華東師範大學 | 软件学院 software engineering institute

个人邮件智能挖掘及可视化

Intelligent Mining and Visualization of Personal Email

黄一夫
软件学院
华东师范大学
上海, 中国
10092510437@ecnu.cn
指导老师：罗远哉

目录

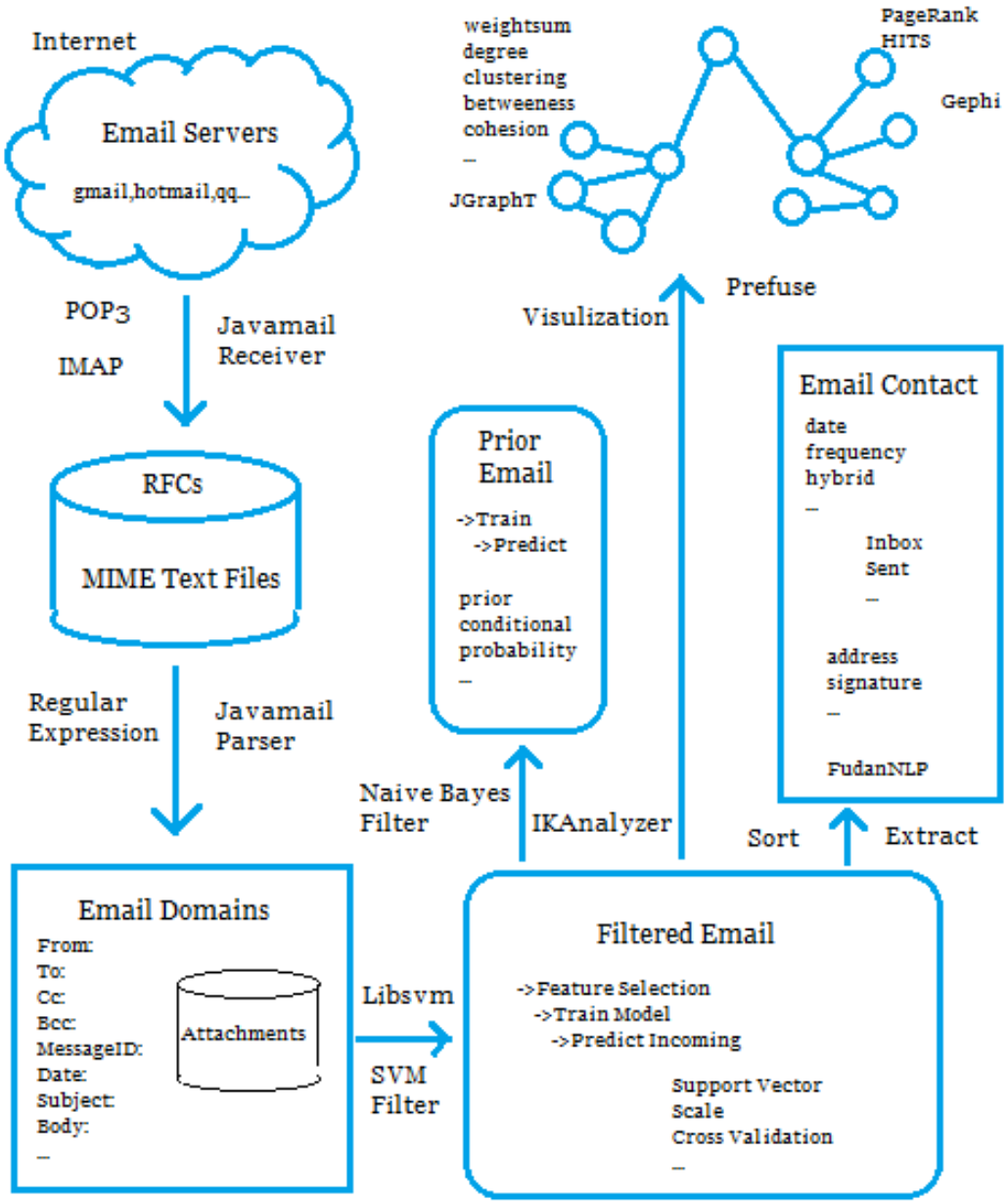
- **研究意义**
- 邮件预处理
- 邮件分类
- 邮件通讯录管理
- 邮件网络分析与可视化
- 代码重构
- 界面设计

研究意义

- 1.　每天我们都会收到大量的电子邮件，但查看它们却变成越来越头疼的问题。对此该项目不仅探讨垃圾邮件过滤，而且进一步向用户推荐重要邮件，从而提升邮件用户的收件效率。

- 2.　维护邮件通讯录是一个很重要的问题。该项目不仅从邮件签名，地址等中提取信息来提升邮件用户通讯录的完整性和可检索性，而且根据已给出的收件人预测推荐额外的收件人来提升邮件用户的发件效率。

- 3.　根据邮件收发关系，可以构建出个人社交网络。该项目分别从网络的结构特点和链接特点入手，计算多种系数值来衡量节点，并提供可视化展示，从而让邮件用户理清社交关系。


Internet
Email Servers
gmail,hotmail,qq...
POP3
IMAP
Javamail
Receiver
RFCs
MIME Text Files
Regular
Expression
Javamail
Parser
Email Domains
From:
To:
Cc:
Bcc:
MessageID:
Date:
Subject:
Body:
...
Attachments
Libsvm
SVM
Filter
Filtered Email
->Feature Selection
->Train Model
->Predict Incoming
Support Vector
Scale
Cross Validation
...
Naive Bayes
Filter
Prior
Email
->Train
->Predict
prior
conditional
probability
...
IKAnalyzer
Sort
Extract
Email Contact
date
frequency
hybrid
...
Inbox
Sent
...
address
signature
...
FudanNLP
Visulization
Prefuse
weightsum
degree
clustering
betweeness
cohesion
...
JGraphT
PageRank
HITS
Gephi

