EmailVis 电子邮件中事件可视分析

罗珣^{1,3)}, 张晴宇^{1,3)*}, 方芳^{2,3)}, 路强^{1,3)}

¹⁾ (合肥工业大学计算机与信息学院 合肥 230000)

²⁾ (合肥工业大学管理学院 合肥 230000)

³⁾ (智能互联系统安徽省实验室(合肥工业大学) 合肥 230009)
(douglas1995@163.com)

摘 要: 对电子邮件文本消息与收发关系的可视化, 能够有效地分析过去发生的重要事件与人员交互. 然而现有的电子邮件可视化软件对电子邮件主题的筛选与聚类较为简单, 使得用户难以挖掘出更深层次的信息. 鉴于这些不足, EmailVis 通过多种角度对电子邮件进行归纳与分类, 最终获取到可能包含重要信息的电子邮件, 并将它们以点、线与邮件线程相结合的方式展现出来, 以便用户快速定位到重要事件的发生时间与相关信息. 最后, 用户研究表明, EmailVis 可以帮助公司管理人员更加清晰地回忆过去发生的项目完成进度、部门交互关系与异常事件信息, 为改进工作流程和协调部门合作提供有力的数据支持, 从而提高人员协作效率.

关键词: 电子邮件可视化; 视觉分析; 主题相似度; 事件挖掘; 协同工作

中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2021.18255

EmailVis Visual Analysis of Events in Emails

Luo Xun^{1,3)}, Zhang Qingyu^{1,3)*}, Fang Fang^{2,3)}, and Lu Qiang^{1,3)}

¹⁾ (School of Computer and Information, Hefei University of Technology, Hefei 230000)

²⁾ (School of Management, Hefei University of Technology, Hefei 230000)

³⁾ (Intelligent Interconnected Systems Laboratory of Anhui Province (Hefei University of Technology), Hefei 230009)

Abstract: Visualization of email text messages and the relationship between sender and receiver can effectively analyze the past important events and human interactions. However, the existing email visualization software is relatively simple to filter and cluster email topics, which makes it difficult for users to excavate deeper information. Given this, EmailVis summarizes and categorizes emails from multiple perspectives, and finally obtains emails which contain important information displaying them in a combination of dots, lines, and email threads, so that users can quickly locate the time and related information of important events. Finally, user research shows that EmailVis can help company managers more clearly memorize the past project completion progress, departmental interactions and abnormal event information, and provide powerful data support for modifying work process and coordinating departmental cooperation, thereby improving the efficiency of personnel collaboration.

Key words: Email visualization; visual analysis; topic similarity; event mining; collaborative work

收稿日期: 2020-02-29; 修回日期: 2020-08-22. 基金项目: 国家自然科学基金(61972130); 国家留学基金(201806695039). 罗珣(1972—), 女, 硕士, 讲师, 主要研究方向为可视化与可视分析; 张晴宇(1995—), 男, 硕士研究生, 论文通讯作者, 主要研究方向为可视化与可视分析; 方芳(1979—), 女, 博士, 副教授, 硕士生导师, 主要研究方向为商务数据分析、资源调度与优化; 路强(1978—), 男, 博士, 副教授, 硕士生导师, CCF 会员, 主要研究方向为可视化与可视分析、计算机图形学.

电子邮件是目前主流的文字交互方式之一, 尤其在公司内部, 电子邮件往往拥有着比即时通信软件更加重要的位置, 因为它的文本内容往往更加凝练简洁, 也更加规整正式, 例如, 公司对全员的通知公告, 或是对某项重要事件的决议结果, 或是项目中若干个部门之间配合协作的纪要总结. 作为一种较为正式的留存文件, 它贯穿于日常工作中, 对于责任分配与工作协同都有着重要意义. 电子邮件中不仅包含丰富的文本、图像和附件信息, 它也具备繁多的类型, 例如, 机密邮件、日常抄送邮件以及垃圾邮件等, 它们的主题、数量和重要性皆不相同. 这也意味着, 当电子邮件数据集较大时, 想通过电子邮件交互来回忆过去的事件信息, 大量的无意义邮件会给人们的分析带来干扰, 也会消耗人们较多的精力. 近些年, 对于邮件数据的可视化进行了广泛的研究, 例如, 垃圾邮件检测、用户行为分析以及社交网络分析等, 如文献[1]通过对邮件主题关键词的筛选, 提炼出一定时间内出现频率较多的词语, 以进行用户行为分析与回顾事件. Kerr^[2]通过线弧来展现邮件的时间顺序, 使用分支树结构来可视化用户的邮件收发信息, 并且优化了线弧之间的排列方式, 减少线弧之间的交叉, 从而给用户带来比较优秀的视觉体验. Luo 等^[3]开发了 EmailMap 电子邮件可视化工具, 通过展示电子邮件中的事件流与接触轨迹, 来挖掘出项目流程中出现的复杂信息. Li 等^[4]提出了一种可视化模型, 对单个用户累积的大量电子邮件内容进行分析, 并可用于检测电子邮件的滥用, 如病毒邮件传播和垃圾邮件传播. 文献[5]创建了 2 种电子邮件可视化视图来描述电子邮件关系, 一种突出了个人社交网络, 另一种描绘了个人邮件互动的时间节奏, 并且可以协助用户进行故事重述和回忆.

