挑战二的可视化方案特点分析

田甜，陈璐，刘松，汪鑫，刘俊荣, 中国科学院信息工程研究所

摘要—本文针对 2016ChinaVis 挑战赛第二题的内容，构建了一套基于 HackingTeam 泄露邮件的可视分析系统。对于第一个问题主要采用散点图多维度展示员工的重要程度，并利用多视图协同的方法验证结果的正确性。对于第二个问题，基于 Hacking Team 公司的主要业务对邮件进行分类，采用 sunburst 图展示业务相关和生活相关邮件等。第三个问题，从多角度以可视化方式推断公司的主要发展阶段，以字符云为主展示公司的主要发展阶段以及各阶段的主要业务，同时利用折线图对业务量进行可视化，验证发展阶段界定的正确性。

关键词— Hacking Team,D3,Echarts,Gephi,Sunburst,字符云,散点图

简介

本次比赛的可视化成果实现主要分为三个阶段，第一阶段是对原始数据预处理及有效特征提取，第二阶段是制定可视化展示及分析策略，实现高价值数据的可视化，第三阶段是对可视化的结果做分析验证。

本文的第一部分介绍主要可视化方案，并对所用到的可视化展示方案做详细的说明。首先提取Hacking Team公司内部员工列表，利用散点图和概念图等从重要程度，工作职责对员工分类。其次，对邮件主题进行聚类分析，提取公司的主营业务，采用Sunburst图对泄露邮件进行分类。最后采用可视化方法推断公司的主要发展阶段，进一步采用字符云展示每个阶段的主要业务。

第二部分从可视化展示方案的实用性、交互性的角度来分别阐述设计的可视化方案的特点。

第三部分将对本文主要内容进行简单的总结。

1 主要的可视化方案

根据挑战二中提供的数据主要有：邮件主题、收发件人名称、收发件人的地址、邮件的重要性权值、邮件发送和接收时间、邮件大小以及附件名称10个属性。对应挑战赛题目的要求，根据原始数据所能提供的信息设计了相关的可视化展示方案。

提取有效的Hacking Team有效员工列表，在对公司员工可视化分类时，采用散点图来体现员工的重要程度，圆圈的大小直接体现重要程度的高低。用散点图和折线图展示每个员工的发邮件时间规律以及有业务来往的公司域名，侧面反映员工的行为特征。采用概念图的方式利用工作职责对员工进行分类，直观阐述每个员工负责的业务。

邮件的分类能够让分析人员通过多层对信息的筛选过滤，获取最有价值的信息。本文采用K-means算法对邮件进行聚类，提取公司主营业务，进一步对邮件主题进行分类，并最终选择Sunburst图进行可视化展示。

对公司业务发展阶段的描述，采用时间线以及标签云的可视化效果。但是在冗长月份中该如何确定关键的时间节点，以对业务进行阶段划分，是个值得考虑的问题，后续我们会进行详细的说明。

下面三个小节将详细论述不同的可视化方案。

1.1 Hacking Team 内部员工列表以及员工分类

1.1.1 数据清洗

提取内部员工列表过程中，主要的数据清洗工作有3部分。第一：针对原始数据中收发件人格式不统一的情况，利用正则表达式对收发件人邮箱地址名进行初步清理，提取名字中的主体部分。第二：对相同的人员使用不同称呼的情况，最终将同一个人的多种称呼统一到同一个名称。第三：筛选有效HT成员名单，将发送邮件数为0并且接收邮件数量小于500的账户过滤掉，最终获得公司有效成员列表，共计132人。最终的公司员工拓扑图如图1.1所示：

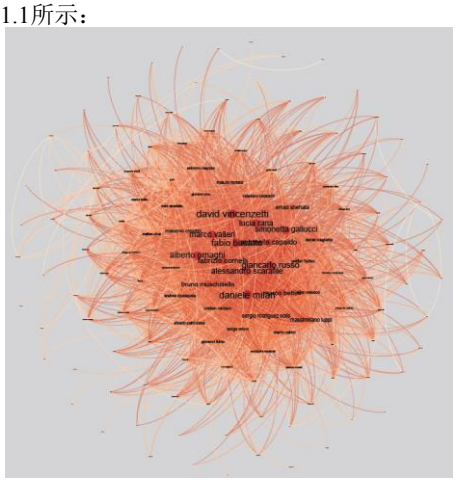

图 1.1 人员重要性可视化展示及个人行为特点

1.1.2 员工分类可视化

按照员工的重要程度进行分类，如图1.2所示。首先采用列状散点图从四个维度衡量员工的重要程度，分别为收到邮件的重要程度、邮件关联的人数、收发邮件的统计数量以及该员工的入职时间，点越大表示在该维度越重要。环状散点图和河流图联动的可视化方案展示员工的行为特征。综合四个权值来看，可以直观辅助分析人员发现公司中地位重要的人物，如：David•Vincenzetti、 Daniele•Milan等。同时点击成员列表，关联该成员工作时间规律分布及负责客户所属的域名信息，进一步验证人员重要程度的正确性。其中环状散点图直线表示24个小时，每一圈代表一年，直观发现成员发邮件的时间习惯。折线图表示与该成员有业务往来域名的收发邮件数量的统计信息，形象而直观的体现这个成员负责的客户所属的公司，以及与该公司联系频率的趋势。