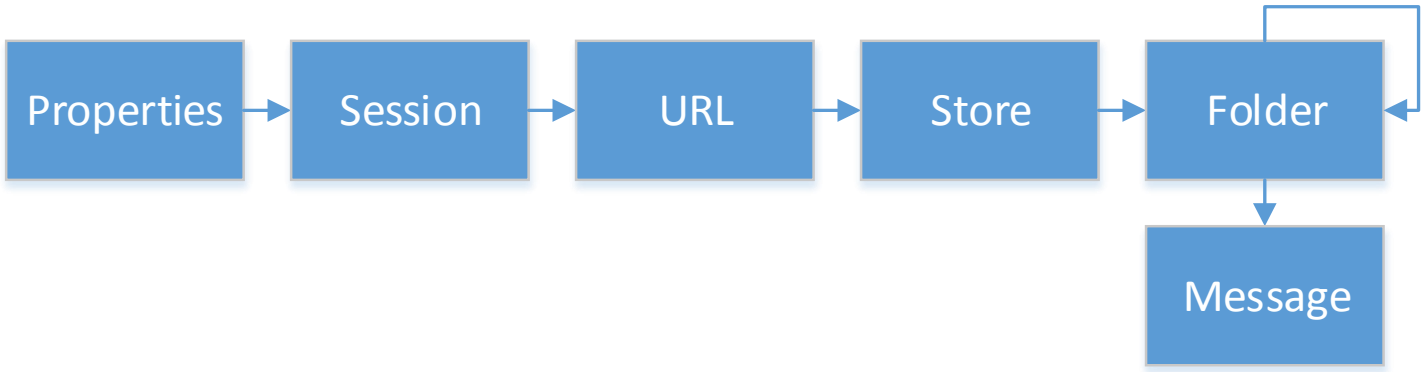
目录

- 研究意义
- **邮件预处理**
- 邮件分类
- 邮件通讯录管理
- 邮件网络分析与可视化
- 代码重构
- 界面设计

邮件预处理

- 邮件获取
  - 输入与输出

  - 调用过程



邮件预处理

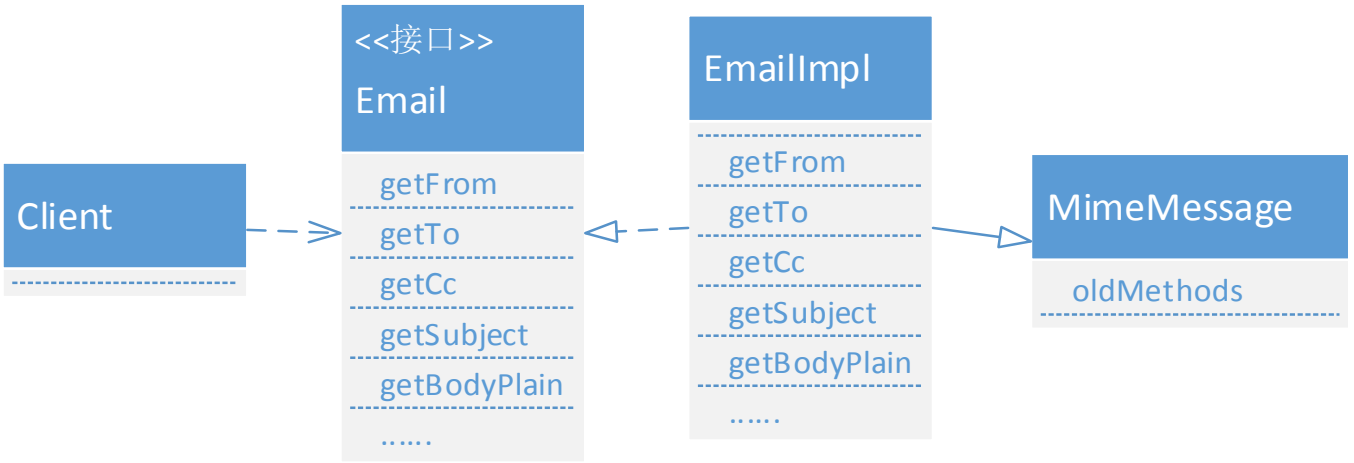
- 邮件获取
  - MIME文本文件示例

```
Subject: =?GB2312?Q?Project_update_=BB=E3=B1=A8?=
From: Lan Man <lanman.sg@gmail.com>
To: =?GB2312?B?wLzC/A==?= <mlan@cs.ecnu.edu.cn>
Content-Type: multipart/alternative; boundary=e89a8ff1c9a0e7631504bc1146ae

--e89a8ff1c9a0e7631504bc1146ae
Content-Type: text/plain; charset=GB2312
Content-Transfer-Encoding: base64

uPfOu9fps6SjrAoKvuDA68/uxL/X99K1tcTX7rrzzOG9u8jVxto01MI4yNXS0b6tuty9/MHLo6zO
qsHLyLexo7j3uPbQodfptcTP7sS/v6rM4rrNvfi2yMTcubvV/bOjvfjQ0KOsztKz9bK9tqjT2s/C
```

邮件预处理

- 邮件解析
  - 输入与输出

  - 适配器模式



邮件预处理

- 邮件解析
  - 解析格式示例
    - 垃圾邮件训练格式

```
0 1:22  2:0 3:1 4:6 5:0 6:0 7:
0 1:35  2:0 3:1 4:48  5:0 6:0
0 1:28  2:0 3:1 4:12  5:0 6:0
0 1:23  2:0 3:1 4:14  5:0 6:0
0 1:33  2:0 3:1 4:20  5:0 6:0
0 1:30  2:0 3:1 4:31  5:0 6:0
```