1 相关工作

时间线叙事可视化是将故事或者文本内容以故事线的形式展现在人们面前, 来体现人物之间的关系或者交互信息, 以此来增进人们对故事情节或者事件的理解. 近些年, 很多学者针对比较连贯的文本, 如书籍和故事等, 都已经做出了比较优秀的作品. 如 Shi 等^[6]对会议进行录音并提取关键词, 然后用故事线的方式展现在用户面前, 来帮助用户回忆会议流程中学者们发言的主题以及讨论内容. Lu 等^[7]结合扇形与线条来展现整个故事线,

从空间上减少用户的视觉混乱的同时, 展现故事中的交互关系, 且所提出的极坐标分层图可视化方法也可以揭示不连续事件和非线性故事. Kim 等^[8]主要是使用叙事时间线与实际时间线这 2 条时间线来展示故事的发展脉络, 通过这种方式可以方便用户十分有效地分析剧情较为复杂的电影, 尤其是当剧情充满了大量的时间节点切换时, 如主角进行反复回忆的时候. Tanahashi 等^[9]通过不同颜色的线条来对应不同的角色, 展示故事中不同人物的事件以及故事脉络. Liu 等^[10]主要是通过优化可视化线条来减少线条之间的交叉以及空间上的空白部分, 从而得到了一个非常精致的线条布局.

对于文本分析, 也提出了许多优秀的分析方法. 其中, 杨谢柳等^[11]针对非规则排列汉字文本时分割正确率低的问题, 提出一种基于连通区域捆绑规则的汉字字符分割方法. 巫英才等^[12]总结了文本可视分析的一般流程, 并展望了文本可视分析面临的 4 大研究挑战: 海量数据规模、复杂数据的不确定性、数据融合以及人机交互. 汤斯亮等^[13]总结了以隐含狄利克雷分配模型(latent Dirichlet allocation, LDA)为主的文本概率主题建模技术及其发展, 讨论了新闻主题建模特点, 对比新闻主题可视化的若干方法, 并对其进行分类. 张怡等^[14]提出一种可视化分析方法, 有效地分析了不同来源的新闻事件评论主题的异同, 以及随着时间推移互相影响的转移趋势; 并使用 LDA 模型提取文本主题, 通过计算主题的内容相关性和时间相关性得到主题间的传递关系.

本文所展示的 EmailVis 邮件事件可视化分析系统的主界面如图 1 所示. 界面上方展示了最终的邮件-事件序列线性视图, 即在一段时间内的邮件交互关系中主要事件与时间的关系. 事件视图的左下方展示了收发邮件数量较为频繁的用户, 通过在线性视图中选择不同的时间的区域, 可以进一步生成不同区域内的主要邮件交互用户, 以便快速定位事件中的涉及的核心人物. 点击具体的时间节点或者相关的邮件收发人, 右下方会展示邮件时间和内容等详细信息.

EmailVis 对邮件数据集进行可视化的主要目的在于挖掘并分析以下 3 个问题:

(1) 在过去的哪些时间段内, 公司可能发生过重要事件? 这里的重要事件包括会议交流、多个部门的交互活动及邮件往来非常频繁的工作事务等.

(2) 在这些重要事件中, 是否存在某些事件属于共性问题, 导致在每个月都需要进行比较频繁

的电子邮件沟通?

(3) 针对所发现的热点问题, 是否能够对公司的业务流程做出优化? 这里的优化包括通过流程

调整来减少跨部门之间的邮件沟通, 或者对于需求较多的热点问题进行修改流程或增派人手, 以提高人员的工作效率。

图 1 EmailVis 主视图

2 邮件数据集处理

2.1 邮件数据集筛选

本文使用的电子邮件数据集取自某制造业公司的 100 名不同部门的员工, 从 2013—2014 年约 9 万封往来电子邮件, 并且对收件人、发件人和抄送人信息进行了脱密处理。EmailVis 所使用的邮件处理与可视化流程图如图 2 所示, 其从 4 个方向对邮件数据集进行筛选分析, 并以故事线的形式将邮件事件呈现在用户面前。