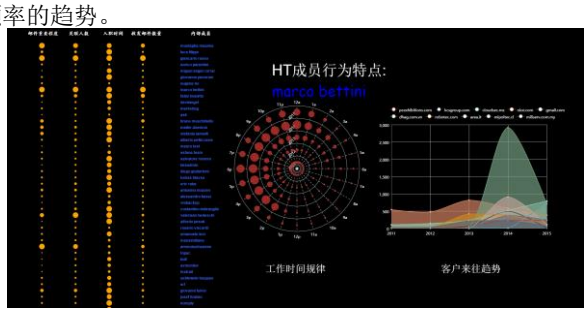

图 1.2 人员重要性可视化展示及个人行为特点

- *田甜. 中国科学院信息工程研究所 研究生. E-mail: tina_sweet877@sina.com.*
- *陈璐. 中国科学院信息工程研究所 研究生. E-mail: chenlu. seu@gmail.com.*
- *刘松. 中国科学院信息工程研究所 研究生. E-mail: lstp678@yeah.net*
- *汪鑫 中国科学院信息工程研究所 研究生. E-mail: wangxin32@126.com*
- *刘俊荣 中国科学院信息工程研究所 助研. E-mail: ljr_0527@163.com*

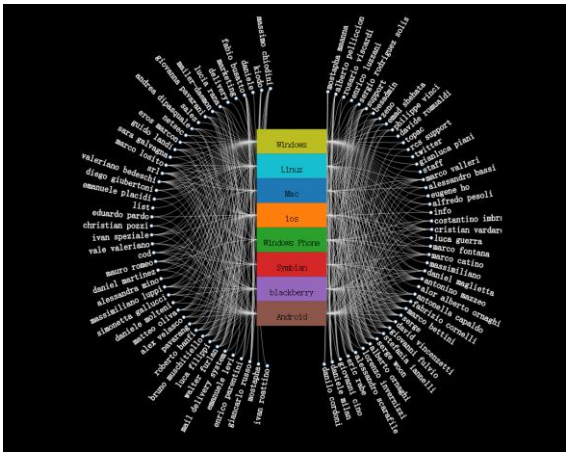

图 1.3 员工工作职责分类

概念图主要通过公司的主营业务来关联涉及到的公司内部成员，这样从负责业务的角度对员工做分类。如同1.3，中间矩形框内表示公司所涉及到的业务，四周均分分布的圆圈表示公司的所有内部成员，我们将光标放在某业务上时，高亮的显示该业务所涉及到的成员。同时，当鼠标光标移动到某Hacking Team成员名字上时，也会高亮显示该成员这几年都接触过哪些业务。

1.2 邮件分类

根据数据清洗的结果对剩余邮件数据提取邮件主题，利用K-means算法进行邮件主题聚类，参数K选择3，选取每个类别频率最高的前20个词作为输出。

根据邮件主题出现的关键字，将邮件依次分为业务邮件及非业务邮件，如图1.4所示。在分类过程中发现非业务主要是生活相关的主题。业务方面，又从操作系统相关和攻击类型两个独立的维度考虑，一个是描述公司的产品本身的功能，另一个则描述它所针对的对象。由图中，我们可以发现，可分类邮件中，业务邮件占据了近四分之三，其中又有近六分之五的比例是在描述攻击类型（方法），包括Exploit，RCS，Botnet，Malware，DDOS这些常见的黑客手法，其中Exploit 和RCS最为显著。RCS作为Hacking Team的最知名的产品，频繁出现在邮件中，包括商务洽谈邮件及开发邮件。操作系统角度，可以看出公司的产品几乎覆盖了所有的PC端和移动端，从Android到Winphone，再到Windows，Linux等均有涉猎。

