2018年9月20日星期四

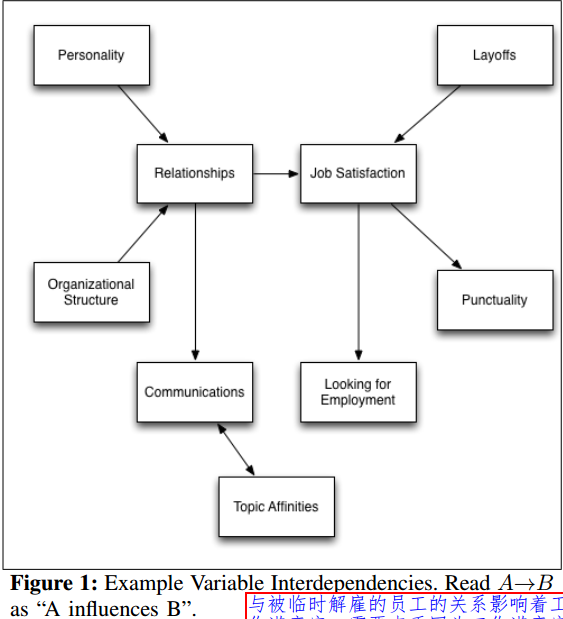
上午10时6分

2018年9月20日星期四

上午9时25分

汲取教训，之前有过先玩乐结果耽误一天的事情。

先做最重要的工作，然后再休息。



继续昨天的分析。

首先，我们现在已经使用到了组织架构数据（LDAP），但是构造人际关系时未使用到人格特征，可以预测的是OCEAN中的E特质影响着用户朋友圈的大小，而朋友圈与工作关系一起构成了用户的Relationship。然而，如何刻画离职人员与该用户的关系呢？

如果是好友关系，可以从邮件通讯频率上进行刻画，包括来往邮件次数，以及双向交流的比例（邮件文字长短等，单向、较短文字通常存在于上下级间工作邮件）

如果仅仅是工作关系，就需要衡量部门距离，如同一团队为距离1，同一部门不同团队为距离2等。。

因此，现在首要工作是：

* 完成CERT5.2的用户数据分用户/时间/行为类型的提取（不包括Web与File以及邮件的内容部门，但是需要包括邮件除去附件的size部分）
* 在上述基础上，学习如何用python构建一个图模型，并针对一个场景二用户的邮件关系进行分析

在实际编写数据提取整理程序前，我们先来做些工作。

* 对于logon数据而言，原始数据格式为：

id,date,user,pc,activity

{Q4D5-W4HH44UC-5188LWZK},01/02/2010 02:24:51,JBI1134,PC-0168,Logon

{G7V0-S4TP95SA-9203AOGR},01/02/2010 02:38:28,JBI1134,PC-0168,Logoff

{B4U7-K4DB84LM-2657VBFY},01/02/2010 04:55:52,JBI1134,PC-2320,Logon

{K2E1-W7VG04OA-3686THSV},01/02/2010 05:02:28,JBI1134,PC-2320,Logoff

我们从中提取的字段为：

Time + User\_id + PC + Activity(Logon/Logoff)

* 对于file数据而言，原始数据为：

id,date,user,pc,filename,activity,to\_removable\_media,from\_removable\_media,content

{Y1W9-R7VJ77IC-9445QFNQ},01/02/2010 08:15:10,TSG0262,PC-9993,R:\79L99n6\H7RHJS5J.zip,File Open,False,True,*50-4B-03-04-14 moved imaging underwent key late appearance span ontario due compiled month 07 sedins final leaders ability doug another presidents improving donation by joseph quadruple 104 agreed 16 brian upon built all to handsome searching track wounded mike march one developer owned 5000 stepping lists orange metacritic second moore supervisor currently initial*