图 2 邮件处理与可视化流程图

(1) 垃圾邮件的过滤。对于较为庞大的邮件数据集, 其中必然充满大量的公告通知或欺诈邮件, 它们的文本内容并不具备分析价值; 但是以故事线形式呈现用户邮件交互行为时, 大量的垃圾邮件可能会导致节点产生额外分支的或者线条数量的增多, 从而引起视觉效果的混乱。通过收发邮件数量, 工作部门对员工加以区分, 结合不同员工节点之间的联系交互, 可以有效地过滤垃圾邮件发件人。当某一发件人短时间内发送了大量电子邮件给其他用户, 却没有收到回复邮件时, 这代表此发件人发送的邮件文本重要性较低, 并且也没有产生后续的回应; 在进行邮件分析时, 此类电子邮件即可归类于垃圾邮件的范畴并予以删除。

EmailVis 使用了力导向图来可视化用户之间的交互关系, 其中, 不同颜色的圆角矩形代表着不同的部门, 圆角矩形的颜色对应着不同部门。每个圆圈节点代表着一名用户, 圆圈的直径对应了该用户积累的电子邮件数据集的大小。如果某一用户的邮件往来仅局限于本部门, 则会在该用户与本部门圆角矩形之间产生一条连接线; 当某一用户与其他部门出现频繁的邮件往来时, 在接收者和发送者之间生成一条新的连接线。如图 3 所示, 可以看出多数员工的邮件往来局限于部门内部, 但是部门之间也存在一定的邮件往来关系。在部门 D_A , D_B , D_C 中, 部门 D_B 的邮件总量最多, 而

电子邮件最多的员工 U_A 来自部门 D_A , 并且用户 U_A 与 U_B 在工作中产生了较为频繁的邮件往来; 同时可以看出存在着用户游离在力导向图主体之外的情况, 如用户 U_C , 它并没有与其他员工产生通信关系, 但是它依然拥有较多的邮件总量, 因此可以合理地认为他们的邮件多数为公告通知或者垃圾邮件, 应考虑从邮件数据集中剔除。

图3 使用力导向图来剔除垃圾邮件信息

(2) 电子邮件的主题分类. EmailVis 使用自然语言处理工具包来将电子邮件文本中的单词数据分割后, 删除其中的停顿词, 再根据筛选过的电子邮件主题关键词计算每个电子邮件主题单词的出现频率, 并填充同义词库. 最后, 使用 LDA 模型进行主题文本分析, 获得若干个主题关键词, 这些关键词体现了电子邮件数据集主题所涉及的业务方向. 提取关键词涉及的业务方向有“风险管控”“固定资产管控”“产能管理”“原材料”“燃料”“弱电设备”和“排放许可”等; 可以看出, 这是一家以制造行业为主, 同时十分注重生产成本、产品质量与生产安全的公司。

(3) 电子邮件发件人部门的筛选. 通过分析不同部门中电子邮件的类型与数量, 可以看出服务部门的电子邮件交互比其他部门要高得多. 这是因为服务部门除了日常的部门内部通信, 还会收到来自其他部门的工作需求或者业务服务类型的电子邮件. 因此, 在进行邮件可视化时, 以服务部门为主体能够获得更加清晰的邮件交互脉络. 因为它们与其他部门之间的业务合作联系更加繁多, 也意味着更有可能挖掘出可供改善的工作流程. 但这并不代表需要放弃其他部门的邮件, 而是仅仅显示与服务部门互动的电子邮件部分即可, 这样可以显著地减少最终可视化的线条, 从而让视图更加简洁明朗。

(4) 对于邮件发信人的筛选. 在部门人员的选取方面, 也应从 3 个维度来进行考量: 邮件数量、邮件主题的业务方向和邮件往来持续时间. 在邮件数量与邮件主题的业务方向, 一个员工在公司的职务越高, 其电子邮件主题所涉及的业务方向会越广泛, 即便处于同一部门, 领导角色的电子邮件往往要涉及部门正常运营的方方面面, 因此在邮件主题范围上也会高于其他普通员工. 本文定义若干名员工针对同一封初始电子邮件的后续若干次电子邮件响应行为, 称为“电子邮件线程”. 在邮件交互时间上, 领导角色的电子邮件交互时间会更久, 电子邮件线程也会更多. 例如, 一位客服角色和一位领导者往往都具备大量的邮件数量, 但是往往客服角色的电子邮件线程数量要明显小于领导者的电子邮件线程数量. 因为客服角色的问题往往可以快速解决, 并不需要长时间, 涉及多名角色的邮件往来; 而领导者角色的人常常需要在持续的一段时间内, 围绕同一主题进行反复的协商与对接。