    - 邮件网络格式

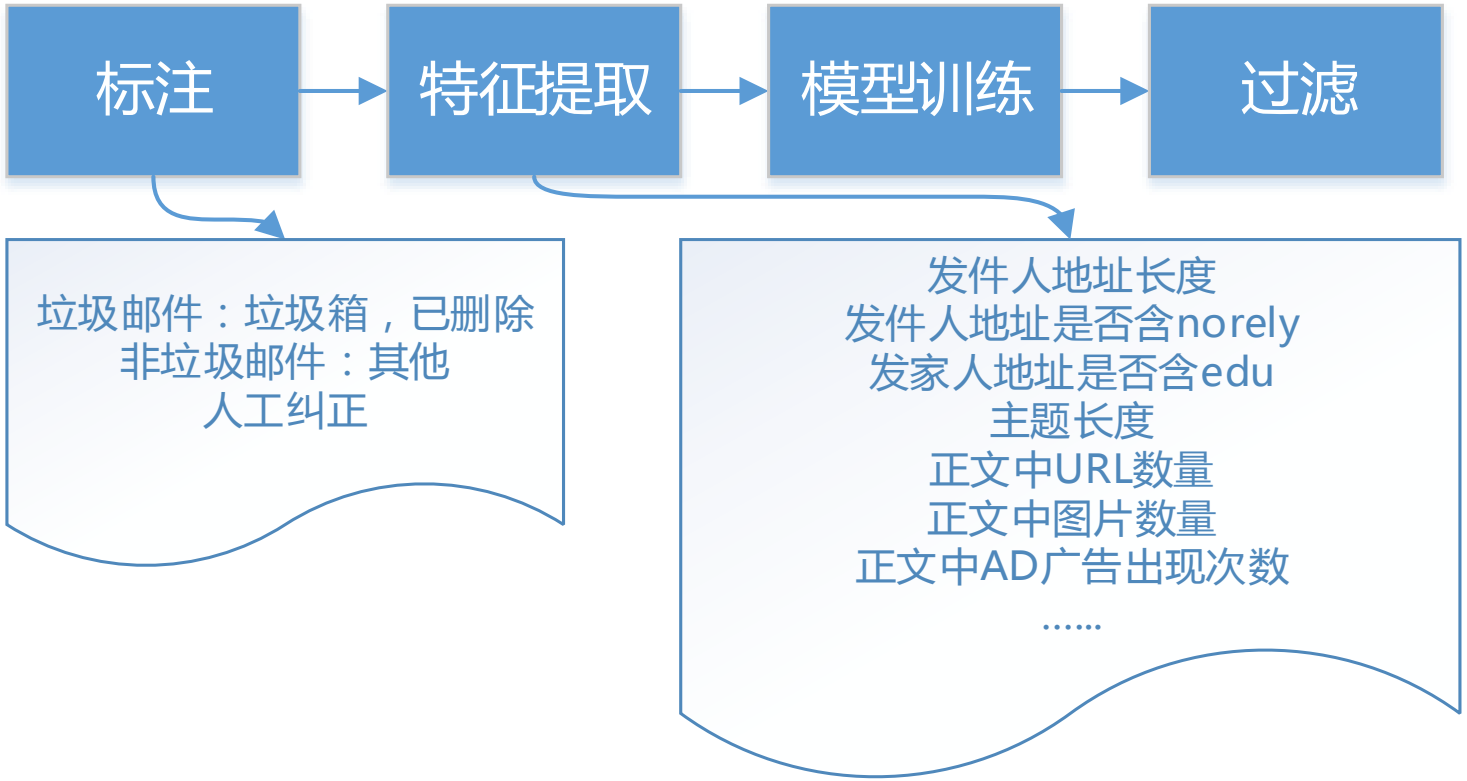
```
328679725@qq.com  leegyao@gmail.com 1 2011-10-08 15
10092510437@ecnu.cn sei_dzp@126.com 2 2012-03-01 16
328679725@qq.com  xiangnanhe@gmail.com  2 2012-01-2
328679725@qq.com  10092510437@ecnu.cn 1 2011-08-30
```

目录

- 研究意义
- 邮件预处理
- **邮件分类**
- 邮件通讯录管理
- 邮件网络分析与可视化
- 代码重构
- 界面设计

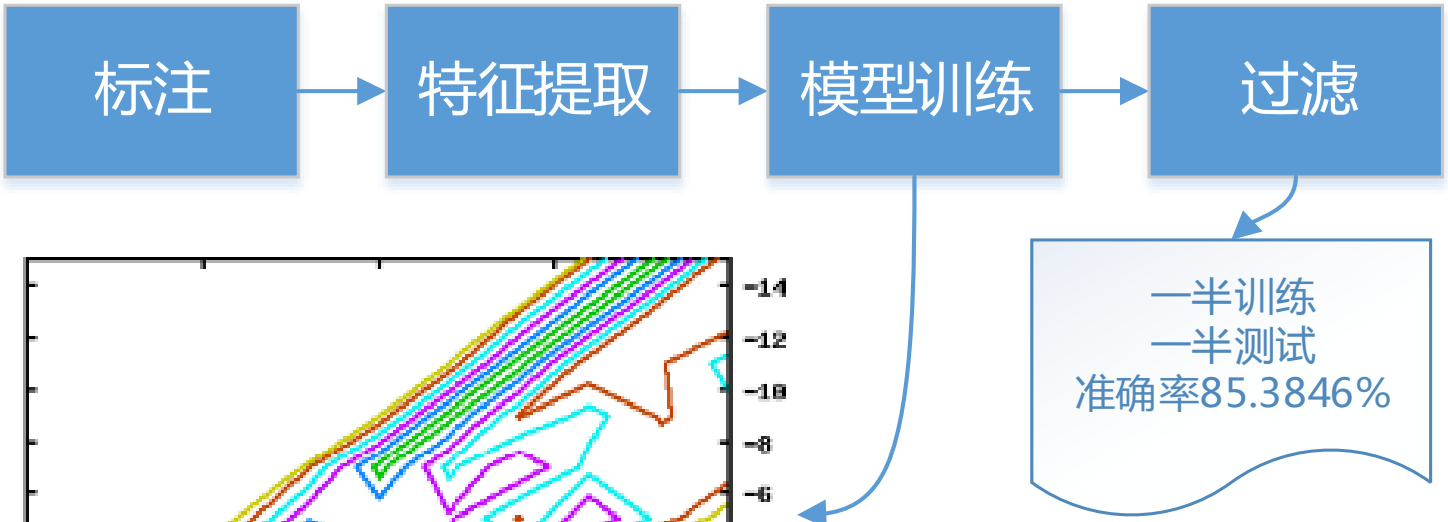
邮件分类

- 垃圾邮件过滤


标注
特征提取
模型训练
过滤
垃圾邮件：垃圾箱，已删除
非垃圾邮件：其他
人工纠正
发件人地址长度
发件人地址是否含norely
发家人地址是否含edu
主题长度
正文中URL数量
正文中图片数量
正文中AD广告出现次数
……