（从USB上打开一个文件R:\79L99n6\H7RHJS5J.zip）

{Y3U8-G5BL42LO-9404XAHI},01/02/2010 08:16:01,TSG0262,PC-9993,R:\79L99n6\H7RHJS5J.zip,File Open,False,True,50-4B-03-04-14 moved imaging underwent key late appearance span ontario due compiled month 07 sedins final leaders ability doug another presidents improving donation by joseph quadruple 104 agreed 16 brian upon built all to handsome searching track wounded mike march one developer owned 5000 stepping lists orange metacritic second moore supervisor currently initial

我们从中提取的字段为：

date,user,pc,filename,activity,to\_removable\_media,from\_removable\_media

* 对于HTTP数据而言，原始数据格式为：

id,date,user,pc,url,content

{D8Q7-C0RU46YI-7391WHNI},01/02/2010 06:46:20,HMI1448,PC-9352,http://nymag.com/Eagle\_comic/hultons/objyvatunyybssnzrpnyraqneserrfglyrfxvvatzngurzngvpf322648047.jsp,eleven 1963 greater literature shorbodolio funding beating treasury both curzon single mourning huq exact visit disobeyed whose not thinking candidates necessary newly elevated eight including head those attempts present had median binds sized replacement colonial databases moderately adaptable symmetrical well drug encourage william 1840 1940s progeny possible variety 1978 on 1987 abandoned

我们提取的格式为：

date,user,pc,url

* 对于Email数据而言，原始数据格式为：

id,date,user,pc,to,cc,bcc,from,activity,size,attachments,content

{N9X0-P3SX99UT-3623QQUN},01/02/2010 06:49:35,KMC1934,PC-2683,Harding.Alden.Lester@dtaa.com;Ora.Clementine.Lester@dtaa.com,Remedios.Kylynn.Weaver@dtaa.com;Hamilton.Palmer.Cortez@dtaa.com,Karina.Melanie.Collins@dtaa.com,Karina.Melanie.Collins@dtaa.com,Send,22578,,lost six speculated individual theoren in ontario pink modeled speculated role most canadian thirty failed including represent shootout owners professional sonne proving around 108 run pled williamson per popular preceded

我们提取的格式为：

date,user,pc,to,cc,bcc,from,activity,size,attachments

* 对于Device数据，原始数据格式为：

id,date,user,pc,file\_tree,activity

{C9S1-Y8GB42VD-2923GATU},01/02/2010 07:27:19,HRE1950,PC-8025,R:\;R:\HRE1950;R:\47yHBn0;R:\54s7J45,Connect

{C3G4-U2ON02HC-9088IHGJ},01/02/2010 07:40:51,EMR0269,PC-6370,R:\;R:\EMR0269;R:\753Cf59;R:\18d36D6;R:\89bc6Q2,Connect

