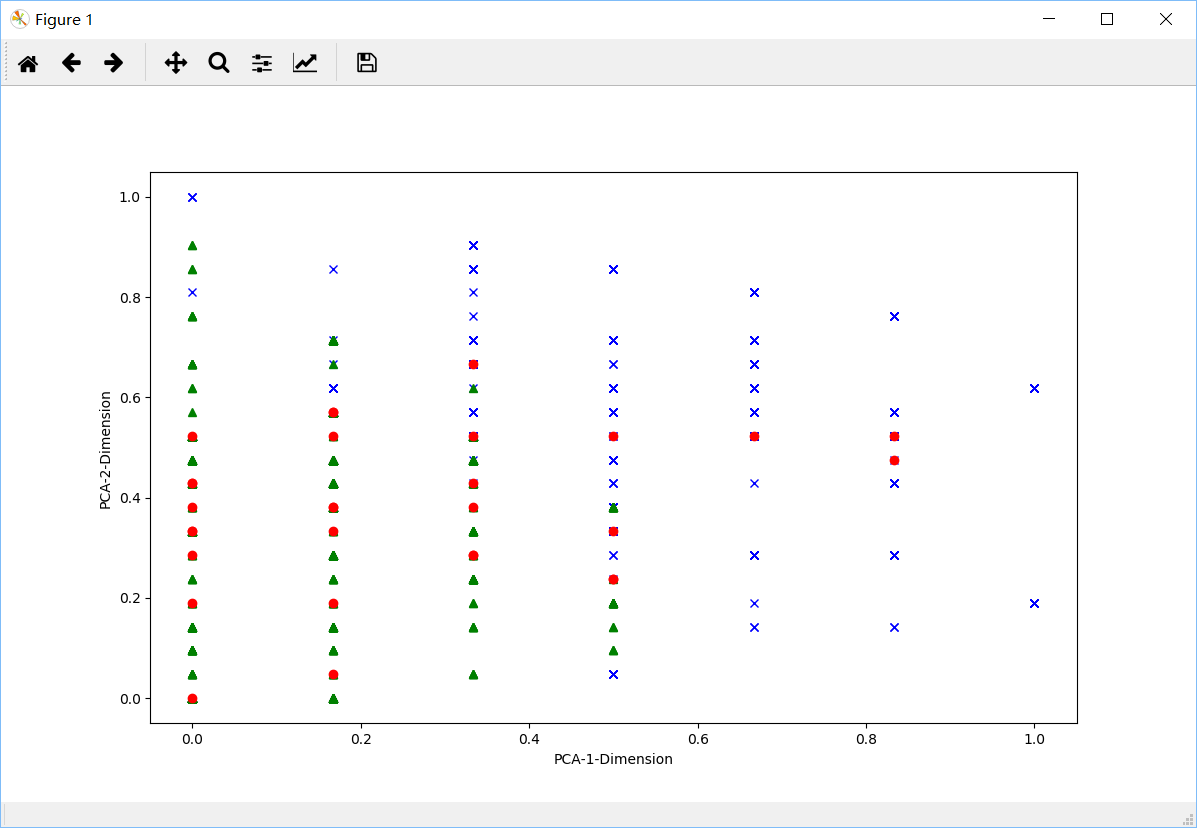
2018年9月28日星期五

下午5时24分

已经下班，自动加班。

随着数据处理开始，自己也着手继续下一步分析。昨天基于CERT5.2中用户离职/解雇的关系数据进行了初步的KMeans聚类分析，但是结果表明无论是PCA还是MinMax都无法挽救该阶段的KMeans，结果如图：



上图为PCA=2时两个维度的图

也许自己不死心吧，决定再次尝试一次，因为论文说“relationship to laid off employees”影响着Job Satisfaction，那么如果直接依据离职人数进行排队呢，不知道场景二的30个用户能排在哪个位置？

首先分析X\_0，即仅考虑同一团队的离职人数，则有：

开始输出Insiders\_2在排序后的索引...

如果考虑X\_1即同一部门内离职人数，依旧不明显

如果考虑LaidOff，依旧结果不明显，此方法作罢。

首先输出X\_0...

30

BYO1846 775

CHP1711 1042

CIF1430 1404

CKP0630 1648

DCC1119 349

GWG0497 1458

HIS1394 463

HMS1658 769

HSN0675 1452

HXP0976 1986

ICB1354 1034

ITA0159 570

JAL0811 1138

KSS1005 1825

LVF1626 105

MCP0611 1555

MDS0680 1654

MGB1235 1964

NAH1366 1557

d'dOSS1463 1063

RRS0056 1702

SIS0042 1383

SNK1280 1459

TMT0851 480

TNB1616 447

TRC1838 1155

VCF1602 557

WDT1634 1577

ZIE0741 1830

......<<<<<<分割线>>>>>>......

2018年9月29日星期六

上午10时42分

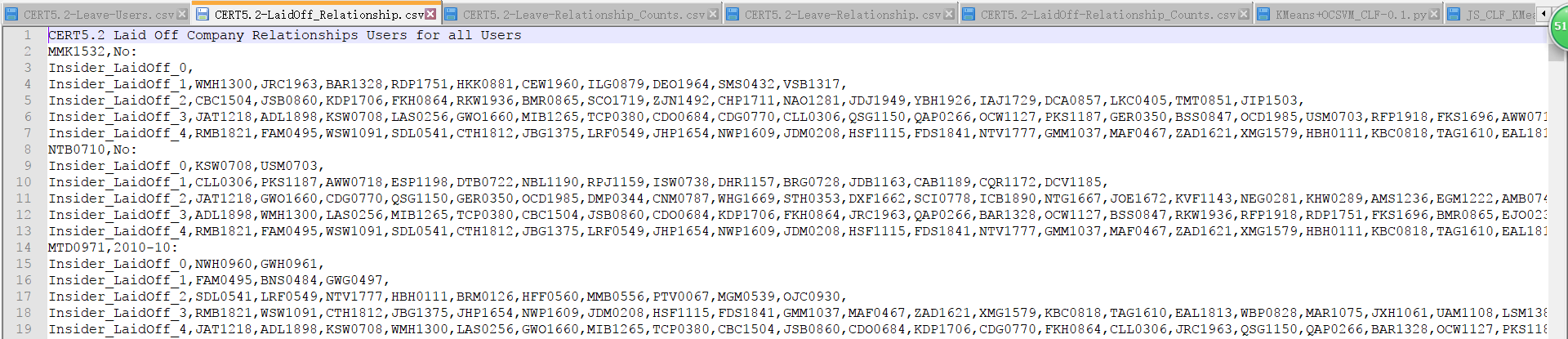
国庆假期调休。

昨天下班时进一步考证了单独从离职人员数量角度无法有效区分，那么看来需要考虑离职人员的亲密度了。

再继续进行下一步前，简单回顾下自己分析离职人员关系的步骤。

1. 统计了LDAP中所有的部门架构层级关系，进而统计了每个部门架构（细化到团队）中的用户成员，得到了CERT5.2-LDAPUsers.csv
2. 在上述基础上，从2009-12的LDAP文件开始，逐个分析到2011--05的LDAP，统计列出这段时间所有离职（工作时间未干全）的员工列表，以及其离职时间（未离职就是NO），得到了CERT5.2-Leave-Users.csv
3. 在上述基础上，重点统计2000个CERT5.2用户的各个层级架构中，出现的离职时间不晚于其离职的用户成员，并分别列入各个层次的列表中，得到CERT5.2-Leave-Relationship.csv；当过滤掉场景二的攻击者后，得到的是CERT5.2-LaidOff-Relationship.csv

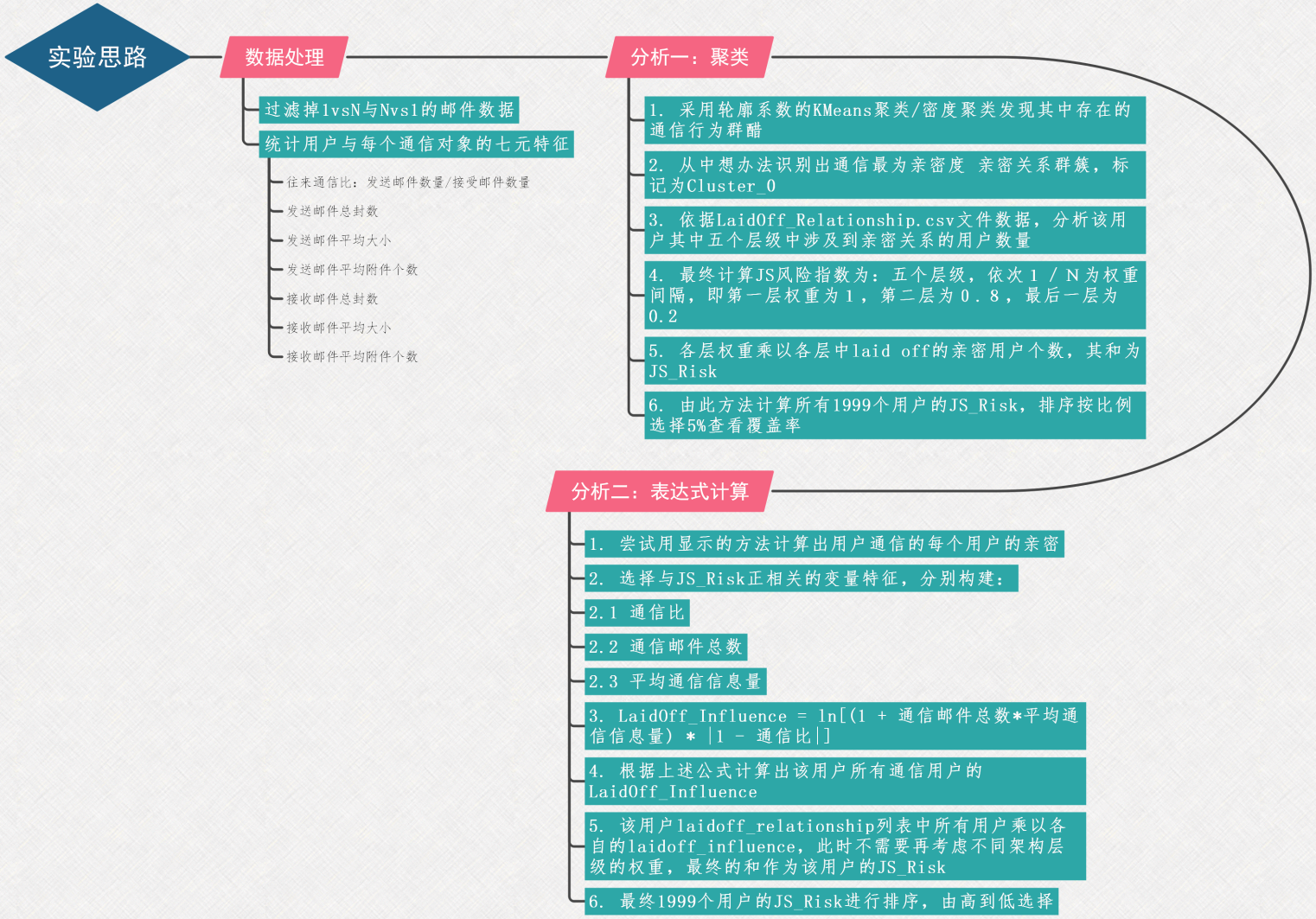
一个简单的示例如：



2018年9月29日星期六

下午2时51分

下午初步整理了下接下来分析的思路，在正式实验前，首先将整体思路放在首位。



完成了单一用户的发送与接收邮件列表，可以看出发送与接收的邮件用户十分不对等

首先统计用户的有效邮件通信列表...