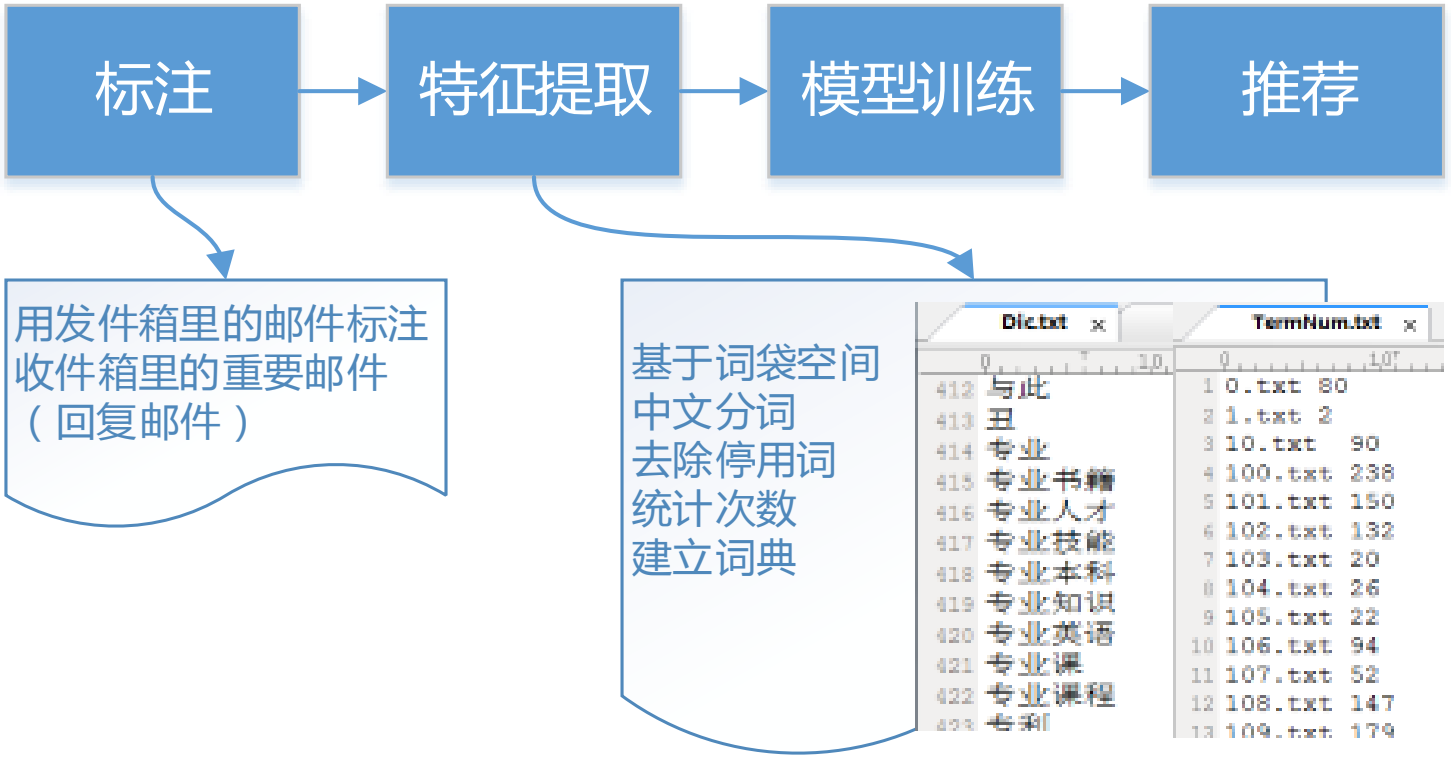
邮件分类

- 垃圾邮件过滤



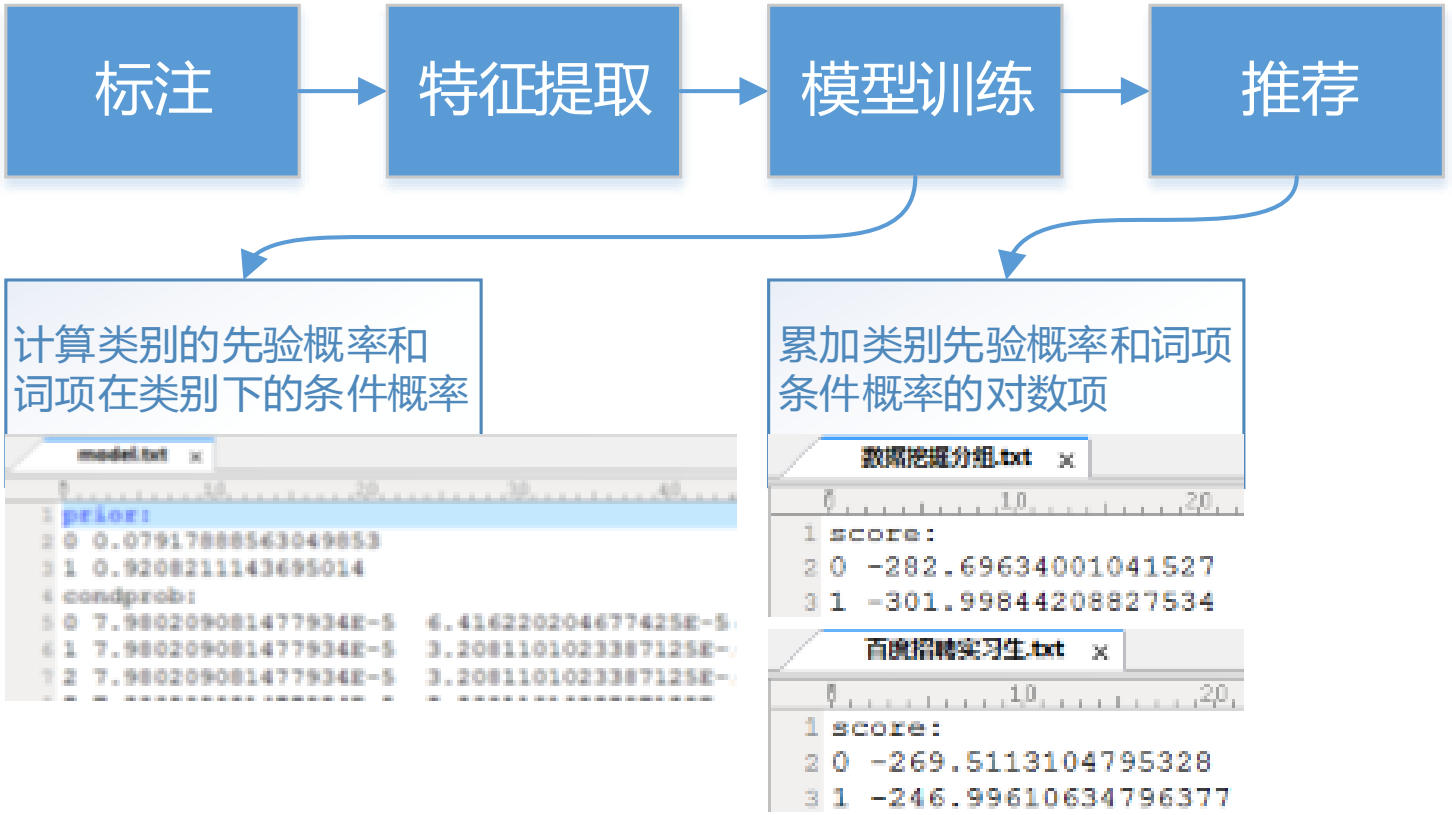
邮件分类

- 重要邮件推荐


标注
特征提取
模型训练
推荐
用发件箱里的邮件标注收件箱里的重要邮件（回复邮件）
基于词袋空间
中文分词
去除停用词
统计次数
建立词典
Dic.txt
TermNum.txt
412 与此
413 丑
414 专业
415 专业书籍
416 专业人才
417 专业技能
418 专业本科
419 专业知识
420 专业英语
421 专业课
422 专业课程
1 0.txt 80
2 1.txt 2
3 10.txt 90
4 100.txt 238
5 101.txt 150
6 102.txt 132
7 103.txt 20
8 104.txt 26
9 105.txt 22
10 106.txt 94
11 107.txt 52
12 108.txt 147