{X4S2-R2YC60OH-9191YYMD},01/02/2010 07:45:00,EMR0269,PC-6370,,Disconnect

{G7C9-E4MQ65OF-8259TKPO},01/02/2010 07:49:50,HRE1950,PC-8025,,Disconnect

我们提取的格式为：

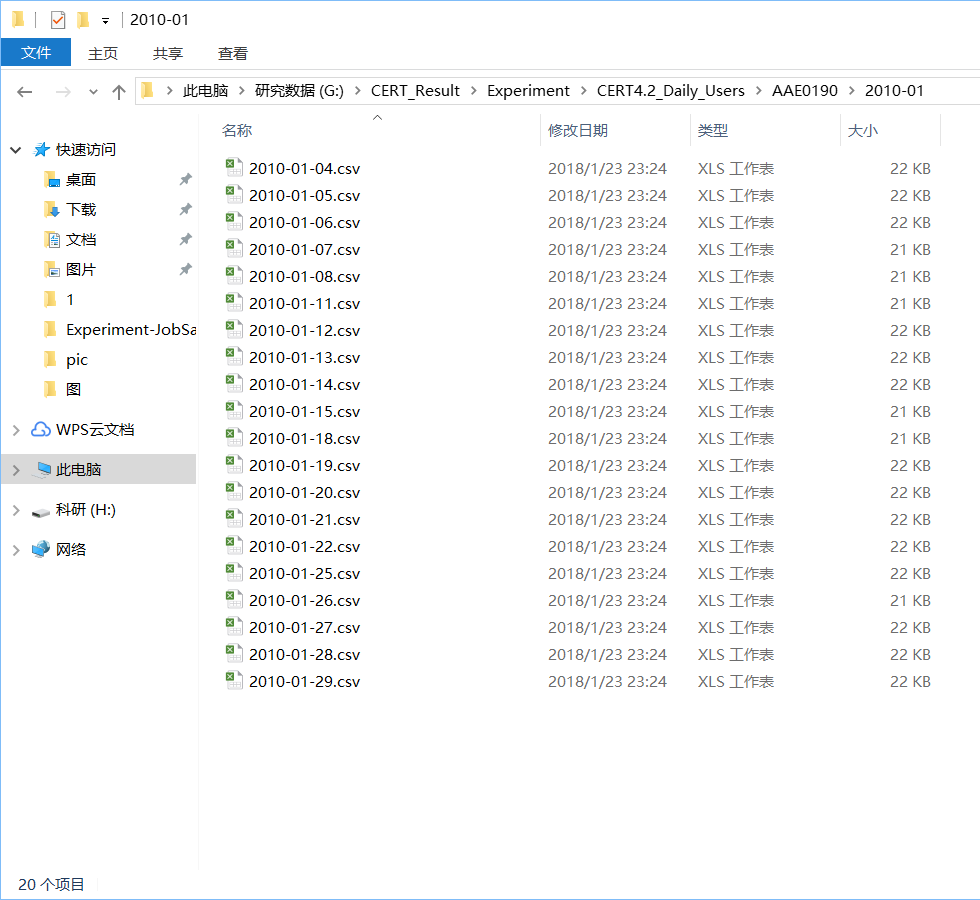
date,user,pc,file\_tree,activity

* 最后综合五种用户行为日志，我们提取的数据仅不包括id与content，基本包含
* Logon: date,user,pc,activity
* File: date,user,pc,filename,activity,to\_removable\_media,from\_removable\_media
* HTTP: date,user,pc,url
* Email: date,user,pc,to,cc,bcc,from,activity,size,attachments
* Device: date,user,pc,file\_tree,activity

然后需要分两个层次对CERT5.2的用户数据进行分析：

1. 从原始的CERT5.2数据中分类出用户以及对应的时间数据
2. 在上步基础上进一步将用户月度数据分类为天数据





2018年9月20日星期四

下午11时46分

经过晚上的初步实验，发现如果对于每个用户的每天遍历全部的HTTP文件实在太费时间了，28个G的HTTP文件读取基本需要10分钟以上。因此我们改变方法，选择采用遍历一次，交替打开写入关闭文件的方式操作。

这样做的前提是CERT5.2数据集是规范数据集，即五个行为数据集均严格按照时间顺序出现，因此编写了一个小程序，检查数据集中先后出现的日期列表的顺序，发现

Logon数据严格按照时间先后记录，即同一天所有用户记录完成后，开始第二天；

文件数据时间假设通过；

Device数据时间假设通过；

Email数据时间假设通过；

2018年9月21日星期五

上午9时20分

继续昨天晚上的工作。

对于python读取大文件而言，如果源文件本身是按照行格式的，即存在’\n’，则可以使用两种方法依次读取：

1. while True:

Line = f.readline()

If Line == ‘’: #最后没有时返回空字符串

Break

1. 推荐，采用Python自动的缓冲区管理，将文件看作一个可迭代对象

With open(filepath, ‘r’) as f:

For line in f:

依次进行的行处理

即可

另外可以一次遍历完成单个用户所有文件内容分类的前提是：CERT5.2数据集每个行为域文件中，严格按照先时间后用户的顺序来排列记录。

即：

* User\_0, file
* User\_0, file
* ....
* User\_1, file
* User\_1, file
* ...
* User\_2, file
* User\_2,file
* ...

其余均同样符合上述规则。

因此，可以在一次读取时，遇到一个新的日期便打开一个对应的文件，写入，直到读入一个新的日期，关闭之前写入的文件对象，重新打开一个新对象，继续即可。