BYO1846 119 发送的邮件用户列表分析完毕...

Macy\_Patterson

Hannah.Mollie.Callahan

Eugenia.Shafira.Mercado

Gareth.Ross.Becker

Xanthus.Kamal.Berger

Micah-King

Tatyana\_A\_Johnston

Jonathan\_Buck

WAS1

Tana\_P\_Orr

BYO1846 42 接收邮件用户列表分析完毕...

Buffy.Yolanda.Ortiz

Kitra.Alexa.Mack

Buffy\_Ortiz

Rose.Maisie.Blackwell

Jonathan.Kasimir.Buck

Olympia.N.Bonner

Kirby.Violet.Pena

Wang.Clark.Rice

Belle.Georgia.Frost

Jemima.Brynne.Pratt

晚上重新修改了程序，结果得到

BYO1846 发送邮件与接收邮件分析完毕...

发送邮件统计示例...

**119 119 （两次结果一样）**

['Macy\_Patterson', 2.0, 5687130.0, 7.0]

['Hannah.Mollie.Callahan', 6.0, 6778336.0, 7.0]

['Eugenia.Shafira.Mercado', 8.0, 11621806.0, 10.0]

['Gareth.Ross.Becker', 5.0, 15104163.0, 12.0]

['Xanthus.Kamal.Berger', 7.0, 14001504.0, 13.0]

['Micah-King', 4.0, 6512878.0, 10.0]

['Tatyana\_A\_Johnston', 1.0, 5938693.0, 6.0]

['Jonathan\_Buck', 2.0, 1627944.0, 2.0]

['WAS1', 6.0, 10403150.0, 10.0]

['Tana\_P\_Orr', 1.0, 2930252.0, 4.0]

接收邮件统计示例...

74 74

['Nelle.Regan.Roman', 14.0, 16731087.0, 20.0] 经过实际验证，NRR确实存在receive邮件

['Nelle.Zephr.Hammond', 6.0, 5106777.0, 7.0]

['Herrod.Bevis.Pearson', 1.0, 1576042.0, 1.0]

['Driscoll.Lewis.Hurst', 1.0, 9091311.0, 8.0]

['YDM6', 4.0, 10587019.0, 12.0]

['Hannah\_M\_Callahan', 11.0, 22461830.0, 24.0]

['WCR66', 1.0, 2053416.0, 2.0]

['Roman-Nelle', 6.0, 9075406.0, 9.0]

['Robert.Merritt.Velez', 26.0, 35519771.0, 41.0]

['Rebekah\_Santos', 15.0, 15771213.0, 17.0]

2018年10月12日星期五

下午2时26分

时隔两周，继续国庆假期前的工作。

由于原有程序终未考虑没有附件信息的邮件数据，因此导致处理的邮件数据长度不一致，重新添加了对于邮件数据长度的判断，程序通过。

新处理得到的Email关系数据，以一个用户为例：

CWW1120,-1.0,1.0,41547.0,0.0,0.0,0.0,0.0,

ZIN1633,1.0,9.0,38190.6666667,0.0,9.0,474093.111111,0.555555555556,

NRM1656,1.0,7.0,28107.8571429,0.0,7.0,186062.857143,0.571428571429,

用户ID表示与目标用户关联的邮件联系人，第一个特征表示发送/接收比，如果是-1则表示只有向该用户的发送没有接收。

必须重申我们现在的研究方向：通过relationships to laid off employees来分析出哪些用户的JS发生了变化。基于“奥卡姆剃刀假设”，在满足实验检测要求的基础上，越少的假设便越好，因此，这里我们首先不考虑用户初始静态人格OCEAN对于反生产行为影响的基础性影响，而是先从场景二用户的“LaidOff Employees Relationship”开始。

接下来有两种分析思路：

1. 对每个用户的邮件通讯用户集合进行聚类，目的是发现通信行为亲密程度不同的用户群簇，然后从其中依据一个正变量公式定性标记出最亲密的关系群簇；然后与用户的离职关系集合比对，按照不同组织层级，分别乘以对应系数，得到离职用户对该用户JS的影响风险，由高到低排序；
2. 第二种思路不再使用聚类，而是直接利用定性公式计算该用户邮件通讯集合每个对象的通信数据量，通过该变量标记其亲密度，然后依据亲密度由高到低对邮件通讯集合排序；最后比对离职用户关系，计算其中离职用户的加权和作为该用户的JS风险指数；

2018年10月18日星期四

下午4时41分

再次回首，发现上次的实验记录竟然是12日，上周五的了！

这周自己考出了摩托驾照，祝贺自己！

但是毕业研究才是自己的头等大事，专业技术的培养材质自己事业发展的关键，因此，自己还是需要着重回到研究层面。

即便自己毕业，即便自己可以试着托托琴姐的关系，明年努力毕业后的择业也未必乐观。

现在开始验证上周五的两个实验假设。

先开始第一个，使用一个聚类+簇类标记的方法选择出一个用户的亲密关系群簇，然后比照该群簇计算场景二30个用户的。

2018年10月19日星期五

上午9时54分

今天上午，周五，继续昨天的工作。

自己重新考虑了一次用户邮件通讯特征中反映的可以表现亲密关系的变量，可以基于以下基本假设：

* 核心假设：朋友关系的用户间存在大量邮件记录
* 推理假设：朋友间的信息流应当尽量均衡，反映为通信的信息量以及邮件次数尽量均衡

在上述假设下，我们提出一个可以标记KMeans聚类结果群簇的定性公式：

上述公式中有几点必要说明：

1. Email\_Ratio是在初始阶段提取特征时计算的，因此依据的邮件发送/接收数量为原始数量；
2. 在进行KMeans聚类前，对每个用户的七元特征进行了MinMax归一化，因此后续在表示通信信息量时，直接将数据量与附件个数进行相加；
3. RelationLevel公式中每一个变量均假设与亲密度高低成正比例关系
4. 需要重新修改邮件特征提取部分代码，更新通信比的计算
5. 分析提取的邮件通讯七元特征中，邮件size与附件个数默认都是平均值；在计算群簇的RelationLevel时，需要先计算列均值，然后再按照上述公式计算中心代表的整体通信量定性特征；

2018年10月19日星期五

下午4时29分

下午终于完成了上午的程序，继续：

首先以分析一个用户为例，即BYO1846

分析得到其邮件通讯聚类得到6个群簇，然后计算中心比较得到朋友群簇

Cluster3

***Macy\_Patterson,WAS1823,HMC1847,ESM1828,GRB1842,YDM1822,CBW1826,HDA364,WCR1830,NRR1835,ONB1833,MUP1819,JKB1843,HDA1824,ACH1831,MPK1844,Rebekah\_Santos,WLH1827,TRC1838,NJV1818,RMV1820,GFV1837,WOS1834,ACH1840,RSS1825,NZH1839,CCB1836,BGF1845,JBP1832,***

对比下该用户的离职人员关系

BYO1846,2010-12:

Insider\_LaidOff\_0,RMB1821,FDS1841,

Insider\_LaidOff\_1,NWP1609,ZAD1621,TAG1610,MAR1075,JXH1061,LSM1382,KBC1390,IVS1411,MFM1400,CTT0639,MZO1066,

Insider\_LaidOff\_2,WSW1091,JHP1654,JDM0208,HSF1115,UAM1108,CIM1095,ZHB1104,PTM1432,WDT1634,DHS0204,DDR1649,HMK0653,HMS1658,

Insider\_LaidOff\_3,FAM0495,SDL0541,CTH1812,JBG1375,LRF0549,NTV1777,GMM1037,MAF0467,XMG1579,HBH0111,KBC0818,EAL1813,WBP0828,NWH0960,BRM0126,GWH0961,HFF0560,MMB0556,BNS0484,CKP0630,JKM1790,PTV0067,MGM0539,OJC0930,KLB0918,MTD0971,TTR1792,MCP0611,VVG0624,LLW0179,RAT0514,NAH1366,PLF1030,TPO1049,RRS0056,MJA1784,EJV0094,GCB0118,

Insider\_LaidOff\_4,JAT1218,ADL1898,KSW0708,WMH1300,LAS0256,GWO1660,MIB1265,TCP0380,CBC1504,JSB0860,CDO0684,KDP1706,CDG0770,FKH0864,CLL0306,JRC1963,QSG1150,QAP0266,BAR1328,OCW1127,PKS1187,GER0350,BSS0847,OCD1985,USM0703,RKW1936,RFP1918,RDP1751,FKS1696,BMR0865,AWW0718,EJO0236,HKK0881,ESP1198,MMR1458,JIB1258,SCO1719,ZJN1492,ZIE0741,DTB0722,CEW1960,ILG0879,DMP0344,DEO1964,CNM0787,NBL1190,ALT1465,WHG1669,SMS0432,WFV0687,STH0353,RPJ1159,JKB0287,ELM1123,DXF1662,NAO1281,SCI0778,AYG1697,LMW0837,ICB1890,NTG1667,PCK0271,DHR1157,ZVW1475,BRG0728,HPM0360,KJG1121,JOE1672,UKM0845,KVF1143,

我们进行了LaidOff与Friends群簇的交叉验证，得到了下述结果：

LaidOff Friends is

0 ['RMB1821', 'FDS1841']

1 ['KBC1390']

2 []

3 ['BRM0126', 'OJC0930']

4 ['TCP0380', 'DTB0722', 'RPJ1159', 'ZVW1475']

即只有第2号邮件联系群簇在离职人员中没有出现

2018年10月22日星期一

下午2时3分

今天周一，继续上周五的实验结果。

上周五最后的实验表明，自己对BYO用户的邮件通讯表进行了KMeans，得到的群簇通过中心标记法选出了最有可能的Friends群簇，然而该群簇与离职用户的交集却为0。今天我们继续该实验的分析。