邮件分类

- 重要邮件推荐

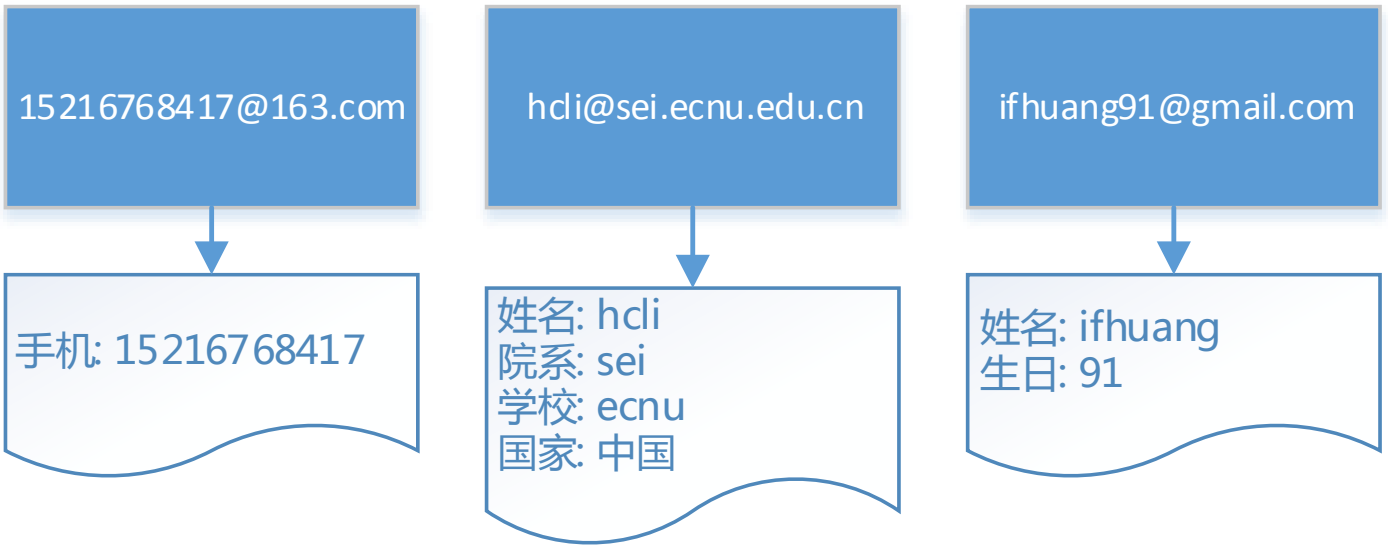

标注
特征提取
模型训练
推荐
计算类别的先验概率和
词项在类别下的条件概率
model.txt
prior:
0 0.07917888563049853
1 0.9208211143695014
condprob:
累加类别先验概率和词项
条件概率的对数项
数据挖掘分组.txt
score:
0 -282.69634001041527
1 -301.99844208827534
百度招聘实习生.txt
score:
0 -269.5113104795328
1 -246.99610634796377