EmailVis 使用玫瑰图来进行员工的筛选, 玫瑰图扇形的高度代表着对应人员总电子邮件数量, 而对应扇形的弧度代表了对应人员电子邮件涉及的业务方向. 从图 4 所示可以看出, 纪强具备发送最多的电子邮件数量, 因为它对应的扇形半径最长, 而车璐的电子邮件所涵盖的业务方向最多, 因为它对应的扇形弧度最长。

图4 使用玫瑰图分析员工邮件重要性

为了量化地计算出比较重要的成员, 即区分出前文所述的客服角色与领导角色. 对于每个成员 x , 其邮件主题涉及的业务方向范围为 f , 个人总邮件数量为 n , 平均邮件交互的天数时间为 d , 计算公式 $N(x) = -\lg(1/n(x))$ 主要是为了削减邮件数量的比重, 以便能够更加合理地体现出人员的重要程度, 成员重要程度计算公式为

$$I(x) = f(x) \cdot N(x) \cdot d.$$

2.2 Jaro 文本相似度

对于同主题电子邮件的筛选也同等重要, 其主要针对处理过的电子邮件数据集的邮件, 将一段时间内文本相似度达到阈值的邮件集合在一起, 这些邮件代表着某一时期不同种类的人员共同参与着某个项目或者事件. 本文使用了邮件主题字符串之间的 Jaro-Winkler 距离来判断邮件主题之间的文本相似度. 首先计算 2 个字符串的 Jaro 距离, 其定义为

$$S_J(m) = \begin{cases} 0, & \text{if } m = 0 \\ \frac{1}{3} \left(\frac{m}{|S_1|} + \frac{m}{|S_2|} + \frac{(m-t)}{|m|} \right), & \text{otherwise} \end{cases}$$

其中, $|S|$ 代表主题字符串长度; m 代表顺序相同的字符匹配的数目; t 代表换位数目, 即相互匹配的字符决定了换位数目 t , 即不同顺序匹配字符数目的一半为换位数目 t . 2 个分别来自 S_1 和 S_2 的字符如果相距不超过 H , 则认为这 2 个字符串是匹配的; 其中, $H = (\max(|S_1|, |S_2|)) / 2 - 1$.

然后计算 Jaro-Winkler 距离. Jaro-Winkler 距离给予了起始部分就相同的字符串更高的分数, 其定义了一个前缀 p , 2 个字符串, 如果前缀部分有相同字符串, 则 Jaro-Winkler 距离为

$$S_{J-W}(m) = S_J(m) - (lp(1 - S_J(m))).$$

其中, S_J 代表 2 个字符串的 Jaro 距离; l 代表前缀的相同的长度, 且规定其最大值为 4; p 是调整分数的常数, 规定其不能超过 0.25, 否则可能出现 $S_{J-W} > 1$ 的情况. Jaro-Winkler 距离定义常数 $p =$

0.1, 将 Jaro-Winkler 距离大于 0.6 的 2 个字符串定义为相似文本.

如 2 个字符串 S_1 与 S_2 分别为 aboard 与 aborader, 使用 Jaro-Winkler 距离来判断二者的相似度, 首先可以得到 $m = 6$, $S_1 = 6$, $S_2 = 8$, $t = 1$, $l = 3$; 然后计算得到二者的 Jaro 距离为

$$S_J(m) = \frac{1}{3} \left(\frac{6}{|6|} + \frac{6}{|8|} + \frac{(6-1)}{|6|} \right) = 0.861.$$

本文使用标准权重 $p = 0.1$, 最后计算得到二者的 Jaro-Winkler 距离为

$$S_{J-W}(m) = 0.861 + (3 \times 0.1(1 - 0.861)) = 0.903.$$

因为 0.903 已经大于阈值 0.6, 则称这 2 个字符串是相似的.

如图 5 所示, 邮件筛选视图主要由力导向图与玫瑰图 2 个部分组成. 图 5a 所示力导向图下面展示了不同部门节点与不同员工节点的邮件数量, 以使用户可以方便地定位到数量较多或者较少的节点并查看其详细信息. 位于总视图右侧的玫瑰图则显示了对应部门内部不同员工的邮件主题涉及的业务方向以及邮件数量. 通过在力导向图中点击不同的部门, 其部门内部的员工邮件信息将重新生成不同的玫瑰图. 用户也可以根据自身需求选择展示不同数量的员工, 以便能获得更加直观、清晰的视图. 玫瑰图下方展示了用户颜色对应的部门以及对应用户涉及的邮件主题的业务方向, 用户可以通过点击不同的员工姓名条目或者邮件关键词来获取更加详细的邮件内容, 对应员工的扇形区域将在玫瑰图中以高亮显示.