上次得到的BYO用户邮件联系者的群簇为

Cluster0

Carol\_Copeland,Juarez.Isaac,Lane-Kylie,Lawrence-Stuart,ACS951,RIK0629,CKM0939,KJH0475,Nora\_Blevins,OWS0279,BLC1263,Ochoa-Albert,Myra.M.Morrow,Jackson.B.Moody,PDH1180,CZK1473,JJF1167,HBO999,Ashely\_H\_Nguyen,Bryant-Upton,ZVW1475,BRM0126,ZWS0755,Justine.R.Skinner,AAC683,ZHB7,IAH5891,Carol\_Gomez,Melissa.Q.Short,Vivian\_Douglas,SRC586,Aileen\_Munoz,Abra\_M\_Campos,PWB1739,QZH57,DJP73,Florence\_N\_Lawson,QKB1893,Butler-Kadeem,KNH588,Myra\_M\_Morrow,BDM178,BTR1410,TCP0380,XAB588,Chastity\_Bonner,Carol\_S\_Copeland,Zena\_A\_Duncan,CLD1435,SRP1576,BHO3872,Mcgee\_Kirestin,IJM621,MAF1623,ARM0805,HBO3,IJM748,Vanna-Fowler,KRC0098,LKM19,Hilda\_Moran,YBD1929,BSD1735,Blair\_Elvis,HAC1527,SJP1023,SEO0946,Laura\_A\_Bowman,RKF0080,Aileen\_R\_Munoz,KLG0576,KBC1390,CWM0618,Ivory\_Mejia,BAR1640,Lacota.H.Livingston,Kirestin-Mcgee,MGC0646,Bruce.G.Kelley,Deirdre\_Clayton,KJA0347,ZWS866,BCG37,SJA57,HTH1001,INW1064,LIJ1581,DAR0139,CEG1467,Webb-Lenore,Willa\_Bullock,LHD1750,Cash-Christopher,YLH0423,LTB0308,CLF1981,IAJ1729,Charde.L.Farmer,Ulric.B.Juarez,JDS115,OTB0008,ISK0116,DDS1090,Cynthia\_Blanchard,Moana.A.Weiss,KRC0842,RPJ1159,Gwendolyn\_Eaton,Adam.C.Haney,Hall.L.Gay,ICH1810,HBW0683,JAW1429,DHR4378,ZUR99,OLM1747,Oneill-Bernard,DTB0722,RCD0299,Eve\_Powell,QEL1515,YNC1336,TCA1183,RJM1,DAL0795,ASH0458,Jayme\_Mcguire,YSH1997,ERM1590,Nora.A.Booth,RLF0472,CIM735,Odette\_Richardson,Lara\_C\_Rich,Cummings-Matthew,Tamekah.C.Cochran,AUH0138,Blair-Tarik,AGB0447,NLC1113,AJR817,PJB0886,HRT0984,RAV0088,Neil\_Dickerson,JEC1427,CEB1879,ACT1903,GRB56,EMM0641,YEG1439,YWF1562,INC0091,FAD1863,Brock.B.Weaver,GTF1464,TSJ0655,JGM1269,OFA1415,FWT1586,BCR0747,CHL0953,JIT0356,KVP1424,XAW13,XRS0498,Tana\_Curry,Florence.N.Lawson,KVC522,Maggie-Keith,Ashely\_Wilkerson,KMO0382,Shannon\_Quin,IAM6518,AGB0901,TAK0501,DCC1119,CAB1189,BCA1517,Tamekah.W.Brady,KHG0205,MNE1698,Vera\_Luna,ABL1229,HRR1154,Moreno-Ann,Teagan.Abbott,Celeste\_Bright,Frost\_Taylor,Morgan\_Dixon,Ariel.L.Barr,Nevada.H.Hood,KJM1303,CPW1,CBD1128,JAC0721,EYM2474,TCC0311,MMM1655,SDH1854,DXB1791,GGH991,HNR0874,Bell.S.Willis,ARF0719,

Cluster1

Macy\_Patterson,WAS1823,HMC1847,ESM1828,GRB1842,XKB1829,YDM1822,Haney-Aaron,CBW1826,Chiquita\_Burns,Micah-King,HDA364,Rose.M.Blackwell,Donna\_Black,YDM6,Ava\_Hebert,Jonathan\_Buck,Fatima\_Santana,WAS1,William.O.Sykes,GRB3,WCR1830,Hannah\_M\_Callahan,NRR1835,Hammond-Nelle,ONB1833,MUP1819,JKB1843,HDA1824,ACH1831,MPK1844,Rebekah\_Santos,WLH1827,TRC1838,NJV1818,DEB1767,Jemima-Pratt,Eugenia.S.Mercado,RMV1820,GFV1837,WOS1834,ACH1840,RSS1825,NZH1839,CCB1836,BGF1845,WCR66,Carly\_B\_Witt,Roman-Nelle,Olympia.N.Bonner,WLH7557,RMV674,Belle\_Frost,JBP1832,

Cluster2

FDS1841,RMB1821,Germane\_Velez,Vasquez-Noah,XKB1,Curry-Tana,Puckett-Regan,Salazar-Jordan,YVJ641,BAG995,Orr\_Arthur,Keelie.I.Sargent,MIC54,CLN0061,HBP0009,Osborn-Jeanette,GCH0470,DOR0935,HYA23,ENB1617,JKC1522,Baker-Thaddeus,CLG236,Morgan\_A\_Fowler,YLR1188,CBN45,Small\_Christopher,Ferris-Horton,KVC1487,AAC0904,HMD1505,MES0966,Gilbert-Christopher,PAP72,HXP0976,Cassady.R.Daniels,AJS1408,ZAR0235,RNC1295,HJB0858,NFH0677,HBP9842,DZA0195,Shelly.L.May,Hasad.S.Knowles,DJC0137,Combs-Louis,CTK0406,DSB0530,TTG0460,DOT0144,YNO0103,Camden-Powell,Stuart.Keefe,CKB0245,UMB0310,HTM1498,Cortez-Harding,KUB1569,Jaden.A.Waters,JJS0013,JJA1164,ERB0104,CTP0071,SWC1392,Alma\_S\_Carpenter,Sawyer-Marcia,Ulla.M.Henry,Madaline\_Tillman,PEM0277,DLM2,CEM1385,Mara\_Farrell,MAB1340,YIN0342,ROB6,JBI25,RMK1771,SFJ0856,HTM57,TAB0519,HBP1076,VCM0992,RDB0546,RRC0891,PHG1778,CAG1416,DPC367,MYB0686,AEG0962,JRB0759,WPL0086,CCW5426,GPS29,Jael.I.Griffith,DEC1939,RDW1710,SJC1563,GCE1147,Wesley-Fletcher,WWW0701,UAB0534,Althea\_M\_Berger,Alice\_Callahan,TDC1086,DZH1867,Sage.E.Ortega,Bree\_Mcclure,Darryl\_J\_Hays,BDB1111,JAC1275,ABM0890,Burt\_Giselle,Conrad-Gannon,Conner-Clayton,BRM1080,WAH1774,MSK0117,BWC0509,HLH9961,Finley-Lynn,BDP0096,ROB0477,Brenda\_P\_Allison,SJA0635,Ursa\_S\_Wood,DDK0995,Aphrodite\_Macias,JMW0038,Brenda\_O\_Carver,CPL0439,ZEP0543,Oliver-Declan,EZB0925,MVM0092,Chavez-Sylvester,ZOF1559,GMB0400,UMB0322,GCD0194,ACB1087,CSE0417,AAW0952,MDR1497,IBR0131,EOG0433,FAJ1122,ENB4947,WDS1286,Jane-Buchanan,BEM1501,

Cluster3

Tatyana\_A\_Johnston,Tana\_P\_Orr,Indigo\_Christensen,Lara.Y.Lopez,OJC0930,Jane.A.Griffin,Wong-Flynn,IKB0691,DLM1699,JQS1350,DJT1534,DES1617,

Cluster4

DLH0679,JBI1134,Colon-Upton,DEM0018,

上述结果中初步的分析发现，同一团队离职人员位于同一个联系者群簇内，继续查看：

我们进行了LaidOff与Friends群簇的交叉验证，得到了下述结果：

LaidOff Friends is

0 ['RMB1821', 'FDS1841'] both in Cluster 2

1 ['KBC1390'] in Cluster 0

2 []

3 ['BRM0126', 'OJC0930'] in Cluster 0 and Cluster 3

4 ['TCP0380', 'DTB0722', 'RPJ1159', 'ZVW1475'] all four in Cluster 0

统计：

Cluster 0: 6;

Cluster 2: 2

Cluster 3: 1

初步的分析可以得到一种改进猜测：

即

当使用“中心标记法”选择群簇时，需要同时具备两点要求：

1. 中心标记法的定性表示应符合同方向变化；
2. 由于是定性选择，因此应按照“二分法”选择，即选择群簇中心定性表示变量的中位数（包含）的群簇，如果K是偶数，则选择了K/2个群簇作为Friends；如果K是个奇数，则选择了K/2 + 1个群簇

于是我们准备重新进行一次改进后的实验，主要改进有：

* 改进“中心标记法”，采用“二分法”进行选择
* 依据“二分法”交叉匹配的离职朋友计算其JS\_Risk指标；
* 在分析用户邮件时去掉了非规范邮件名，因为数据集说明中企业员工全部使用规范邮件名通讯，对于分析离职前朋友影响而言，非企业联系无需考虑；

重新计算后得到

Cluster0

FDS1841,RMB1821,CLN0061,HBP0009,DLH0679,GCH0470,DOR0935,ENB1617,JKC1522,YLR1188,JBI1134,KVC1487,AAC0904,HMD1505,MES0966,HXP0976,AJS1408,ZAR0235,RNC1295,HJB0858,NFH0677,HBP9842,DZA0195,DJC0137,CTK0406,DSB0530,TTG0460,DOT0144,YNO0103,CKB0245,UMB0310,HTM1498,KUB1569,JJS0013,JJA1164,ERB0104,CTP0071,SWC1392,PEM0277,CEM1385,MAB1340,YIN0342,RMK1771,SFJ0856,TAB0519,HBP1076,VCM0992,RDB0546,RRC0891,PHG1778,CAG1416,MYB0686,AEG0962,JRB0759,WPL0086,DEM0018,CCW5426,DEC1939,RDW1710,SJC1563,GCE1147,WWW0701,UAB0534,TDC1086,DZH1867,BDB1111,JAC1275,ABM0890,BRM1080,WAH1774,MSK0117,BWC0509,HLH9961,BDP0096,ROB0477,SJA0635,DDK0995,JMW0038,CPL0439,ZEP0543,EZB0925,MVM0092,ZOF1559,GMB0400,UMB0322,GCD0194,ACB1087,CSE0417,AAW0952,MDR1497,IBR0131,EOG0433,FAJ1122,ENB4947,WDS1286,BEM1501,

Cluster1

RIK0629,CKM0939,KJH0475,OWS0279,BLC1263,PDH1180,CZK1473,JJF1167,ZVW1475,BRM0126,ZWS0755,IAH5891,PWB1739,QKB1893,BTR1410,TCP0380,CLD1435,SRP1576,BHO3872,MAF1623,ARM0805,KRC0098,YBD1929,BSD1735,OJC0930,HAC1527,SJP1023,SEO0946,RKF0080,KLG0576,KBC1390,CWM0618,BAR1640,MGC0646,KJA0347,HTH1001,INW1064,LIJ1581,DAR0139,CEG1467,LHD1750,YLH0423,LTB0308,CLF1981,IAJ1729,OTB0008,ISK0116,DDS1090,KRC0842,RPJ1159,ICH1810,HBW0683,JAW1429,DHR4378,OLM1747,DTB0722,RCD0299,QEL1515,YNC1336,TCA1183,DAL0795,ASH0458,YSH1997,ERM1590,RLF0472,AUH0138,AGB0447,NLC1113,PJB0886,HRT0984,RAV0088,JEC1427,CEB1879,ACT1903,EMM0641,YEG1439,YWF1562,INC0091,FAD1863,GTF1464,TSJ0655,JGM1269,OFA1415,FWT1586,BCR0747,CHL0953,JIT0356,KVP1424,XRS0498,KMO0382,IAM6518,AGB0901,TAK0501,DCC1119,CAB1189,BCA1517,KHG0205,MNE1698,ABL1229,HRR1154,KJM1303,CBD1128,JAC0721,EYM2474,TCC0311,MMM1655,SDH1854,DXB1791,HNR0874,ARF0719,