目录

- 研究意义
- 邮件预处理
- 邮件分类
- **邮件通讯录管理**
- 邮件网络分析与可视化
- 代码重构
- 界面设计

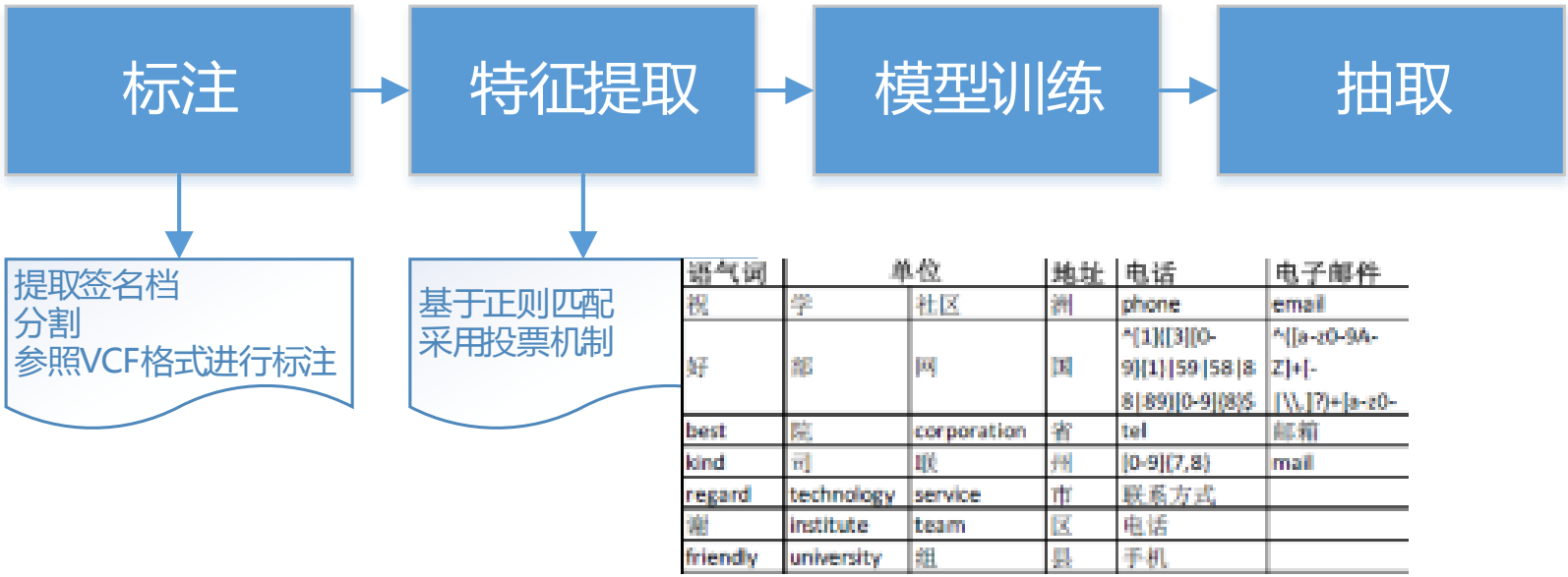
邮件通讯录管理

- 联系人信息抽取
  - 邮件地址信息抽取
    - 基于规则，正则匹配



邮件通讯录管理

- 联系人信息抽取
  - 邮件签名抽取



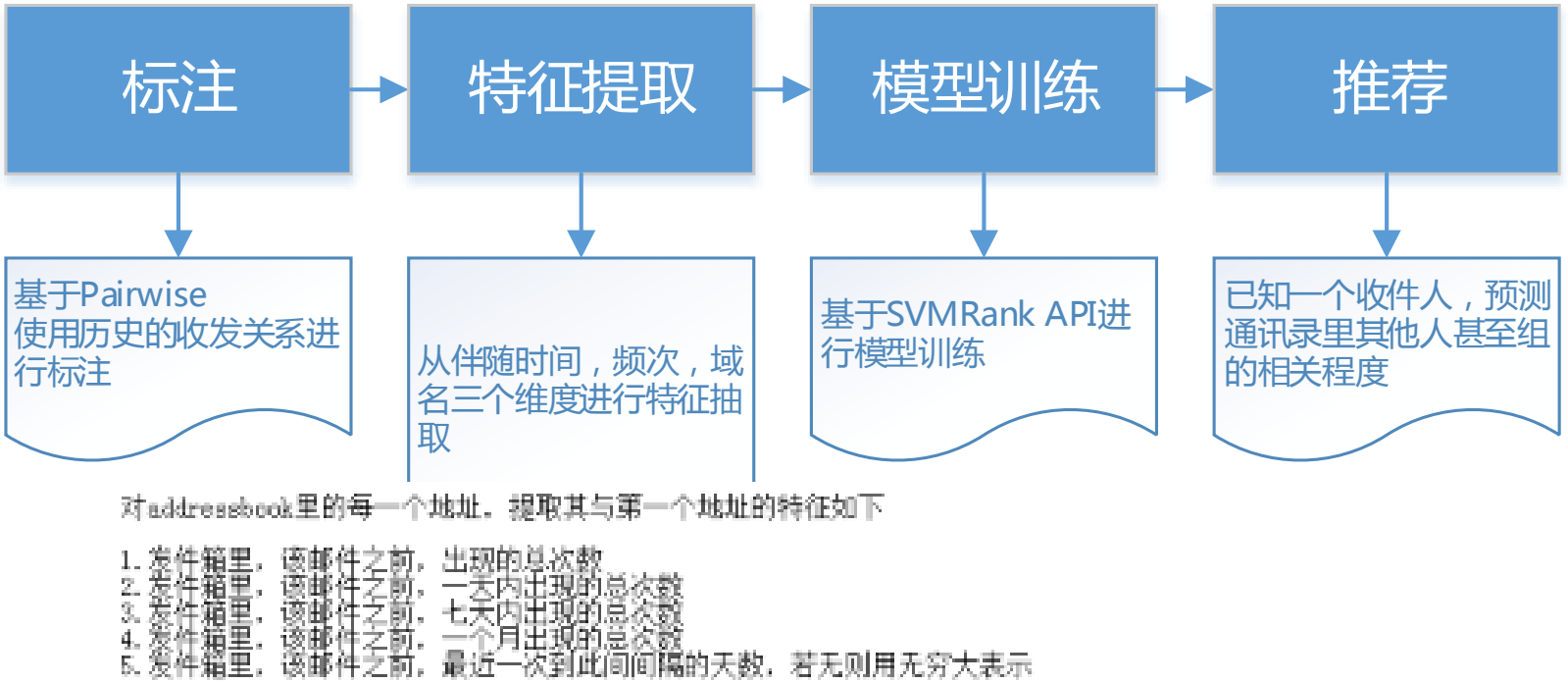
语气词	单位		地址	电话	电子邮件
祝	学	社区	洲	phone	email
好	部	网	国	^([1][3][0-9][1]|59|58|88|89)[0-9]{8}$	^([a-z0-9A-Z]+[-|\.])+[a-z0-
best	院	corporation	省	tel	邮箱
kind	司	联	州	[0-9]{7,8}	mail
regard	technology	service	市	联系方式	
谢	institute	team	区	电话	
friendly	university	组	县	手机	

邮件通讯录管理

- 联系人信息抽取
  - 邮件签名抽取

邮件通讯录管理

- 收件人推荐

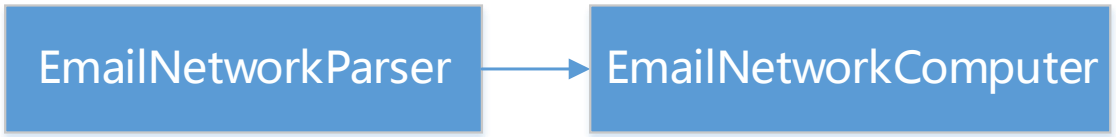


目录

- 研究意义
- 邮件预处理
- 邮件分类
- 邮件通讯录管理
- **邮件网络分析与可视化**
- 代码重构
- 界面设计

邮件网络分析与可视化

- 邮件网络分析



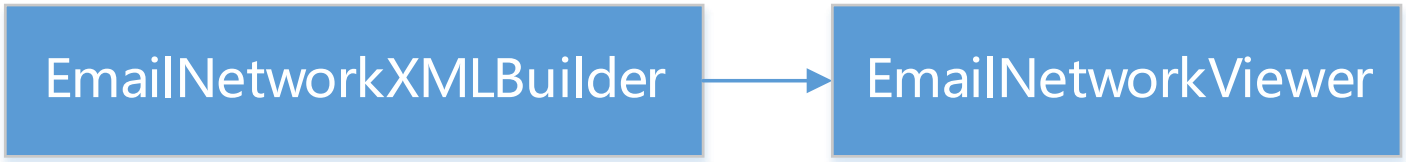
  - 结构
    - 度数：衡量节点的活跃程度
    - 聚集系数：衡量节点的聚团程度
    - 介数：衡量节点的中介程度
    - 凝聚度：衡量节点的中心程度
  - 链接
    - PageRank：衡量节点的重要程度
    - HITS：衡量节点的枢纽程度和权威程度

邮件网络分析与可视化

- 邮件网络分析
  - 重要节点
    - 活跃节点：与很多节点通信的节点
    - 领导节点：仅与一些活跃节点通信的节点
    - 意外节点：外围节点和桥接节点
  - 社团发现
    - 话题发现：不同社团的话题不同
    - 专家发现：汇总并分类历史知识