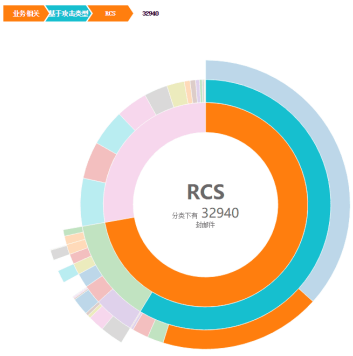

图 1.4 邮件分类

1.3 公司发展阶段

问题三主要是要阐述Hacking Team业务随时间的更迭状况。我们首先采用折线图和河流图来分别对邮件数量、人员数量及涉及关联客户所属的国家数量进行趋势分析。效果如图1.5所示。

从河流图中可以直观的看到，2013年的7月开始，公司的客户逐渐增多，开始第一个增长高峰期，到2014年5月有个较小的波动，此后数量几乎稳步提升，直到2015年4-5月达到峰值。再往后便是邮件泄露事件，各国与Hacking team的联系数量急剧减少。从客户所属国家分布趋势与邮件数量趋势图可以得出：客户增长与邮件数目增长趋势是同步的。最终确定公司业务发展的关键节点。分别为2013年7月、2014年5月和2015年6月。

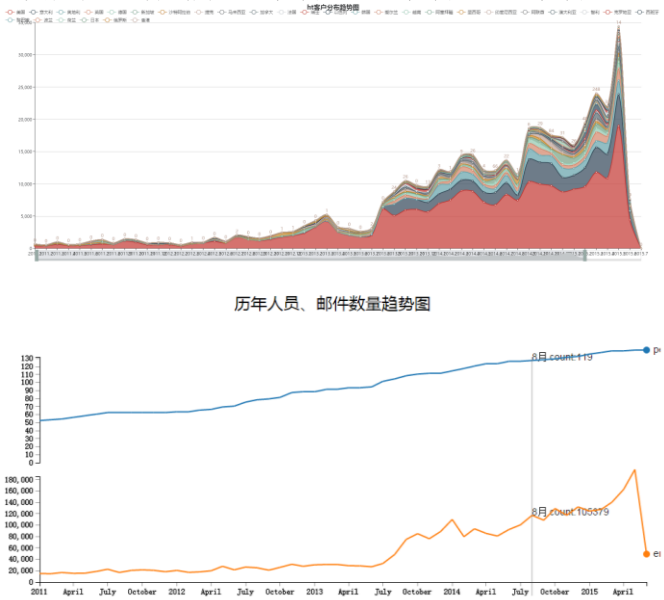

图 1.5 对邮件数量、人员数量及涉及到的国家数量的趋势分析图

主营业务发展阶段可视化采用标签云和时间轴的方式展示，如图1.6所示。从上一个阶段到当前的所有业务及该业务线下的主题词，直观展示公司业务增删及量级的变化。同时采用折线图协同展示不同业务随时间变化的趋势。字符云显示了各发展阶段的主流业务及变更情况。例如，2012年5月之前公司的业务以及邮件频繁度都较低，之后开始有起色，2013年7月左右Android成为主要支撑业务。从Android的业务的发展趋势可视化结果进一步验证了利用折线图和河流图推断公司主要发展阶段时间节点的正确性。

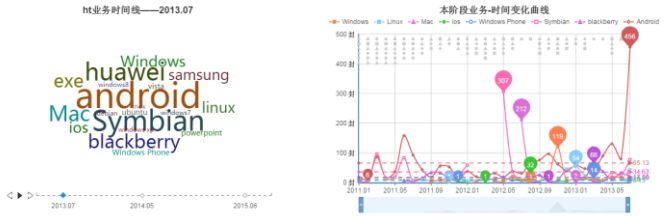

图1.6 公司的三个主要阶段以及各阶段主要业务

2 可视化交互性

对于成员重要程度的可视化交互，首先体现在员工重要程度可视化中，增加每个成员个人行为特点可视化协同联动，包括工作时间规律分布及关联客户所属公司。

其次在邮件分类中，不同的扇区显示该分类的信息，以及它的父类，以层级的方式展示，对用户来说信息的获取更为直观。

最后，业务发展阶段的展示方式采用字符云和时间轴，时间轴设置为自动滚动，用户也可以选择暂停查看。在时间变化的同时，右边的折线图支持实时加载。用户可以点击上面的图例设置该维度的折线开关状态，也可以拉动时间轴对时间范围进行缩放。

3 结论

本次挑战所选取的可视化方案形象且简洁明了的对应了挑战赛的要求内容。同时，该可视化设计方案有着良好的扩展性，对于数据的处理，接口的编写也具有普适性，能够更好的帮助用户发现隐含规律与有价值的信息链。

参考文献

[1] 邱南森. 数据之美. 中国人民大学出版社,北京,2014.
[2] https://wikileaks.org/hackingteam/emails/