Cluster2

WAS1823,HMC1847,ESM1828,GRB1842,XKB1829,YDM1822,CBW1826,WCR1830,NRR1835,ONB1833,MUP1819,JKB1843,HDA1824,ACH1831,MPK1844,WLH1827,TRC1838,NJV1818,DEB1767,RMV1820,GFV1837,WOS1834,ACH1840,RSS1825,NZH1839,CCB1836,BGF1845,WLH7557,JBP1832,

Cluster3

IKB0691,DLM1699,JQS1350,DJT1534,DES1617,

Friends is

Cluster2

WAS1823,HMC1847,ESM1828,GRB1842,XKB1829,YDM1822,CBW1826,WCR1830,NRR1835,ONB1833,MUP1819,JKB1843,HDA1824,ACH1831,MPK1844,WLH1827,TRC1838,NJV1818,DEB1767,RMV1820,GFV1837,WOS1834,ACH1840,RSS1825,NZH1839,CCB1836,BGF1845,WLH7557,JBP1832,

Cluster3

IKB0691,DLM1699,JQS1350,DJT1534,DES1617,

0 ['RMB1821', 'FDS1841'] both in Cluster 0

1 ['KBC1390']

2 []

3 ['BRM0126', 'OJC0930']

4 ['TCP0380', 'DTB0722', 'RPJ1159', 'ZVW1475']

2018年10月22日星期一

下午4时50分

针对BYO1846的新分析结果

自动划分为4个群簇，且每个群簇中心点指标为：

Cluster\_RL is [1.569104524037839e-07, 3.0128794257443135e-07, 0.0005568606666948341, 0.00010915803161659686]

Cluster 2: 0.000556860666695

Cluster 3: 0.000109158031617

而将BYO1846邮件通讯用户与其周围离职用户交叉比对，发现以下用户

注意：由于这部分源于组织离职数据以及BYO1846的邮件通讯列表用户，因此结果与KMeans聚类结果无关

LaidOff Friends is

0 ['RMB1821', 'FDS1841'] both in Cluster 0

1 ['KBC1390'] in Cluster 1

2 []

3 ['BRM0126', 'OJC0930'] both in Cluster 1

4 ['TCP0380', 'DTB0722', 'RPJ1159', 'ZVW1475'] all in Cluster 1

我们发现，Friends选择的类别完美地避开了所有的匹配用户：

KBC1390,-1.0,1.0,34149.0,0.0,0.0,0.0,0.0,

BRM0126,-1.0,1.0,23259.0,0.0,0.0,0.0,0.0,

OJC0930,-1.0,1.0,3331657.0,3.0,0.0,0.0,0.0,

TCP0380,-1.0,1.0,35257.0,0.0,0.0,0.0,0.0,

DTB0722,-1.0,1.0,25142.0,0.0,0.0,0.0,0.0,

RPJ1159,-1.0,1.0,32813.0,0.0,0.0,0.0,0.0,

ZVW1475,-1.0,1.0,28742.0,0.0,0.0,0.0,0.0,

对于离职的员工来说：

TRC1838,-0.269841269841,23.0,273414.782609,0.304347826087,40.0,30483.5,0.0,（数据相当高）

感觉不应该排除掉Insiders用户，而是全部考虑在内：

1. 如果离职用户是被解雇，那么好友关系会引发本人不满；
2. 如果离职用户是主动离职，也会促使本人考虑离职寻求更好待遇；

如果使用Leave数据，则有通信聚类，朋友选择了1和3群簇

Cluster0

RIK0629,CKM0939,KJH0475,OWS0279,BLC1263,PDH1180,CZK1473,JJF1167,ZVW1475,BRM0126,ZWS0755,IAH5891,PWB1739,QKB1893,BTR1410,TCP0380,CLD1435,SRP1576,BHO3872,MAF1623,ARM0805,KRC0098,YBD1929,BSD1735,OJC0930,HAC1527,SJP1023,SEO0946,RKF0080,KLG0576,KBC1390,CWM0618,BAR1640,MGC0646,KJA0347,HTH1001,INW1064,LIJ1581,DAR0139,CEG1467,LHD1750,YLH0423,LTB0308,CLF1981,IAJ1729,OTB0008,ISK0116,DDS1090,KRC0842,RPJ1159,ICH1810,HBW0683,JAW1429,DHR4378,OLM1747,DTB0722,RCD0299,QEL1515,YNC1336,TCA1183,DAL0795,ASH0458,YSH1997,ERM1590,RLF0472,AUH0138,AGB0447,NLC1113,PJB0886,HRT0984,RAV0088,JEC1427,CEB1879,ACT1903,EMM0641,YEG1439,YWF1562,INC0091,FAD1863,GTF1464,TSJ0655,JGM1269,OFA1415,FWT1586,BCR0747,CHL0953,JIT0356,KVP1424,XRS0498,KMO0382,IAM6518,AGB0901,TAK0501,DCC1119,CAB1189,BCA1517,KHG0205,MNE1698,ABL1229,HRR1154,KJM1303,CBD1128,JAC0721,EYM2474,TCC0311,MMM1655,SDH1854,DXB1791,HNR0874,ARF0719,

Cluster1

WAS1823,HMC1847,ESM1828,GRB1842,XKB1829,YDM1822,CBW1826,WCR1830,NRR1835,ONB1833,MUP1819,JKB1843,HDA1824,ACH1831,MPK1844,WLH1827,TRC1838,NJV1818,DEB1767,RMV1820,GFV1837,WOS1834,ACH1840,RSS1825,NZH1839,CCB1836,BGF1845,WLH7557,JBP1832,

Cluster2

FDS1841,RMB1821,CLN0061,HBP0009,DLH0679,GCH0470,DOR0935,ENB1617,JKC1522,YLR1188,JBI1134,KVC1487,AAC0904,HMD1505,MES0966,HXP0976,AJS1408,ZAR0235,RNC1295,HJB0858,NFH0677,HBP9842,DZA0195,DJC0137,CTK0406,DSB0530,TTG0460,DOT0144,YNO0103,CKB0245,UMB0310,HTM1498,KUB1569,JJS0013,JJA1164,ERB0104,CTP0071,SWC1392,PEM0277,CEM1385,MAB1340,YIN0342,RMK1771,SFJ0856,TAB0519,HBP1076,VCM0992,RDB0546,RRC0891,PHG1778,CAG1416,MYB0686,AEG0962,JRB0759,WPL0086,DEM0018,CCW5426,DEC1939,RDW1710,SJC1563,GCE1147,WWW0701,UAB0534,TDC1086,DZH1867,BDB1111,JAC1275,ABM0890,BRM1080,WAH1774,MSK0117,BWC0509,HLH9961,BDP0096,ROB0477,SJA0635,DDK0995,JMW0038,CPL0439,ZEP0543,EZB0925,MVM0092,ZOF1559,GMB0400,UMB0322,GCD0194,ACB1087,CSE0417,AAW0952,MDR1497,IBR0131,EOG0433,FAJ1122,ENB4947,WDS1286,BEM1501,

Cluster3

IKB0691,DLM1699,JQS1350,DJT1534,DES1617,

而此时如果直接检验交叉匹配

0 ['RMB1821', 'FDS1841', 'TRC1838']

1 ['KBC1390']

2 []

3 ['BRM0126', 'OJC0930', 'HXP0976']

4 ['TCP0380', 'DTB0722', 'RPJ1159', 'ZVW1475']

BYO1846离职员工列表

BYO1846,2010-12:

Insider\_LaidOff\_0,RMB1821,FDS1841,

Insider\_LaidOff\_1,NWP1609,ZAD1621,TAG1610,MAR1075,JXH1061,LSM1382,KBC1390,IVS1411,MFM1400,CTT0639,MZO1066,

Insider\_LaidOff\_2,WSW1091,JHP1654,JDM0208,HSF1115,UAM1108,CIM1095,ZHB1104,PTM1432,WDT1634,DHS0204,DDR1649,HMK0653,HMS1658,

Insider\_LaidOff\_3,FAM0495,SDL0541,CTH1812,JBG1375,LRF0549,NTV1777,GMM1037,MAF0467,XMG1579,HBH0111,KBC0818,EAL1813,WBP0828,NWH0960,BRM0126,GWH0961,HFF0560,MMB0556,BNS0484,CKP0630,JKM1790,PTV0067,MGM0539,OJC0930,KLB0918,MTD0971,TTR1792,MCP0611,VVG0624,LLW0179,RAT0514,NAH1366,PLF1030,TPO1049,RRS0056,MJA1784,EJV0094,GCB0118,

Insider\_LaidOff\_4,JAT1218,ADL1898,KSW0708,WMH1300,LAS0256,GWO1660,MIB1265,TCP0380,CBC1504,JSB0860,CDO0684,KDP1706,CDG0770,FKH0864,CLL0306,JRC1963,QSG1150,QAP0266,BAR1328,OCW1127,PKS1187,GER0350,BSS0847,OCD1985,USM0703,RKW1936,RFP1918,RDP1751,FKS1696,BMR0865,AWW0718,EJO0236,HKK0881,ESP1198,MMR1458,JIB1258,SCO1719,ZJN1492,ZIE0741,DTB0722,CEW1960,ILG0879,DMP0344,DEO1964,CNM0787,NBL1190,ALT1465,WHG1669,SMS0432,WFV0687,STH0353,RPJ1159,JKB0287,ELM1123,DXF1662,NAO1281,SCI0778,AYG1697,LMW0837,ICB1890,NTG1667,PCK0271,DHR1157,ZVW1475,BRG0728,HPM0360,KJG1121,JOE1672,UKM0845,KVF1143,

BYO1846周围所有离职员工

BYO1846,2010-12:

Insider\_LaidOff\_0,RMB1821,FDS1841,TRC1838,

Insider\_LaidOff\_1,NWP1609,ZAD1621,TAG1610,MAR1075,JXH1061,KEW0198,LSM1382,VCF1602,KBC1390,TNB1616,IVS1411,SLL0193,MFM1400,CTT0639,HIS1394,MZO1066,

Insider\_LaidOff\_2,WSW1091,JHP1654,JDM0208,HSF1115,UAM1108,CIM1095,ZHB1104,PTM1432,CIF1430,WDT1634,DHS0204,DDR1649,HMK0653,MIB0203,HMS1658,

Insider\_LaidOff\_3,FAM0495,SDL0541,CTH1812,JBG1375,LRF0549,NTV1777,GMM1037,MAF0467,XMG1579,HBH0111,KBC0818,EAL1813,WBP0828,NWH0960,BRM0126,GWH0961,HFF0560,MMB0556,BNS0484,GFM1815,CKP0630,JKM1790,PBC0077,PTV0067,MGM0539,IHC0561,OJC0930,SIS0042,KLB0918,MTD0971,TTR1792,GWG0497,MCP0611,VVG0624,LLW0179,RAT0514,GKW0043,NAH1366,PLF1030,TPO1049,ICB1354,RRS0056,MJA1784,EJV0094,GCB0118,HXP0976,