邮件网络分析与可视化

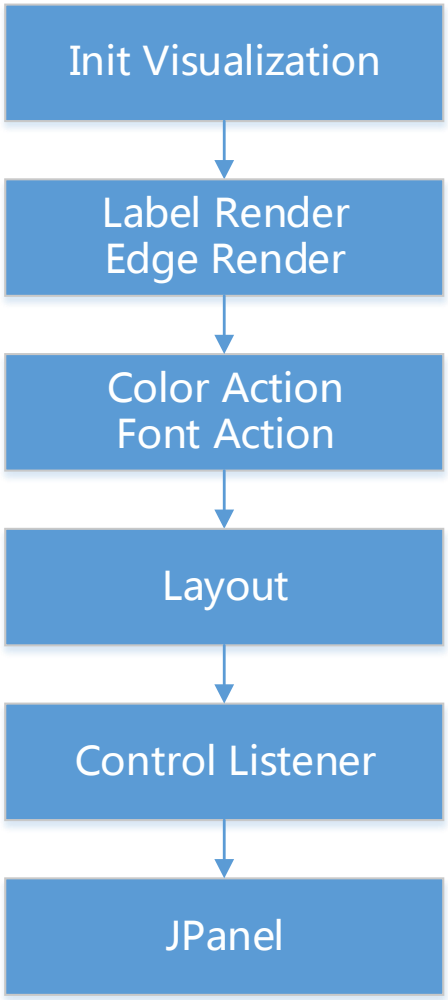
- 邮件网络可视化



  - 根据From, To, Cc建立邮件网络
  - 采用物理拉力引擎进行绘制
  - 使用不同颜色代表不同类型的节点
  - 边长与通信次数成反比
  - 从边可以查看相关通信邮件

