

University of Chinese Academy of Sciences



软件与系统安全分析

设计文档

小组成员：

王健 201528015029027

王建飞 201528015029026

赵伟 201528015029034

杨玲 2015E8015061088

曹佳欣 2015E8015061080

2015年12月30日

**原始数据处理：**

1：使用DataAnalyze.cpp 和GenerateHash.cpp 对所有邮件进行分析，统计每个人与bush的邮件往来次数。

1）由于往来任务邮箱地址只会出现在收件人和发件人中，故DataAnalyze.cpp通过正则表达式"^((From)|(To)):\\s+.+" 提取邮件中“FROM: ”和“TO: ”这两行写入文件mediumOutput.txt

2）GenerateHash.cpp使用"[\_a-zA-Z\\d\\-\\./]+@[\_a-zA-Z\\d\\-]+(\\.[\_a-zA-Z\\d\\-]+)+"提取“FROM: ”和“TO: ”中的邮箱地址，以邮箱地址为关键字建立无序的哈希表，哈希表的值即为与bush的邮件往来次数，若哈希表中已有该邮箱地址，将其值加一，若没有则置一，最后将哈希表写入文件FinalOutput.txt。

3）pre\_process\_before\_vis.py将来往次数小于800的联系人过滤掉，剩下的即为绘制关系图谱所需的数据集。

2：splitFile.cpp 和tfidf.cpp使用tf-idf计算每个词项的权重，形成空间向量

3：人工标注300封邮件所属分类：

**生成人物关系图谱：**

1：运行 pre\_process\_before\_vis.py 生成 画图需要的数据的格式

2: 将步骤1 生成的数据拷贝到1.html 中（由于画图使用的js，然而这个程序只是一段代码，不是一个完整的项目，处于安全性考虑，js机制不能读取本地文件）

3：使用Vis.js绘制关系图谱，Vis.js 是一个动态的基于浏览器的可视化库,特点是易用,可处理大量的动态数据,并与这些数据进行交互操作。由于画图使用的js，处于安全性考虑，js机制不能读取本地文件，故将之前生成的数据拷贝到Network\_images.html中，以Job.Bush作为图的中心，根据往来次数，设置边和节点的大小和颜色，往来越密切则点越大，最后根据生成都点和边调用vis.js库绘制关系图谱。

4：在浏览器中打开 Network-images.html 即可看到绘制的关系图谱

**生成兴趣图谱：**

1： 利用K-means算法对所有邮件进行聚类分析，初始K 点的选择可以使用人工标注的300封邮件

K-means是一种基于形心的聚类方法。

假设我们提取到原始数据的集合为，并且每个为d维的向量，K-means聚类的目的就是，在给定分类组数k（k ≤ n）值的条件下，将原始数据分成k类 **S** = {S1, S2, …, Sk}，在数值模型上，即对以下表达式求最小值：

\underset{\mathbf{S}} {\operatorname{arg\,min}} \sum_{i=1}^{k} \sum_{\mathbf x_j \in S_i} \left\| \mathbf x_j - \boldsymbol\mu_i \right\|^2 

这里***μ***i 表示分类Si的平均值。

其算法步骤一般如下：

1、从D中随机取k个元素，作为k个簇的各自的中心。

2、分别计算剩下的元素到k个簇中心的相异度，将这些元素分别划归到相异度最低的 簇。

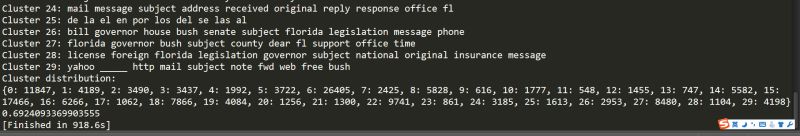
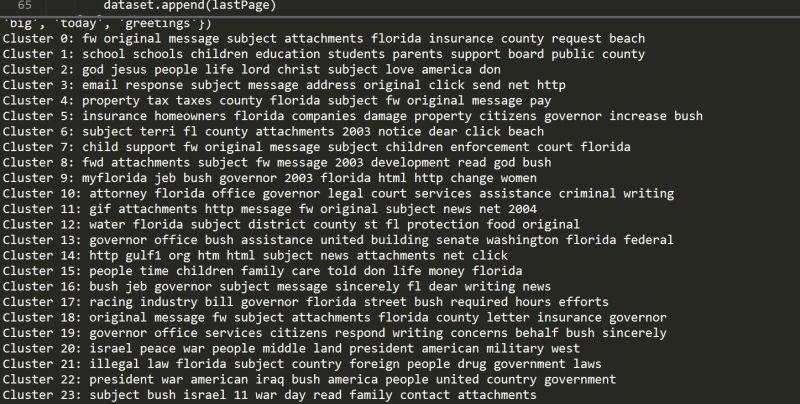
3、根据聚类结果，重新计算k个簇各自的中心，计算方法是取簇中所有元素各自维度 的算术平均数。