Insider\_LaidOff\_4,JAT1218,ADL1898,KSW0708,WMH1300,LAS0256,GWO1660,MIB1265,TCP0380,CBC1504,JSB0860,CDO0684,KDP1706,CDG0770,FKH0864,CLL0306,JRC1963,QSG1150,QAP0266,BAR1328,OCW1127,PKS1187,GER0350,BSS0847,OCD1985,MPF0690,USM0703,RKW1936,RFP1918,RDP1751,FKS1696,BMR0865,CRD0272,AWW0718,EJO0236,DAS1320,HKK0881,ESP1198,MMR1458,JIB1258,SCO1719,ZJN1492,SAF1942,ZIE0741,DTB0722,CEW1960,EPG1196,ILG0879,DMP0344,DEO1964,MDS0680,CNM0787,NBL1190,OSS1463,ALT1465,WHG1669,SMS0432,WFV0687,STH0353,RPJ1159,CHP1711,JKB0287,DNJ0740,ELM1123,DXF1662,NAO1281,SCI0778,ISW0738,AYG1697,LMW0837,ICB1890,NTG1667,REF1924,PCK0271,DHR1157,ZVW1475,BRG0728,HPM0360,ACA1126,KJG1121,JOE1672,UKM0845,KVF1143,

而我们此时得到的4个Cluster为：

Cluster0

FDS1841,RMB1821,CLN0061,HBP0009,DLH0679,GCH0470,DOR0935,ENB1617,JKC1522,YLR1188,JBI1134,KVC1487,AAC0904,HMD1505,MES0966,HXP0976,AJS1408,ZAR0235,RNC1295,HJB0858,NFH0677,HBP9842,DZA0195,DJC0137,CTK0406,DSB0530,TTG0460,DOT0144,YNO0103,CKB0245,UMB0310,HTM1498,KUB1569,JJS0013,JJA1164,ERB0104,CTP0071,SWC1392,PEM0277,CEM1385,MAB1340,YIN0342,RMK1771,SFJ0856,TAB0519,HBP1076,VCM0992,RDB0546,RRC0891,PHG1778,CAG1416,MYB0686,AEG0962,JRB0759,WPL0086,DEM0018,CCW5426,DEC1939,RDW1710,SJC1563,GCE1147,WWW0701,UAB0534,TDC1086,DZH1867,BDB1111,JAC1275,ABM0890,BRM1080,WAH1774,MSK0117,BWC0509,HLH9961,BDP0096,ROB0477,SJA0635,DDK0995,JMW0038,CPL0439,ZEP0543,EZB0925,MVM0092,ZOF1559,GMB0400,UMB0322,GCD0194,ACB1087,CSE0417,AAW0952,MDR1497,IBR0131,EOG0433,FAJ1122,ENB4947,WDS1286,BEM1501,

Cluster1

RIK0629,CKM0939,KJH0475,OWS0279,BLC1263,PDH1180,CZK1473,JJF1167,ZVW1475,BRM0126,ZWS0755,IAH5891,PWB1739,QKB1893,BTR1410,TCP0380,CLD1435,SRP1576,BHO3872,MAF1623,ARM0805,KRC0098,YBD1929,BSD1735,OJC0930,HAC1527,SJP1023,SEO0946,RKF0080,KLG0576,KBC1390,CWM0618,BAR1640,MGC0646,KJA0347,HTH1001,INW1064,LIJ1581,DAR0139,CEG1467,LHD1750,YLH0423,LTB0308,CLF1981,IAJ1729,OTB0008,ISK0116,DDS1090,KRC0842,RPJ1159,ICH1810,HBW0683,JAW1429,DHR4378,OLM1747,DTB0722,RCD0299,QEL1515,YNC1336,TCA1183,DAL0795,ASH0458,YSH1997,ERM1590,RLF0472,AUH0138,AGB0447,NLC1113,PJB0886,HRT0984,RAV0088,JEC1427,CEB1879,ACT1903,EMM0641,YEG1439,YWF1562,INC0091,FAD1863,GTF1464,TSJ0655,JGM1269,OFA1415,FWT1586,BCR0747,CHL0953,JIT0356,KVP1424,XRS0498,KMO0382,IAM6518,AGB0901,TAK0501,DCC1119,CAB1189,BCA1517,KHG0205,MNE1698,ABL1229,HRR1154,KJM1303,CBD1128,JAC0721,EYM2474,TCC0311,MMM1655,SDH1854,DXB1791,HNR0874,ARF0719,

Cluster2

WAS1823,HMC1847,ESM1828,GRB1842,XKB1829,YDM1822,CBW1826,WCR1830,NRR1835,ONB1833,MUP1819,JKB1843,HDA1824,ACH1831,MPK1844,WLH1827,TRC1838,NJV1818,DEB1767,RMV1820,GFV1837,WOS1834,ACH1840,RSS1825,NZH1839,CCB1836,BGF1845,WLH7557,JBP1832,

Cluster3

IKB0691,DLM1699,JQS1350,DJT1534,DES1617,

2018年10月23日星期二

上午10时15分

继续昨天的分析，我们发现了一个奇怪的现象：

前提：用户的离职关联用户关系数据与邮件通讯数据过滤结果的交叉匹配是一定的，对于BYO1846而言，如果考虑的是laid-off relationships有：

邮件联系人形成4个聚类，且LaidOff Friends is

0 ['RMB1821', 'FDS1841'] both in Cluster 0

1 ['KBC1390'] in Cluster 1

2 []

3 ['BRM0126', 'OJC0930'] both in Cluster 1

4 ['TCP0380', 'DTB0722', 'RPJ1159', 'ZVW1475'] all in Cluster 1

我们发现，Friends选择的类别完美地避开了所有的匹配用户：

KBC1390,-1.0,1.0,34149.0,0.0,0.0,0.0,0.0,

BRM0126,-1.0,1.0,23259.0,0.0,0.0,0.0,0.0,

OJC0930,-1.0,1.0,3331657.0,3.0,0.0,0.0,0.0,

TCP0380,-1.0,1.0,35257.0,0.0,0.0,0.0,0.0,

DTB0722,-1.0,1.0,25142.0,0.0,0.0,0.0,0.0,

RPJ1159,-1.0,1.0,32813.0,0.0,0.0,0.0,0.0,

ZVW1475,-1.0,1.0,28742.0,0.0,0.0,0.0,0.0,

FDS1841: FDS1841,-0.333333333333,4.0,158562.0,0.25,8.0,489182.875,0.875,

RMB1821: RMB1821,-0.428571428571,2.0,42436.0,0.0,5.0,145618.6,0.6,

Cluster-0: CLN0061,0.0,1.0,33099.0,0.0,1.0,1620201.0,2.0,

可以看出，cluster-0中的用户才是我们想要的交互频繁，却又离职的关系用户

Cluster2:

WAS1823,0.0,47.0,157937.87234,0.148936170213,47.0,118080.361702,0.170212765957,

HMC1847,-0.232323232323,38.0,203589.263158,0.184210526316,61.0,383001.163934,0.44262295082,Cluster3:

DJT1534,-1.0,1.0,4176791.0,5.0,0.0,0.0,0.0,

IKB0691,-1.0,1.0,7399054.0,9.0,0.0,0.0,0.0,

从数据上看，符合我们预期的应该是Cluster0与Cluster2两个，但是实际却选择了Cluster2与Cluster3

同样，此时得到四个群簇为：

针对BYO1846的新分析结果

自动划分为4个群簇，且每个群簇中心点指标为：

Cluster\_RL is [1.569104524037839e-07, 3.0128794257443135e-07, 0.0005568606666948341, 0.00010915803161659686]

Cluster 2: 0.000556860666695

Cluster 3: 0.000109158031617

而我们此时得到的4个Cluster为：

Cluster0

FDS1841,RMB1821,CLN0061,HBP0009,DLH0679,GCH0470,DOR0935,ENB1617,JKC1522,YLR1188,JBI1134,KVC1487,AAC0904,HMD1505,MES0966,HXP0976,AJS1408,ZAR0235,RNC1295,HJB0858,NFH0677,HBP9842,DZA0195,DJC0137,CTK0406,DSB0530,TTG0460,DOT0144,YNO0103,CKB0245,UMB0310,HTM1498,KUB1569,JJS0013,JJA1164,ERB0104,CTP0071,SWC1392,PEM0277,CEM1385,MAB1340,YIN0342,RMK1771,SFJ0856,TAB0519,HBP1076,VCM0992,RDB0546,RRC0891,PHG1778,CAG1416,MYB0686,AEG0962,JRB0759,WPL0086,DEM0018,CCW5426,DEC1939,RDW1710,SJC1563,GCE1147,WWW0701,UAB0534,TDC1086,DZH1867,BDB1111,JAC1275,ABM0890,BRM1080,WAH1774,MSK0117,BWC0509,HLH9961,BDP0096,ROB0477,SJA0635,DDK0995,JMW0038,CPL0439,ZEP0543,EZB0925,MVM0092,ZOF1559,GMB0400,UMB0322,GCD0194,ACB1087,CSE0417,AAW0952,MDR1497,IBR0131,EOG0433,FAJ1122,ENB4947,WDS1286,BEM1501,

Cluster1

RIK0629,CKM0939,KJH0475,OWS0279,BLC1263,PDH1180,CZK1473,JJF1167,ZVW1475,BRM0126,ZWS0755,IAH5891,PWB1739,QKB1893,BTR1410,TCP0380,CLD1435,SRP1576,BHO3872,MAF1623,ARM0805,KRC0098,YBD1929,BSD1735,OJC0930,HAC1527,SJP1023,SEO0946,RKF0080,KLG0576,KBC1390,CWM0618,BAR1640,MGC0646,KJA0347,HTH1001,INW1064,LIJ1581,DAR0139,CEG1467,LHD1750,YLH0423,LTB0308,CLF1981,IAJ1729,OTB0008,ISK0116,DDS1090,KRC0842,RPJ1159,ICH1810,HBW0683,JAW1429,DHR4378,OLM1747,DTB0722,RCD0299,QEL1515,YNC1336,TCA1183,DAL0795,ASH0458,YSH1997,ERM1590,RLF0472,AUH0138,AGB0447,NLC1113,PJB0886,HRT0984,RAV0088,JEC1427,CEB1879,ACT1903,EMM0641,YEG1439,YWF1562,INC0091,FAD1863,GTF1464,TSJ0655,JGM1269,OFA1415,FWT1586,BCR0747,CHL0953,JIT0356,KVP1424,XRS0498,KMO0382,IAM6518,AGB0901,TAK0501,DCC1119,CAB1189,BCA1517,KHG0205,MNE1698,ABL1229,HRR1154,KJM1303,CBD1128,JAC0721,EYM2474,TCC0311,MMM1655,SDH1854,DXB1791,HNR0874,ARF0719,

Cluster2

WAS1823,HMC1847,ESM1828,GRB1842,XKB1829,YDM1822,CBW1826,WCR1830,NRR1835,ONB1833,MUP1819,JKB1843,HDA1824,ACH1831,MPK1844,WLH1827,TRC1838,NJV1818,DEB1767,RMV1820,GFV1837,WOS1834,ACH1840,RSS1825,NZH1839,CCB1836,BGF1845,WLH7557,JBP1832,

