选择自动生成字段提取规则

120 8,904,389 根据相关子组件的性能状态确定虚拟机环境中组件的性能状态

121 8,874,755 为服务提供云网络

122 8,849,779 数据量的弹性缩放

123 8,826,434 基于大数据访问新注册域的指示进行安全威胁检测

124 8,825,664 索引预览

125 8,806,361 机器数据事件的多通道时间同步可视化

126 8,793,225 处理包括外部数据源和混合模式的系统搜索请求

127 8,788,526 用于语义搜索的机器数据的数据模型

128 8,788,525 用于语义搜索的机器数据的数据模型

129 8,788,459 高可用性和灾难恢复的聚类

130 8,756,614 使用动态指针绑定特征的中央注册表

131 8,756,593 用于表示由动态指针伪造的应用程序功能之间的相互关系的地图生成器

132 8,756,262 通用数据中实数的近似订单统计

133 8,752,178 将安全相关事件列入黑名单

134 8,751,963 正则表达式之前提取的数据字段的实时指示

135 8,751,529 分布式数据的可扩展交互式显示

136 8,751,499 资源限制下的变量代表抽样

137 8,751,486 对非结构化数据执行结构化查询

138 8,745,109 通用数据中实数的近似订单统计

139 8,738,629 外部结果提供用于检索使用不同配置或协议存储的数据的进程

140 8,738,587 通过从本机索引和虚拟索引检索结果来处理系统搜索请求

141 8,694,450 机床数据网

142 8,683,467 根据相关子组件的性能状态确定虚拟机环境中父组件的性能状态

143 8,682,930 数据量管理

144 8,682,925 分布式高性能分析商店

145 8,682,906 基于手动编辑正则表达式实时显示数据字段值

146 8,682,886 使用中间事件摘要报告加速

147 8,682,860 数据量管理

148 8,589,876 检测影响动态指针和应用程序功能依赖关系的中心注册表事件

149 8,589,432 实时搜索和报告

150 8,589,403 用于元数据恢复和复制的事件跟踪文件中的压缩日记

151 8,589,375 实时搜索和报告

152 8,589,321 机床数据网

153 8,589,304 控制网络设备之间音量索引的系统和方法

154 8,583,631 流水线搜索语言的元数据跟踪（字段的数据建模）

155 8,566,336 文件识别管理和跟踪

156 8,548,961 快速文件跟踪和更改监控的系统和方法

157 8,516,008 灵活模式列存储

158 8,515,963 索引预览

159 8,412,696 实时搜索和报告

160 8,112,425 时间序列搜索引擎

161 7,937,344 机床数据网