图 5 邮件筛选视图

3 邮件可视化结构

在对电子邮件数据集进行过滤之后,如第2节所述,选取主要的4个服务性部门作为可视化的主体部分,分别是财务部、信息管理部、行政管理部与人力资源部。由于主要目的不是观察员工的电子邮件,而是查看部门之间的交互,因此,对同一部门的员工使用相同的配色方案。线性电子邮件可视化遵循以下规则:

(1) 同一部门的电子邮件使用相同颜色的平行线条来表示,线条的粗细与该部门的电子邮件总量成正比,线条的颜色对应不同的部门,线条的延伸对应时间跨度。图6所示为部门 D_A 与部门 D_B 在某一段时间节点内的交互,部门 D_B 于点 a 向部门 D_A 发送电子邮件,该邮件与部门 D_A 的交点 b 取决于此次产生的邮件线程的主题文本相似度,即2条平行的线夹角的 $\sin x$ 取决于发送的电子邮件数据集的主题文本相似性均值。其中,邮件主题的相似性越高,则 $\sin x$ 越大,且部门邮件数据的线条向其他部门聚拢的幅度也就越大。通过比较线条的聚拢速率可以看出,邮件线程的主题紧密性,即在点 c 由部门 D_B 产生的邮件线程的文本相似度要高于在点 a 由部门 D_B 发往部门 D_A 时产生的电子邮件线程。而分离线条的宽度取决于该段时间内电子邮件线程的邮件总数量,如 ab 段弧对应的邮件数量要小于 cd 段弧对应的电子邮件数量。通过这样的视觉隐喻,用户可以有效地了解邮件数据集之间的主题相似度,也可以知道在某一项目过程中不同部门员工之间的邮件交互行为。

图6 部门之间的电子邮件收发关系

(2) 在某个时间段内发生大量邮件交互时,会生成相应的邮件线程,其中邮件线程跨度的长度与该线程的第1封邮件和最后1封邮件的时间有关。用户可以从线程上方的节点获取事件的发起者信息,其节点的中心颜色代表其发送者的相应

部门。为了减少视觉混乱,将同一部门的若干封电子邮件可视化为单个节点,这可以显著地减少视图中的节点数,来自同一线程的邮件会汇聚至同一起始邮件节点,通过这样的方式可以快速定位邮件线程交互的时间段。电子邮件线程的出现意味着多个员工正在围绕同一主题进行交流或者项目合作,从而产生了交互行为。电子邮件线程也可以有效地筛选出更有价值的事件信息,如当邮件交互数量较多而邮件交互时间较短时,它有可能是对于通知类消息的回复,类似于前文所述的客服角色类型的邮件。而重要事件往往意味着邮件数量较多的同时,邮件往来时间也更长,类似前文所述的领导角色类型邮件。

如图7所示,从电子邮件线程 T_B 中可以看出,此次邮件线程的发起方来自部门 D_Z ,并且在一段时间后收到了来自部门 D_Y 员工的回复,随着时间推进,部门 D_X 也加入该邮件线程的讨论之中。电子邮件线程 T_A 包含来自部门 D_X 和部门 D_Y 的员工通信,部门 D_Z 与部门 D_Y 稍后的时间也产生新的电子邮件交互 T_C 和 T_D 。但是线程之间是相对独立的,即它们的邮件交互可能发生于同一时间段,邮件讨论的主题可能相似,但收发件人之间并无明显联系。

图7 部门之间的电子邮件线程

(3) 黄色矩形区域表示所有消息符合主题相似性的消息的阈值要求。即说明它们邮件往来围绕着同一主题,但是人员之间可能没有产生邮件线程。这也说明,即便是围绕着同一项目进行工作,部门之间的合作沟通也可能是相对独立的。通过了解黄色矩形区域中的电子邮件线程与消息,可以帮助用户筛选出重要事件发生的时间段。

(4) 对于不属于主体部门但仍与主体部门存在邮件往来以及相似邮件主题的电子邮件集群,EmailVis中也会显示这些电子邮件的相应部门。这些部门包括研发部门、管理部门和电力技术部门等。当这些部门通常与主体部门的邮件交流较少,但是一旦有了交流,通常就意味着重要事件的发生,非主体部门会在图形上以较短的彩色线条表示,线条的长度对应了沟通时间。

4 邮件可视化结果

通过可视化该公司 2013 年 6 月份与 7 月份的电子邮件数据集, 并分析其中的事件信息. 可以看出, 当频繁地跨部门电子邮件往来或者电子邮件线程出现时, 通常意味着发生了重要事件或跨部门的合作. 如图 8 所示, 可以将事件信息分为 4 类:

(1) 不同部门正常业务往来引起的邮件交互. 此类事件也是占比最多的, 如图 8 所示的 A, C, F, I 和 J 区域. 如图 8 所示, A 区域的邮件交互是由信息管理部首先发起的, 通过研究电子邮件的内容, 此处的邮件内容起始于主题为“网络防火墙的采购策划方案”的邮件, 邮件中提及初步的需求方案已经收集完毕, 申请财务进行审批流程通过. 但是在后续的邮件中, 也可以看到财务对此进行的回复“双签文件中需要添加设备型号, 请确认相关型号并盖章反馈”这样的字眼; 可以看出, 在随后一段时间, 2 个部门进行了方案细节的研讨以及对文件补充或者更改, 由此产生的往来邮件也是形成该邮件的线程的主要原因之一. 在 E 区域, 由研发部门首先向财务部门发送了主题为“第 2 季度维保验收”的邮件, 邮件中提及了 2 点需求: 服务器出现异常故障未能及时解决, 因此需要申请更多的维修预算费用, 以及后续增加维护工程师的人数. 在该线程后续邮件中, 收件人开始转变为人力资源部的员工, 并且在其回复邮件中简要

说明了未来的招聘计划以及技术人员的到岗时间. 在 J 区域, 可以看出是由电力技术部与财务部 2 个部门之间的交互, 该线程的起始邮件名为“新增电算用品预算”, 邮件中说明了新增电算用品的原因以及物品类型, 对于经费使用, 因为成本中心的变动也申请进行此类项目经费的预算调整. 从财务部门回复的邮件来看, 可以合理地认为此类财务预算调整是属于比较常见的业务流程. 在 I 区域, 动力技术部对于网络的点位存在新增需求, 因此需要联系信息管理部来进行对应的网络点位建设以及相关交换机的配置, 信息管理部后续的邮件中给出了施工人员的人员信息以及施工时间.

在 C 区域与 F 区域中, 可以看出大量的邮件交互, 但是并未产生邮件线程. 通过查看相关邮件, C 区域中, 群发了生产合作经理人会议通知, 该邮件收件人数与抄送人数均较多, 邮件提出了会场布置的若干要求, 并要求与会者提供本人参会方式与 IP 地址, 这也导致该起始邮件短时间内收到大量回复邮件, 但是并无后续的邮件往来. 在 F 区域, 邮件中显示这段时间内生产管理部由于电脑显示器需要进行修理, 申请对本部门内的电脑显示器进行大批量更换, 因此在月末这段时间不同的生产管理部的员工持续与信息管理部的员工进行了邮件往来; 从邮件往来的持续时间可以看出, 此类硬件设备更换流程简单, 属于比较常见的业务.

图 8 2 个月份的邮件可视化结果

(2) 异常情况事件. 这类事件通常会出现多个部门的涉入, 从 H 区域的邮件可以看出, 信息管理部对行政管理部进行了例行信息审计, 包括审核审批密级是否符合要求和外访人员信息登记等, 并要求行政管理部提高对信息保密的要求. 在 K 区域, 首先由信息管理部发送的邮件中提及有筛查到异常的上传流量, 从而将该异常情况告知对应的工艺制程部与工程管理部, 并要求对应部门进行自我审计, 检查即时通信软件是否出现上传文件权限管理异常以及外发文件密级. 在后续的邮件收件人中, 可以看到信息管理部門员工的介入, 由此 4 个部门进行了大量的邮件往来, 包括公布了文件泄露时间以及相应的员工惩罚结果, 同时要求相关部门加强对即时通信软件的管控. 这段时间内的邮件, 其主题也多与信息安全相关, 并且邮件内容要求各部门反馈信息权限的相关人员表格.

对于管理者而言, 发生此类事件后, 其相应的部门应加大对信息安全工作的管控, 以杜绝此类现象再次发生.

(3) 部门之间的项目合作. 如图 8 所示的 B 区域和 D 区域. 它们均与多个部门围绕某个项目进行了沟通, 在 B 区域, 由信息管理部首先提出“IT 自助服务需求调研”的主题项目与行政管理部进行了沟通, 由于需要进行相关网络点位的装配, 服务器的部署与管理人員的分配, 动力技术部与人力资源部也在本项目的中途参与进来. 在 D 区域, 主要是行政管理部与财务部就新增薪资查询系统的可行性进行研讨.

可以看出, 对于业务支援体系的部门, 如信息管理部和行政管理部等与其他研发或者生产为主的部门之间涉及的合作项目相对较多, 项目类型也多数围绕着智能化工厂、简化流程及提高员工办公效率等.