Cluster3

IKB0691,DLM1699,JQS1350,DJT1534,DES1617,

**为了统一，重新运行了一次程序，针对BYO1846有：**

用户群簇聚类结果：4个

1. 本群簇大小为： 29 中心点坐标为： **0.607078018223716 0.7718268525311811 0.04443840214335777 0.04122093309218429 0.6964386659129452 0.03159659789503671 0.03759174755156466**
2. 本群簇大小为： 110 中心点坐标为： 0.0 0.02205029013539651 0.04130266388545403 0.03454545454545455 0.0 0.0 0.0
3. 本群簇大小为： 96 中心点坐标为： 0.6419234360408992 0.0017730496453900709 0.00033359306203451015 0.00028935185185185184 0.018954918032786903 0.051326241931948346 0.05465494791666667
4. 本群簇大小为： 5 中心点坐标为： **0.0 0.02127659574468085 0.7139472906610507 0.7777777777777777 0.0 0.0 0.0**

**BYO1846 4 选中的friends群簇标号为：**

Cluster\_RL is [**0.0005568606666948341**, 3.0128794257443135e-07, 1.569104524037839e-07, **0.00010915803161659686**]

* 0 0.000556860666695
* 3 0.000109158031617

由下述具体4个聚类结果分析，同时比照中心点的定性表示看，我们应当选择的无疑应当是第0和2号群簇，因为第1和3号群簇的接收数据部分中心值为0

------------------------------------------------------------------------------------------------------------------------------

此时KMeans聚类的四个群簇用户分别为（BYO1846的邮件通讯用户）：

**Cluster0**

WAS1823,HMC1847,ESM1828,GRB1842,XKB1829,YDM1822,CBW1826,WCR1830,NRR1835,ONB1833,MUP1819,JKB1843,HDA1824,ACH1831,MPK1844,WLH1827,TRC1838,NJV1818,DEB1767,RMV1820,GFV1837,WOS1834,ACH1840,RSS1825,NZH1839,CCB1836,BGF1845,WLH7557,JBP1832,

**Cluster1**

RIK0629,CKM0939,KJH0475,OWS0279,BLC1263,PDH1180,CZK1473,JJF1167,ZVW1475,BRM0126,ZWS0755,IAH5891,PWB1739,QKB1893,BTR1410,TCP0380,CLD1435,SRP1576,BHO3872,MAF1623,ARM0805,KRC0098,YBD1929,BSD1735,OJC0930,HAC1527,SJP1023,SEO0946,RKF0080,KLG0576,KBC1390,CWM0618,BAR1640,MGC0646,KJA0347,HTH1001,INW1064,LIJ1581,DAR0139,CEG1467,LHD1750,YLH0423,LTB0308,CLF1981,IAJ1729,OTB0008,ISK0116,DDS1090,KRC0842,RPJ1159,ICH1810,HBW0683,JAW1429,DHR4378,OLM1747,DTB0722,RCD0299,QEL1515,YNC1336,TCA1183,DAL0795,ASH0458,YSH1997,ERM1590,RLF0472,AUH0138,AGB0447,NLC1113,PJB0886,HRT0984,RAV0088,JEC1427,CEB1879,ACT1903,EMM0641,YEG1439,YWF1562,INC0091,FAD1863,GTF1464,TSJ0655,JGM1269,OFA1415,FWT1586,BCR0747,CHL0953,JIT0356,KVP1424,XRS0498,KMO0382,IAM6518,AGB0901,TAK0501,DCC1119,CAB1189,BCA1517,KHG0205,MNE1698,ABL1229,HRR1154,KJM1303,CBD1128,JAC0721,EYM2474,TCC0311,MMM1655,SDH1854,DXB1791,HNR0874,ARF0719,

**Cluster2**

FDS1841,RMB1821,CLN0061,HBP0009,DLH0679,GCH0470,DOR0935,ENB1617,JKC1522,YLR1188,JBI1134,KVC1487,AAC0904,HMD1505,MES0966,HXP0976,AJS1408,ZAR0235,RNC1295,HJB0858,NFH0677,HBP9842,DZA0195,DJC0137,CTK0406,DSB0530,TTG0460,DOT0144,YNO0103,CKB0245,UMB0310,HTM1498,KUB1569,JJS0013,JJA1164,ERB0104,CTP0071,SWC1392,PEM0277,CEM1385,MAB1340,YIN0342,RMK1771,SFJ0856,TAB0519,HBP1076,VCM0992,RDB0546,RRC0891,PHG1778,CAG1416,MYB0686,AEG0962,JRB0759,WPL0086,DEM0018,CCW5426,DEC1939,RDW1710,SJC1563,GCE1147,WWW0701,UAB0534,TDC1086,DZH1867,BDB1111,JAC1275,ABM0890,BRM1080,WAH1774,MSK0117,BWC0509,HLH9961,BDP0096,ROB0477,SJA0635,DDK0995,JMW0038,CPL0439,ZEP0543,EZB0925,MVM0092,ZOF1559,GMB0400,UMB0322,GCD0194,ACB1087,CSE0417,AAW0952,MDR1497,IBR0131,EOG0433,FAJ1122,ENB4947,WDS1286,BEM1501,

**Cluster3**

IKB0691,DLM1699,JQS1350,DJT1534,DES1617,

上午11时19分

通过回溯特征计算时EmailRatio的计算，发现修改后没有更新之前的特征计算，导致后续都是按照1/-1来表示只发送无接收和无发送只接收的单向邮件行为；

故重新修改后运行程序

---------------------------------------------------------------------------------------------------------------------------

**BYO1846邮件联系人的聚类中心指标**

..<<开始计算各个群簇中心点坐标，并依据公式计算其对应的RelationLevel定性指标>>..

1. 本群簇大小为： 29 中心点坐标为： *0.4691057413548276 0.7718268525311811 0.04443840214335777 0.04122093309218429 0.6964386659129452 0.03159659789503671 0.03759174755156466*
2. 本群簇大小为： 110 中心点坐标为： 1.0 0.02205029013539651 0.04130266388545403 0.03454545454545455 0.0 0.0 0.0
3. 本群簇大小为： 96 中心点坐标为：  *0.016865079365083333 0.0017730496453900709 0.00033359306203451015 0.00028935185185185184 0.018954918032786903 0.051326241931948346 0.05465494791666667*
4. 本群簇大小为： 5 中心点坐标为： 1.0 0.02127659574468085 0.7139472906610507 0.7777777777777777 0.0 0.0 0.0

**选中的亲密群簇中心为：**

BYO1846 4 选中的friends群簇标号为：

Cluster\_RL is [0.0007522298043701955, 0.0, 4.308131612704546e-07, 0.0]

* **0 0.00075222980437**
* **2 4.3081316127e-07**

此时邮件联系人聚类结果为，终于选中了交互符合我们预期的两个群簇

Cluster 0 与 Cluster 1

**Cluster0**

WAS1823,HMC1847,ESM1828,GRB1842,XKB1829,YDM1822,CBW1826,WCR1830,NRR1835,ONB1833,MUP1819,JKB1843,HDA1824,ACH1831,MPK1844,WLH1827,TRC1838,NJV1818,DEB1767,RMV1820,GFV1837,WOS1834,ACH1840,RSS1825,NZH1839,CCB1836,BGF1845,WLH7557,JBP1832,

**Cluster1**

RIK0629,CKM0939,KJH0475,OWS0279,BLC1263,PDH1180,CZK1473,JJF1167,ZVW1475,BRM0126,ZWS0755,IAH5891,PWB1739,QKB1893,BTR1410,TCP0380,CLD1435,SRP1576,BHO3872,MAF1623,ARM0805,KRC0098,YBD1929,BSD1735,OJC0930,HAC1527,SJP1023,SEO0946,RKF0080,KLG0576,KBC1390,CWM0618,BAR1640,MGC0646,KJA0347,HTH1001,INW1064,LIJ1581,DAR0139,CEG1467,LHD1750,YLH0423,LTB0308,CLF1981,IAJ1729,OTB0008,ISK0116,DDS1090,KRC0842,RPJ1159,ICH1810,HBW0683,JAW1429,DHR4378,OLM1747,DTB0722,RCD0299,QEL1515,YNC1336,TCA1183,DAL0795,ASH0458,YSH1997,ERM1590,RLF0472,AUH0138,AGB0447,NLC1113,PJB0886,HRT0984,RAV0088,JEC1427,CEB1879,ACT1903,EMM0641,YEG1439,YWF1562,INC0091,FAD1863,GTF1464,TSJ0655,JGM1269,OFA1415,FWT1586,BCR0747,CHL0953,JIT0356,KVP1424,XRS0498,KMO0382,IAM6518,AGB0901,TAK0501,DCC1119,CAB1189,BCA1517,KHG0205,MNE1698,ABL1229,HRR1154,KJM1303,CBD1128,JAC0721,EYM2474,TCC0311,MMM1655,SDH1854,DXB1791,HNR0874,ARF0719,

**Cluster2**

FDS1841,RMB1821,CLN0061,HBP0009,DLH0679,GCH0470,DOR0935,ENB1617,JKC1522,YLR1188,JBI1134,KVC1487,AAC0904,HMD1505,MES0966,HXP0976,AJS1408,ZAR0235,RNC1295,HJB0858,NFH0677,HBP9842,DZA0195,DJC0137,CTK0406,DSB0530,TTG0460,DOT0144,YNO0103,CKB0245,UMB0310,HTM1498,KUB1569,JJS0013,JJA1164,ERB0104,CTP0071,SWC1392,PEM0277,CEM1385,MAB1340,YIN0342,RMK1771,SFJ0856,TAB0519,HBP1076,VCM0992,RDB0546,RRC0891,PHG1778,CAG1416,MYB0686,AEG0962,JRB0759,WPL0086,DEM0018,CCW5426,DEC1939,RDW1710,SJC1563,GCE1147,WWW0701,UAB0534,TDC1086,DZH1867,BDB1111,JAC1275,ABM0890,BRM1080,WAH1774,MSK0117,BWC0509,HLH9961,BDP0096,ROB0477,SJA0635,DDK0995,JMW0038,CPL0439,ZEP0543,EZB0925,MVM0092,ZOF1559,GMB0400,UMB0322,GCD0194,ACB1087,CSE0417,AAW0952,MDR1497,IBR0131,EOG0433,FAJ1122,ENB4947,WDS1286,BEM1501,

**Cluster3**

IKB0691,DLM1699,JQS1350,DJT1534,DES1617,

2018年10月23日星期二

下午2时46分

上午修改了部分的代码，发现出现了嵌套列表计算时直接引用子列表计算的错误。然后重新计算出CERT5.2中场景二30个用户的JS\_Risk

JS\_Risks for CERT5.2

BYO1846,0.6

CHP1711,0.0

CIF1430,0.0

CKP0630,0.0

DCC1119,0.8

GWG0497,0.0

HIS1394,4.6

HMS1658,1.8

HSN0675,2.8

HXP0976,2.0

ICB1354,0.8

ITA0159,0.0

JAL0811,0.0

KSS1005,0.2

LVF1626,0.2

MCP0611,0.0

MDS0680,1.2

MGB1235,0.0

NAH1366,1.2

OKM1092,5.8

OSS1463,0.0

RRS0056,0.0

SIS0042,0.0

SNK1280,1.4

TMT0851,5.0

TNB1616,0.2

TRC1838,0.0

VCF1602,0.0

WDT1634,0.0

ZIE0741,0.6

初步分析发现，按照现有的方法计算，竟然有16个用户的JS\_Risk为0！这不符合我们之前的预期，因此需要重新审视下JS\_Risk为0的用户。

以CHP1711,0.0为例，发现竟然是为组织层次赋予权值时颠倒了顺序，应该顺序为：[1, 0.8, 0.6, 0.4, 0.2]， 而非[0.0, 0.2, 0.4, 0.6, 0.8]，如此重新计算即可。