4、将D中全部元素按照新的中心重新聚类。

5、重复第4步，直到聚类结果不再变化。

6、将结果输出。

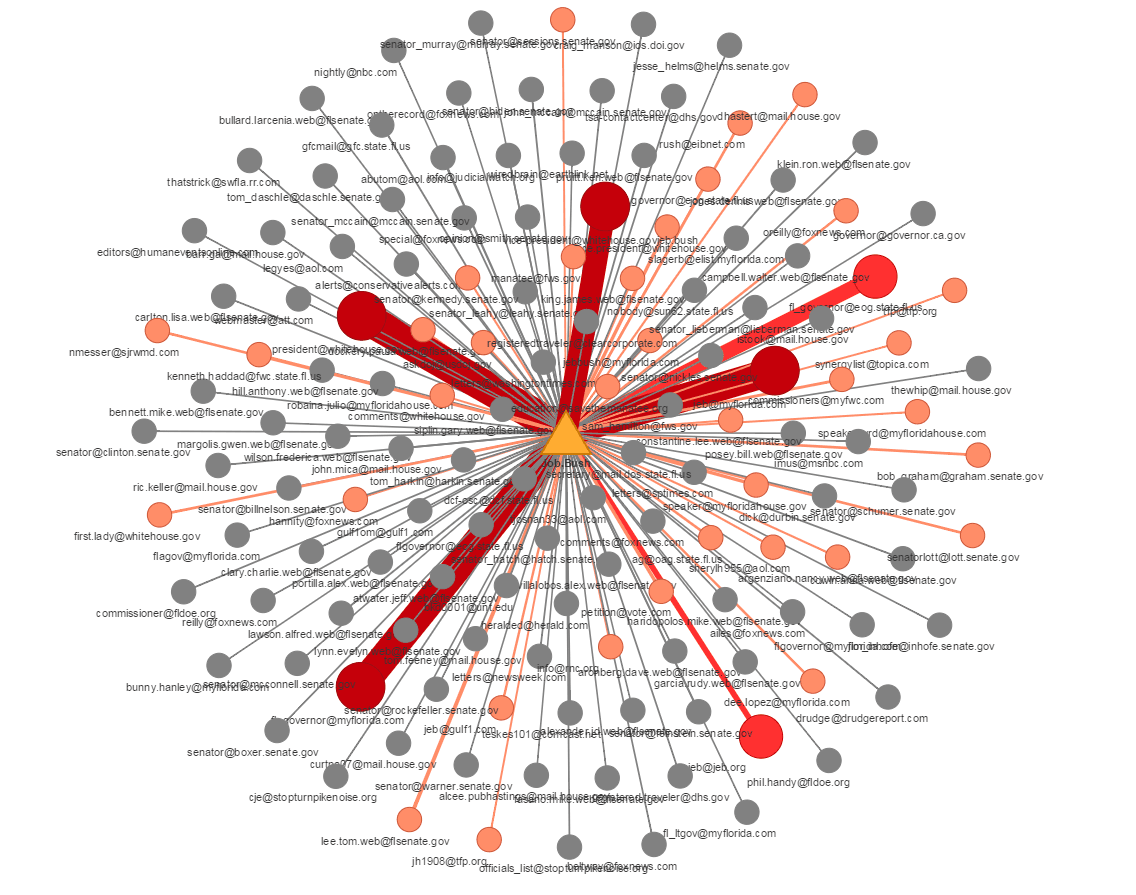
聚类结果如下图所示：



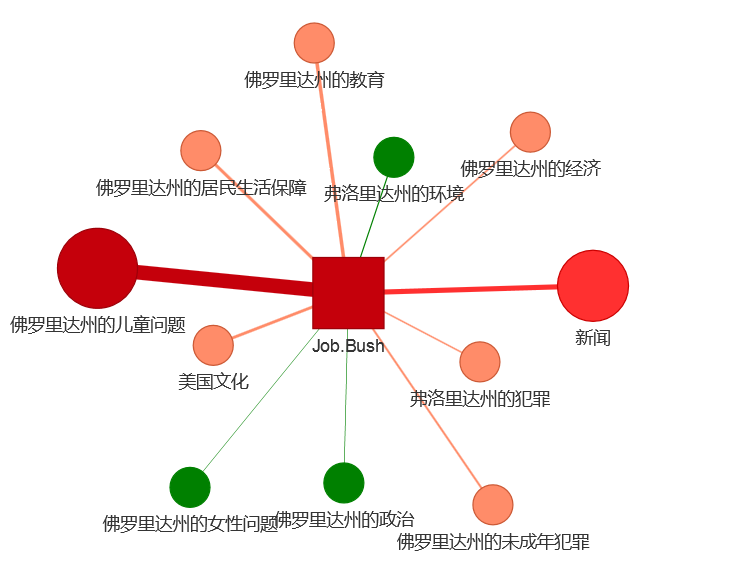
2：将聚类结果数据传到2.html 中，用浏览器打开即可。

**结果展示：**

以bush为中心的人物关系图谱，通过点和线段的颜色，大小和宽度标识此人与bush的往来密切程度。



以bush为中心的兴趣图谱如下图



**项目Github 地址：**https://github.com/danny-wang/mailsOfBushAnalys