(4) 每月在固定日期进行的业务沟通. 如 E 和 L 区域以及 G 和 M 区域所示. 其中 E 和 L 区域均是开发部门和财务部的会议纪要, 结合会议纪要内容与会议开始时间可知, 该研讨会属于月度例会, 会议容包括本月研发成本、预算调整和供应物料采购等项目, 内容多以总结为主. 在 G 和 M 区域, 相应的电子邮件其邮件主题多与月末财务结算相关, 邮件主题也开始存在“申请 SAP 系统接口修改”, “系统单位显示异常”这样的文本. 可以看出, 多个部门在月末共同合作以确保核算与冲账库存资金, 包括采购部门、生产管理部和财务部等;

而冲账与月结过程中一个问题也往往需要多个部门的配合与修正才能得到解决. 分析邮件的收发时间, 以月结为主题的邮件持续时间约为 3 天; 而财务部邮件存在发送时间在深夜或者凌晨的情况, 而且即便周末的邮件交互数量依然会与日常工作期间的数量相当. 可以合理地认为, 在此段时间内, 由于月结工作的要求, 财务部工作量会较为显著地增加, 这也是对员工体力和工作效率的双重考验. 因此, 管理人员可以考虑调整研讨会的会议时间, 或者在月底增加人员来协助财务部的月结工作, 以保证月结工作业务的有序进行.

5 用户分析

为了进一步验证 EmailVis 的有效性和实用性, 本文设计了一个用户调查和评估, 将 EmailVis 可视化方法与文献[1]中较为经典的“时间-电子邮件关键词”可视化方法进行了比较.

邀请了 20 名参与者参加评估, 包括 13 名男性, 7 名女性; 其中 14 名 IT 员工, 3 名硕士研究生, 1 名管理学博士研究生和 2 名项目开发人員. 他们当中, 100% 使用电子邮件, 65% 拥有项目开发经验, 50% 拥有项目管理背景知识, 40% 知道信息可视化.

整个调查过程包括: 对于每个参与对象, 首先向参与者解释视觉交互的含义, 对经典的时间-电子邮件关键词可视化方法与 EmailVis 可视化方法, 展示两者详细的例子并进行讲解. 最后, 参与者可以自由地探索这 2 种方法, 在 EmailVis 中查看不同时间的邮件事件视图, 并且对参与者探索过程中仍然存疑的部分进行解答; 最后由参与者完成对问卷的问题评估.

本文使用 5 等级李克特量表进行评估. 要求用户使用 1~5 分在问卷上对 6 个问题进行评分; 5 分代表非常满意, 4 分代表满意, 3 分代表普通, 2 分代表不满意, 1 分代表非常不满意. 问题描述如下:

问题 1. 视图是否美观且易于理解?

问题 2. 视图能否清楚地显示发送和接收电子邮件的时间?

问题 3. 视图能否知道电子邮件中的关键词?

问题 4. 视图能否知道过去一段时间内公司发生的重要事件?

问题 5. 视图是否可以显示事件与电子邮件主题的关系?

问题 6. 该视图可以为现有公司流程提出改进

建议吗?

所有参与者完成问卷后, 对问卷中的分数进行计算. 图9显示了参与者对2种不同的可视化方法的6个调研问题的看法. 结果表明, 超过90%的参与者对EmailVis视图的美观性非常满意, 并认为它和时间-电子邮件关键词可视化方法都可以清楚地显示邮件主题和时间之间的对应关系. 在可视化事件方面, 用户更倾向于EmailVis, 10名用户对它的可视化方式比较满意, 因为它对于事件邮件的筛选更加科学, 可视化隐喻的方式也更加醒目; 而对于时间-电子邮件关键词的可视化方法, 有9名参与者认为其对邮件内容展示不

够详细, 且不能较好地分析邮件中包含的事件信息. 在问题5上调查者也有较大的分歧, 有13人认为EmailVis可以有效地通过邮件线程快速定位事件的主要信息; 有5人认为经典方法可以展示事件与主题的关系, 他们主要通过同一关键词出现多次时侧面推测出邮件发送者可能在做的事情. 从这点也可以看出, EmailVis可较为有效地对该经典可视化方法起到了补充的作用. 问题6的结果表明, 通过展示的案例, 参与者可以根据邮件事件可视化系统为公司流程完善提出自己的看法, 也说明他们对本文中各个视觉视图的有效性 with 视图交互持肯定态度.