而且发现了CHP1711用户也是同一团队有交往密切的两个用户离职。

重新计算后，发现新的结果： 30个Insiders中有9个用户的JS\_Risk为0

JS\_Risks for CERT5.2

BYO1846,3.4

CHP1711,2.0

CIF1430,1.0

CKP0630,0.2

DCC1119,4.2

GWG0497,0.0

HIS1394,5.4

HMS1658,2.0

HSN0675,2.2

HXP0976,1.0

ICB1354,1.2

ITA0159,2.0

JAL0811,2.0

KSS1005,0.8

LVF1626,4.8

MCP0611,0.0

MDS0680,0.0

MGB1235,0.0

NAH1366,0.8

OKM1092,6.2

OSS1463,0.0

RRS0056,0.0

SIS0042,0.0

SNK1280,1.6

TMT0851,6.0

TNB1616,4.8

TRC1838,0.0

VCF1602,3.0

WDT1634,0.0

ZIE0741,0.4

接下来我们重点分析下JS\_Risk为0的Insiders，就先以用户GWG0497,0.0为例子

**分析其离职与朋友的关系数据：**

LaidOff Friends is

0 ['FAM0495'] Cluster 1

1 ['PBC0077'] Cluster 1

2 ['MMB0556', 'IHC0561'] Cluster 1, Cluster 1

3 ['WBP0828', 'ZHB1104', 'IVS1411', 'MFM1400'] Cluster 0, Cluster0, Cluster 0, Cluster 1

4 ['LAS0256', 'OCD1985', 'FKS1696', 'BMR0865', 'AWW0718'] Cluster 0, Cluster 1, Cluster 0, Cluster 1, Cluster 1

可以发现，主要的交叉匹配发生在Cluster 0 与 Cluster 1之间，因此，继续检查该群簇的用户通讯特征

FAM0495,-1.0,0.0,0.0,0.0,1.0,31852.0,0.0, Cluster 1

PBC0077,-1.0,0.0,0.0,0.0,1.0,21867.0,0.0, Cluster 1

WBP0828,1.0,1.0,33971.0,0.0,0.0,0.0,0.0, Cluster 0

注意到，这里仅仅表示出了单向通讯，

我们猜测的一种可能是：

也许收发均衡，通讯的信息量是独立的定性标志，因此不能用log(A\*B\*C)的形式，而应采用log(A + B + C)的形式

**选择的亲密群簇以及所有群簇的中心标记为：**

*GWG0497* 10 选中的friends群簇标号为：

Cluster\_RL is [0.0, 0.0, 9.464493549532657e-05, 7.667337003100332e-05, 0.0, 0.004821449737770455, 8.372550716876508e-06, 2.1522034206294337e-05, 0.0006847606968232683, 0.0024174486120135544]

**5 0.00482144973777**

**9 0.00241744861201**

**8 0.000684760696823**

**2 9.46449354953e-05**

**3 7.6673370031e-05**

**该用户的离职人员关系为Leave：**

**GWG0497,2010-10:**

Insider\_LaidOff\_0,FAM0495,

Insider\_LaidOff\_1,NWH0960,GWH0961,BNS0484,PBC0077,MTD0971,

Insider\_LaidOff\_2,SDL0541,LRF0549,NTV1777,HBH0111,BRM0126,HFF0560,MMB0556,PTV0067,MGM0539,IHC0561,OJC0930,

Insider\_LaidOff\_3,RMB1821,WSW1091,CTH1812,JBG1375,JHP1654,NWP1609,JDM0208,HSF1115,FDS1841,GMM1037,MAF0467,ZAD1621,XMG1579,KBC0818,TAG1610,EAL1813,WBP0828,MAR1075,JXH1061,UAM1108,KEW0198,LSM1382,GFM1815,CIM1095,VCF1602,CKP0630,JKM1790,KBC1390,ZHB1104,PTM1432,SIS0042,CIF1430,KLB0918,TRC1838,TNB1616,IVS1411,WDT1634,DHS0204,SLL0193,DDR1649,MFM1400,TTR1792,MCP0611,

Insider\_LaidOff\_4,JAT1218,ADL1898,KSW0708,WMH1300,LAS0256,GWO1660,MIB1265,TCP0380,CBC1504,JSB0860,CDO0684,KDP1706,CDG0770,FKH0864,CLL0306,JRC1963,QSG1150,QAP0266,BAR1328,OCW1127,PKS1187,GER0350,BSS0847,OCD1985,MPF0690,USM0703,RKW1936,RFP1918,RDP1751,FKS1696,BMR0865,CRD0272,AWW0718,EJO0236,DAS1320,HKK0881,ESP1198,MMR1458,JIB1258,SCO1719,ZJN1492,SAF1942,ZIE0741,DTB0722,CEW1960,EPG1196,ILG0879,DMP0344,DEO1964,MDS0680,CNM0787,NBL1190,OSS1463,ALT1465,WHG1669,SMS0432,WFV0687,STH0353,RPJ1159,CHP1711,JKB0287,DNJ0740,ELM1123,

而该用户对应的10个邮件通讯群簇为：

**Cluster0**

QJF0725,CLN0061,KWM0365,HMK0653,CRD0987,IDC1775,JFC0075,LDM0399,ZHP0989,ASR0150,SJG0228,HBS1299,MAM1293,LAS0256,JJP0798,HWS1772,TPO1049,MXM0184,KRM1306,NCC0454,AMC1724,UHM1273,HTH1001,HRR0079,QGB0554,JDR1106,FKS1696,MSS0799,ZHB1104,SSG0726,IVS1411,HRT0984,MCT0695,AGB0447,CDD1124,NZL1395,KJH0804,FMH1018,AMG1628,BDP1162,LSO1783,KEM1211,TAJ0023,WBP0828,DKS0563,DEM0018,EWP1316,GAT1343,CBS1856,SGH0834,MRH0407,SFA1912,PTH1889,EAP0312,ILM1441,GJC0870,TME0887,ATO0307,TSG0262,TRP0892,RMK1771,HOS1349,DSH1074,MHP1377,GNV1082,FAM0211,ISS0508,RSB1054,PDH1180,KHG0205,ZUR1449,

**Cluster1**

SEO0946,MHM1645,WIW1212,EZB0925,SKV0041,IHC0561,FAM0495,NFH0677,LDF0644,DAD0331,SME1271,HMC1847,MMB0556,PAA0121,LMG0065,BTD0296,AWW0718,NLC1113,JRB0759,CRG0063,NMF1088,JDJ1949,TMT0851,QZH0367,RHB1193,OCD1985,DAE1219,MOW1888,LVC0408,UJC0329,BHM1707,EAC0769,PWF0709,RXH1139,OHR0583,ACM1770,JTF0313,BMR0865,ESM1828,ACC0950,JOR1495,AKD0332,PBC0077,EPB0893,BLK0698,JSB0803,TAW0689,HRC0977,ASC1297,KRS0172,HNJ1407,RHM0775,JCC0016,MFM1400,DSB0530,DAC3768,LJH1027,SMW1131,FCG0693,LKP1714,CSB0327,RJW0348,

**Cluster2**

RHD1155,HST0084,EZG9638,JMA0540,HTR0239,PRL0585,ESW0928,BBB0012,NKH0704,BHG1947,CLZ1937,FMP0459,KAC1209,DPC1618,

**Cluster3**

CDL0189,CGS0309,KCM0402,EZG0968,MAB0500,ICW0083,RDM0489,TAK0501,ABL1229,MTS0958,JDB1163,WRH0526,SKC0074,CLL0494,ANR0502,FYC0060,AHH0085,XRS0498,AKP0503,ZAC0499,ICW6286,WCR1830,

**Cluster4**

RAA0263,DCD0715,AKS0586,SJC0232,JRH0993,EXB0090,HNJ0638,

**Cluster5**

NMV0716,JCB0791,AFC0792,EBG1553,JLC0742,SEC0985,

**Cluster6**

MHS1277,MCS0242,KOL1923,TDC0800,

**Cluster7**

BMC0650,HQG0905,QKB0562,

**Cluster8**

CAD0354,UTM0493,RMT0496,

**Cluster9**

MMM0082,AFC9839,

考虑到可能是单向通讯导致的信息缺失，略微修改了原先的中心标记公式：

RL = math.log10(1 + (2 - abs(sr)) \* (s1 \* s2 + r1 \* r2) \* (s1 \* s3 + r1 \* r3))

通过将1-->2-abs(sr)，使得该部分最小值为1

此时依然Kmeans得到10个群簇

Cluster\_RL is [**0.001950735264600607**, 1.9035617310074998e-08, 0.0, **0.014522260955346457**, 3.9487632622008215e-05, **0.00016308030661314258**, 7.400436803929033e-06, 0.0023406954082099743, **4.30430019127308e-05**, 1.6744940026558036e-05]

**3 0.0145222609553**

**7 0.00234069540821**

**0 0.0019507352646**

**5 0.000163080306613**

**8 4.30430019127e-05**

**FAM0495,-1.0,0.0,0.0,0.0,1.0,31852.0,0.0,**

此时再分析离职与通讯的交叉匹配

LaidOff Friends is

0 ['FAM0495'] Cluster 2

1 ['PBC0077'] Cluster 2

2 ['MMB0556', 'IHC0561'] Cluster 2, Cluster 2

3 ['WBP0828', 'ZHB1104', 'IVS1411', 'MFM1400'] Cluster 1, Cluster1, Cluster 1, Cluster 2

4 ['LAS0256', 'OCD1985', 'FKS1696', 'BMR0865', 'AWW0718'] Cluster 1, Cluster 2, Cluster 1, Cluster

继续调整公式

RL = math.log10(1 + (1 - abs(sr)) + (s1 \* s2 + r1 \* r2) + (s1 \* s3 + r1 \* r3))

Cluster\_RL is [0.29310058570888764, 7.199126589863809e-05, 0.3010587805269192, 0.2770165336963357, 0.3040783874349082, 0.009423621954285059, 0.19786969492102494, 0.21752518828554698, 0.30315030139165783, 0.004268124514905118]

**4 0.304078387435**

**8 0.303150301392**

**2 0.301058780527**

**0 0.293100585709**

**3 0.277016533696**

FAM0495 ： Cluster 2

WBP0828 : Cluster 1

2018年10月23日星期二

下午4时40分

再次按照线性组合方式计算中心标记，得到新的结果：此时只有三个用户的JS\_Risk为0（30个里面漏报3个，召回率为90%）

JS\_Risks for CERT5.2

BYO1846,3.4

CHP1711,2.0

CIF1430,0.6

CKP0630,0.2

DCC1119,5.4

GWG0497,4.0

HIS1394,5.4

HMS1658,2.2

HSN0675,3.4

HXP0976,0.6

ICB1354,1.6

ITA0159,4.2

JAL0811,2.4

KSS1005,2.4

LVF1626,4.6

MCP0611,0.2

MDS0680,1.8

MGB1235,0.0

NAH1366,1.6

OKM1092,6.2

OSS1463,2.0

RRS0056,1.8

SIS0042,0.4

SNK1280,2.0

TMT0851,6.0

TNB1616,2.4

TRC1838,0.0

VCF1602,2.2

WDT1634,0.2

ZIE0741,0.0

试着分析最后一个JS\_Risk的用户

其选中的KMeans群簇为

ZIE0741 5 选中的friends群簇标号为：

Cluster\_RL is [0.3011279016521203, 0.006626718455519201, 0.19920439313306965, 0.3040776887459866, 0.310645401171046]

4 0.310645401171

3 0.304077688746

0 0.301127901652

本群簇大小为： 63 中心点坐标为： 0.0 0.0 0.0 0.0 0.007154035323049402 0.03128476735502727 0.031746031746031744

本群簇大小为： 57 中心点坐标为： 0.9954592363261053 0.021703743895822008 0.49921219732076966 0.0 0.00024709661477637757 0.00021172703211610235 0.0

本群簇大小为： 4 中心点坐标为： 0.448049327212825 0.9329896907216495 0.47075824100130265 0.0 0.7975352112676056 0.029223763278823898 0.03799728193722501

本群簇大小为： 31 中心点坐标为： 0.4502822357835661 0.05487196541403391 0.46526724863122143 0.0 0.05633802816901408 0.056211289222412864 0.06350373970490322

本群簇大小为： 1 中心点坐标为： 0.0 0.0 0.0 0.0 0.007042253521126761 1.0 1.0

其中，

LaidOff Friends is

0 []

1 ['BNS0484'] Cluster 1

2 ['MGM0539'] Cluster 2

3 ['JXH1061'] Cluster 1

4 ['CDO0684', 'DAS1320', 'ZJN1492', 'DNJ0740'] Cluster 2, Cluster 1, Cluster 1, Cluster 2

BNS0484,1.0,2.0,29827.0,0.0,0.0,0.0,0.0,

MGM0539,-0.2,2.0,19107.0,0.0,3.0,1715107.0,1.33333333333,

CDO0684,0.4,7.0,31978.1428571,0.0,3.0,1446771.0,2.0,

由于中心标记中log函数中的变量部分都很小，因此使用log2进行考察

RL = math.log(1 + (1 - abs(sr)) + (s1 \* s2 + r1 \* r2) + (s1 \* s3 + r1 \* r3), 2)

Cluster\_RL is [1.000325236652657, 0.022013482214297823, 1.0319416856976533, 0.6617426701737312, 1.0101242172737077]

2 1.0319416857

4 1.01012421727

0 1.00032523665

如此得到三个

最后的实验，下班前突然意识到EmailRatio通过MinMax归一化后只有0-1，公式中处理容易得到1和0的极值，因此，需要依据s1与r1的中心值重新计算

JS\_Risks for CERT5.2 （30个Insiders中6个为0）

BYO1846,3.4

CHP1711,4.0

CIF1430,1.0

CKP0630,0.0

DCC1119,4.2

GWG0497,0.0

HIS1394,5.2

HMS1658,3.0

HSN0675,1.8

HXP0976,1.2

ICB1354,1.0

ITA0159,2.0

JAL0811,2.0

KSS1005,1.0

LVF1626,9.4

MCP0611,1.0

MDS0680,1.8

MGB1235,0.0

NAH1366,0.0

OKM1092,6.2

OSS1463,0.0

RRS0056,1.4

SIS0042,0.0

SNK1280,3.8

TMT0851,6.0

TNB1616,6.2

TRC1838,3.2

VCF1602,3.0

WDT1634,1.0

ZIE0741,1.6

2018年10月23日星期二

下午10时47分

阶段小结

本文档主要记录的实验是关于力图解析用户的relationship to laid off employees的特征以识别出潜在具备满意度风险的高危用户。

其整体处理流程主要有以下几步：

* 将CERT5.2数据按照用户、时间日期以及类别进行提取整理，方便依据用户自身行为进行追踪；
* 利用EmailFilter-0.1.py来分析每个用户的邮件通讯列表，并提取七元特征（注意1.仅考虑A--B的1vs1用户通信；2. 七元特征（邮件通讯比EmailRatio，发送次数、平均发送字数大小、平均发送的附件个数以及接收次数、平均接收字数大小、平均接收的附件个数）；3. ER计算通过（X-Y）/(X+Y)的形式定义）
* 通过调用模块KMeans\_Module-0.1.py来运行KMeans\_for\_EmailContactFeats-0.1.py，从而实现归一化后七元特征的自动KMeans聚类，并通过中心标记的方法计算出不同群簇代表的与用户的邮件通讯关系深浅
* 中心标记公式
* 可以用于交叉匹配用户离职与邮件通讯的交集test.py

上述方案分析30个场景二的用户，却发现有6个用户的JS\_Risk为0

**SIS0042** 6 选中的friends群簇标号为：

Cluster\_RL is [0.000505088073401518, 0.12568448512810043, 0.6455707128484369, 0.012799302874754784, 0.26840709330492446, 0.644182886231868]

**2 0.645570712848**

**5 0.644182886232**

**4 0.268407093305**

其对应的离职与联系的交叉匹配结果为：

LaidOff Friends is

0 []

1 []

2 ['BRM0126']

3 ['MAF0467']

4 ['BSS0847', 'ESP1198', 'SCO1719']

BRM0126,-1.0,0.0,0.0,0.0,6.0,510148.333333,0.333333333333,0

MAF0467,0.8,9.0,27750.4444444,0.0,1.0,26186.0,0.0,1

BSS0847,1.0,1.0,40599.0,0.0,0.0,0.0,0.0,1

ESP1198,0.75,7.0,1344392.0,1.14285714286,1.0,24471.0,0.0,1

SCO1719,1.0,1.0,24648.0,0.0,0.0,0.0,0.0,1

-----------------------------------------------------------------------------------------------------------------------------

2018年10月24日星期三

上午9时40分

早上重新思考了昨天实验的中心标记公式，发现其实自己搞混了一个问题：

原始特征计算的EmailRatio取值范围在[-1, 1]之间，然后直接使用原始的结果做MinMax归一化后进行自动KMeans聚类；

1. 随后使用公式计算中心标记时，由于计算的是群簇中心点，也就是所有群簇点ER的均值，很有可能出现在一个单纯的Send/Recv群簇中，ER的均值为1或者0（此时MinMax已经将数值变化到了[0,1]）；而此时如果使用原始的公式

来表示ER的影响，而是转而使用新的公式

来表示。

1. 除了上述方法之外，还有一种自然的方法，即使用中心点的坐标特征重新计算中心点的er，然后参与原始公式计算即可，这样ERavg重新变化为了[-1,1]之间

sr = float((s1 - r1) / (s1 + r1))  
RL = math.log(1 + (1 - abs(sr)) + (s1 \* s2 + r1 \* r2) + (s1 \* s3 + r1 \* r3), math.e)

然后我们分别针对两种方法进行实验。

首先是自然的方法（2）的结果，即重新使用中心点坐标计算er，然后再计算中心点的Relationlevel定性表示：

JS\_Risks for CERT5.2 （此时30个用户中有6个用户的JS\_Risk为0，无法区分）

BYO1846,3.4

CHP1711,4.0

CIF1430,1.0

CKP0630,0.0

DCC1119,4.2

GWG0497,0.0

HIS1394,5.2

HMS1658,3.0

HSN0675,1.8

HXP0976,1.2

ICB1354,1.0

ITA0159,2.0

JAL0811,2.0

KSS1005,1.0

LVF1626,9.4

MCP0611,1.0

MDS0680,1.8

MGB1235,0.0

NAH1366,0.0

OKM1092,6.2

OSS1463,0.0

RRS0056,1.4

SIS0042,0.0

SNK1280,2.0

TMT0851,6.0

TNB1616,6.2

TRC1838,3.2

VCF1602,3.0

WDT1634,1.0

ZIE0741,1.6

其次是第一种方法，即使用自然指数变化

JS\_Risks for CERT5.2 （强调了均衡的重要性，有6个insiders的JS\_Risk为0）

BYO1846,1.0

CHP1711,2.0

CIF1430,1.0

CKP0630,0.0

DCC1119,4.2

GWG0497,0.0

HIS1394,5.2

HMS1658,3.0

HSN0675,1.8

HXP0976,1.2

ICB1354,1.0

ITA0159,2.0

JAL0811,2.0

KSS1005,1.0

LVF1626,9.4

MCP0611,1.0

MDS0680,0.0

MGB1235,0.0

NAH1366,0.0

OKM1092,2.6

OSS1463,0.0

RRS0056,1.4

SIS0042,0.0

SNK1280,1.0

TMT0851,1.2

TNB1616,4.4

TRC1838,3.2

VCF1602,3.0

WDT1634,1.0

ZIE0741,0.4

按照我们的理解，relationship影响着communication行为，也就是email行为，因此从email上应该可以建模出用户的relationship。但是对于我们最后的方法而言（采取的KMeans聚类与中心标记选择法）发现有6个Insiders丢失。

考虑是否可以引入personality因素？

本0.1版本主要就KMeans+中心标记方法进行实验分析，0.2版本将研究全部用户的权值计算排序；

如果根据物以类聚，人以群分的理念，我们基于CERT5.2的用户群簇进行基于OCEAN的聚类，然后选定与目标用户存在的聚类，并将其与leave/laidioff用户交叉比对；

2018年10月25日星期四

上午12时23分

利用睡前的丁点时间，就0.1版本中使用的JS\_Risk方法做下分析。

自己的初衷是假设虽然一个用户的组织结构中有不少中途离职的用户，看组织结构貌似有交集；然而，当然不能因为在一个部门、一个团队工作就认为该用户离职/解雇对目标有重要影响。

因此自己选择了可以表现刻画用户邮件通信的7元特征，特别的，EmailRatio用于刻画A与B通信的方向特征；在此基础上借助自动KMeans聚类和中心标记方法得到“关系较为亲密”的群簇，以二分法筛选与离职/解雇用户结果进行交叉匹配，并统计不同组织层次中关系亲密人员个数，乘以权值便得到JS\_Risk。

然而，上述方法在后续实验中发现，无论如何调整中心标记公式，总不能避免有JS\_Risk=0的Insiders出现，这也就意味着该方法筛选得到的用户并不能包括所有关联需分析的用户全集。

即：

* 即便JS\_Risk > 0， 也无法说明所有的关联用户都考虑到了；因为分析BYO1846用户时，就发现得到的两个群簇竟然是单向发送和单向接收用户，由此需要重新定义Relationship表现为communication而非friends；
* 有部分Insiders的JS\_Risk = 0，明显地，该方法丢失了需要分析的用户的重要信息；

虽然0.1版本方法实验证明讲不通，不符合数据自身的特征，然而其中心标记公式可以用来参考在后续的0.2版本中进行RelationLevel的计算，并且将该值与组织层次权值结合，得到该用户因为周围同事离职造成的JS\_Risk影响；

然后，需要根据上述工作针对CERT5.2的全部用户提取邮件特征，并进一步计算所有用户的JS\_Risk，然后从高到低排序，查看排序质量；