图9 2种可视化方法的调查问卷结果

6 结 语

通过对公司内部电子邮件数据集多种维度的筛选, 本文有效地提取了事件信息, 并将它们以事件线条的形式呈现给用户. 从结果来看, 它从多种维度来衡量发信人邮件的重要性, 有效地筛选出职务较高的员工邮件, 剔除分析价值较低的垃圾邮件. 在电子邮件的可视化分析方面, 它展示了电子邮件之间的交互关系, 方便人们快速地挖掘出过去一段时间内发生的异常与冗余流程, 如较为严重的泄密事件, 或者某些部门在月末工作量迅速提升可能带来的隐患, 协助公司管理者快速地了解公司内部事件与时间的对应关系, 从而对部门协作进行改善与优化. 未来, EmailVis 主要的改善方向在于对邮件事件模块的完善, 对于邮件的解析不局限于文本内容, 还要包括对邮件附件的分析; 在事件分析视图上, 能够表现出不同邮件线程之间存在的潜在关系, 从而更好地协助使用者

分析在持续数月的大型项目中重要事件的时间节点与项目合作关系.

参考文献(References):

- [1] Viégas F B, Golder S, Donath J. Visualizing email content: portraying relationships from conversational histories[C] //Proceedings of the Conference on Human Factors in Computing Systems. New York: ACM Press, 2006: 979-988
- [2] Kerr B. Thread Arcs: an email thread visualization[C] //Proceedings of the IEEE Symposium on Information Visualization. Los Alamitos: IEEE Computer Society Press, 2007: 211-218
- [3] Luo S J, Huang L T, Chen B Y, et al. EmailMap: visualizing event evolution and contact interaction within email archives[C] //Proceedings of the IEEE Pacific Visualization Symposium. Los Alamitos: IEEE Computer Society Press, 2014: 320-324
- [4] Li W J, Hershkop S, Stolfo S J. Email archive analysis through graphical visualization[C] //Proceedings of the ACM Workshop on Visualization and Data Mining for Computer Security. New York: ACM Press, 2004: 128-132

- [5] Viégas F B, Boyd D, Nguyen D H, *et al.* Digital artifacts for remembering and storytelling: posthistory and social network fragments[C] //Proceedings of the 37th Annual Hawaii International Conference on System Sciences. Los Alamitos: IEEE Computer Society Press, 2004: 1753-1762
- [6] Shi Y, Bryan C, Bhamidipati S, *et al.* MeetingVis: visual narratives to assist in recalling meeting context and content[J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(6): 1918-1929
- [7] Lu Q, Chai B J, Zhang H B. Storytelling by the StoryCake visualization[J]. The Visual Computer, 2017, 33(10): 1241-1252
- [8] Kim N W, Bach B, Im H, *et al.* Visualizing nonlinear narratives with story curves[J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(1): 595-604
- [9] Tanahashi Y, Ma K L. Design considerations for optimizing storyline visualizations[J]. IEEE Transactions on Visualization and Computer Graphics, 2012, 18(12): 2679-2688
- [10] Liu S X, Wu Y C, Wei E X, *et al.* StoryFlow: tracking the evolution of stories[J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12): 2436-2445
- [11] Yang Xieliu, Niu Xihui, Liang Wenfeng. Character segmentation method for irregularly arranged text in Chinese[J]. Journal of Computer-Aided Design & Computer Graphics, 2019, 31(9): 1542-1548(in Chinese)
(杨谢柳, 牛玺辉, 梁文峰. 面向非规则排列汉字文本的字符分割方法[J]. 计算机辅助设计与图形学学报, 2019, 31(9): 1542-1548)
- [12] Wu Yingcai, Cui Weiwei, Song Yangqiu, *et al.* A survey on topic-based visual text analytics[J]. Journal of Computer-Aided Design & Computer Graphics, 2012, 24(10): 1266-1272(in Chinese)
(巫英才, 崔为炜, 宋阳秋, 等. 基于主题的文本可视分析研究[J]. 计算机辅助设计与图形学学报, 2012, 24(10): 1266-1272)
- [13] Tang Siliang, Cheng Lu, Shao Jian, *et al.* A survey on news text visualization via probabilistic topic modeling[J]. Journal of Computer-Aided Design & Computer Graphics, 2015, 27(5): 771-782(in Chinese)
(汤斯亮, 程璐, 邵健, 等. 基于概率主题建模的新闻文本可视化综述[J]. 计算机辅助设计与图形学学报, 2015, 27(5): 771-782)
- [14] Zhang Yi, Shao Yudong, Zhang Jiawan. Visual analysis of topic transition among different sources of text corpora[J]. Journal of Computer-Aided Design & Computer Graphics, 2017, 29(12): 2265-2272(in Chinese)
(张怡, 邵裕东, 张加万. 多源媒体文本主题演变的可视分析[J]. 计算机辅助设计与图形学学报, 2017, 29(12): 2265-2272)