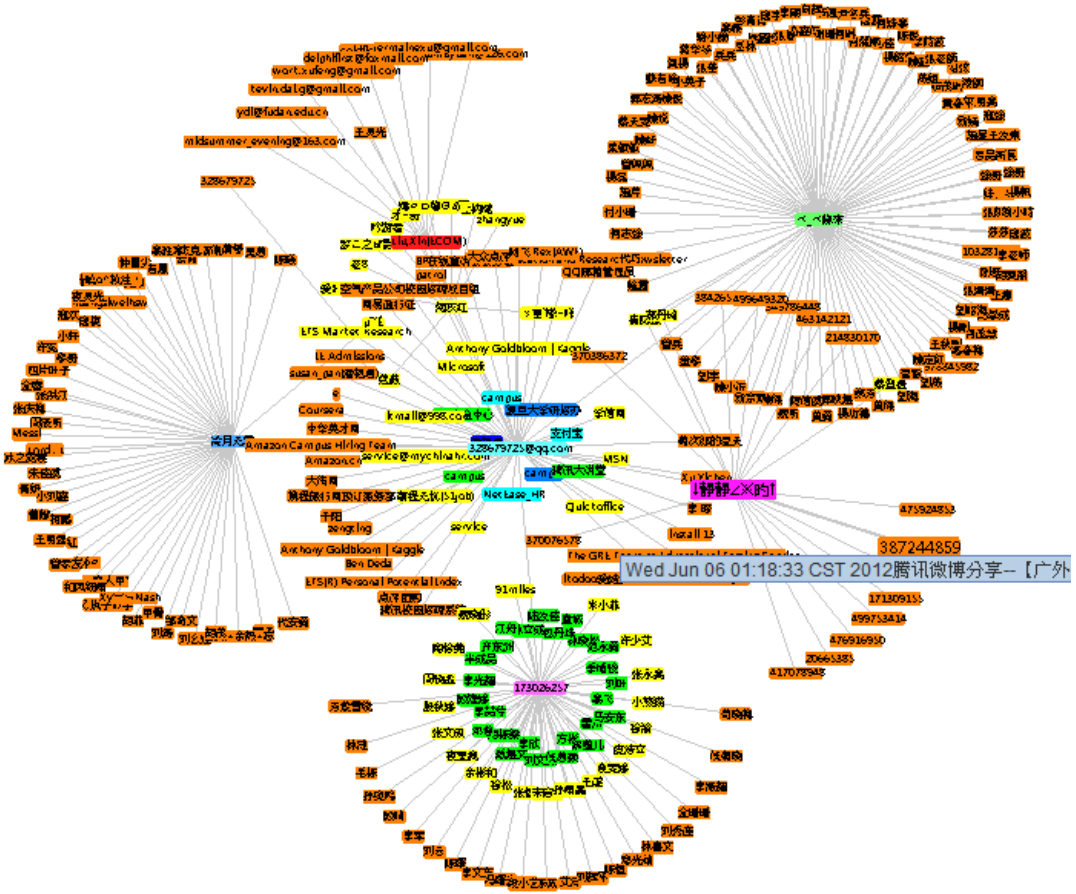
邮件网络分析与可视化

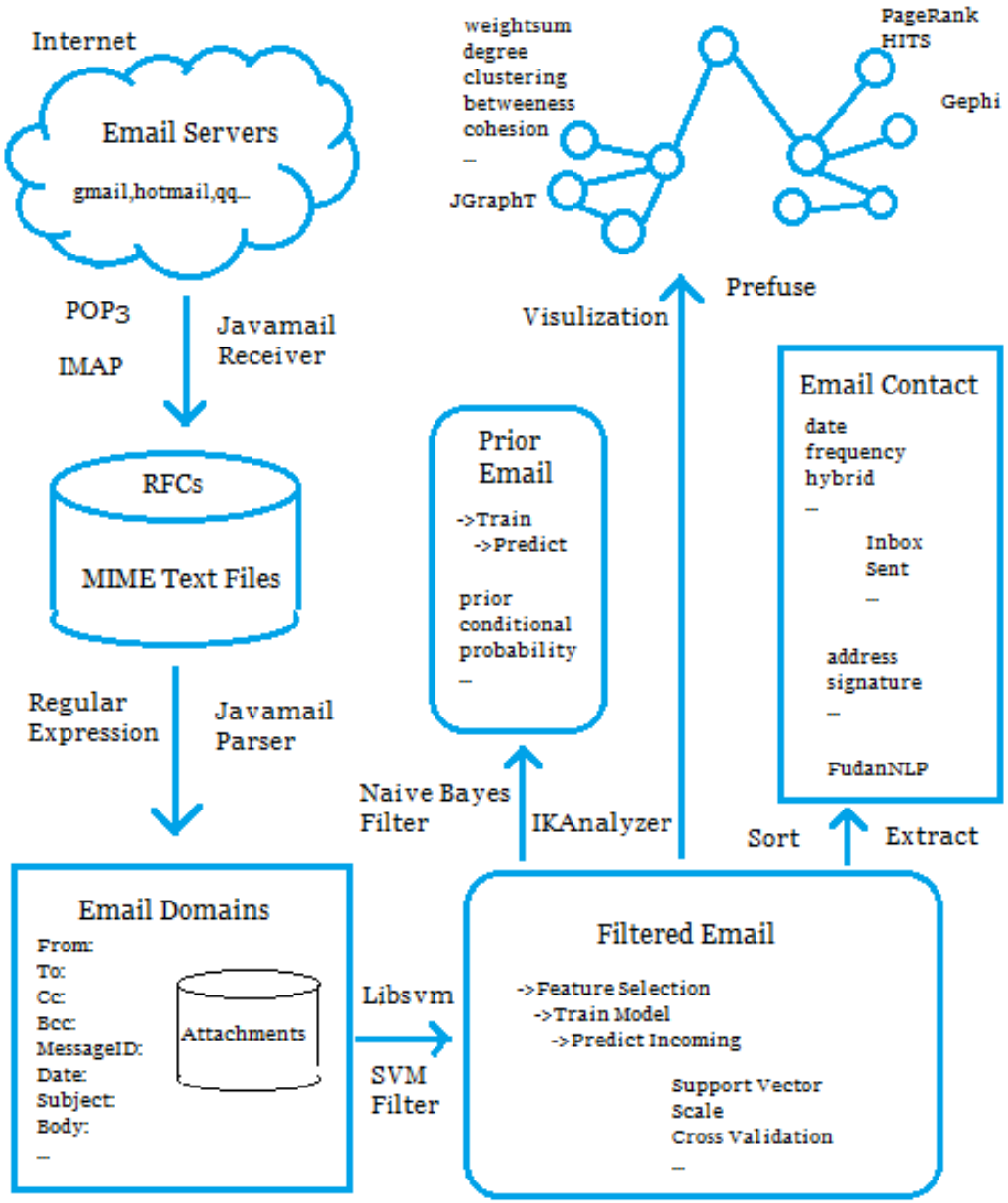
- 邮件网络可视化



邮件网络分析与可视化

- 邮件网络可视化



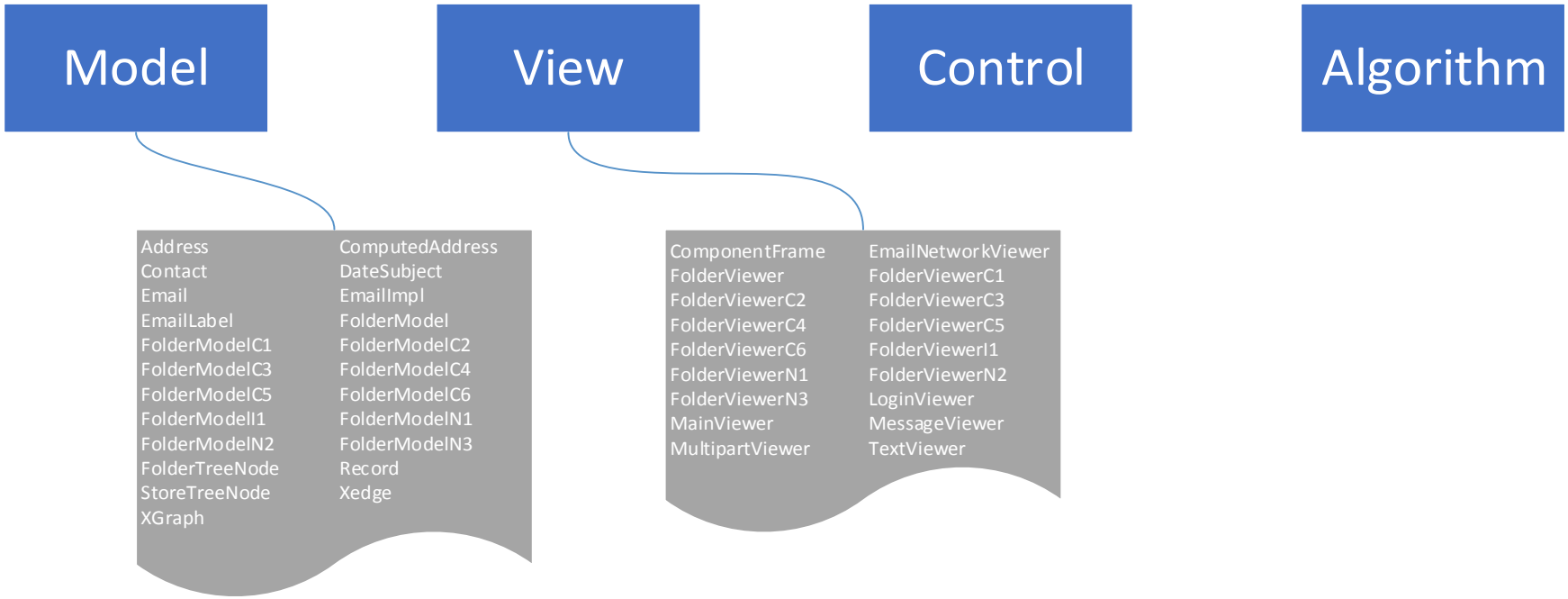

Internet
Email Servers
gmail,hotmail,qq...
POP3
IMAP
Javamail
Receiver
RFCs
MIME Text Files
Regular
Expression
Javamail
Parser
Email Domains
From:
To:
Cc:
Bcc:
MessageID:
Date:
Subject:
Body:
...
Attachments
Libsvm
SVM
Filter
Filtered Email
->Feature Selection
->Train Model
->Predict Incoming
Support Vector
Scale
Cross Validation
...
Naive Bayes
Filter
Prior
Email
->Train
->Predict
prior
conditional
probability
...
IKAnalyzer
Sort
Extract
Email Contact
date
frequency
hybrid
...
Inbox
Sent
...
address
signature
...
FudanNLP
Visulization
Prefuse
weightsum
degree
clustering
betweeness
cohesion
...
JGraphT
PageRank
HITS
Gephi

目录

- 研究意义
- 邮件预处理
- 邮件分类
- 邮件通讯录管理
- 邮件网络分析与可视化
- **代码重构**
- 界面设计

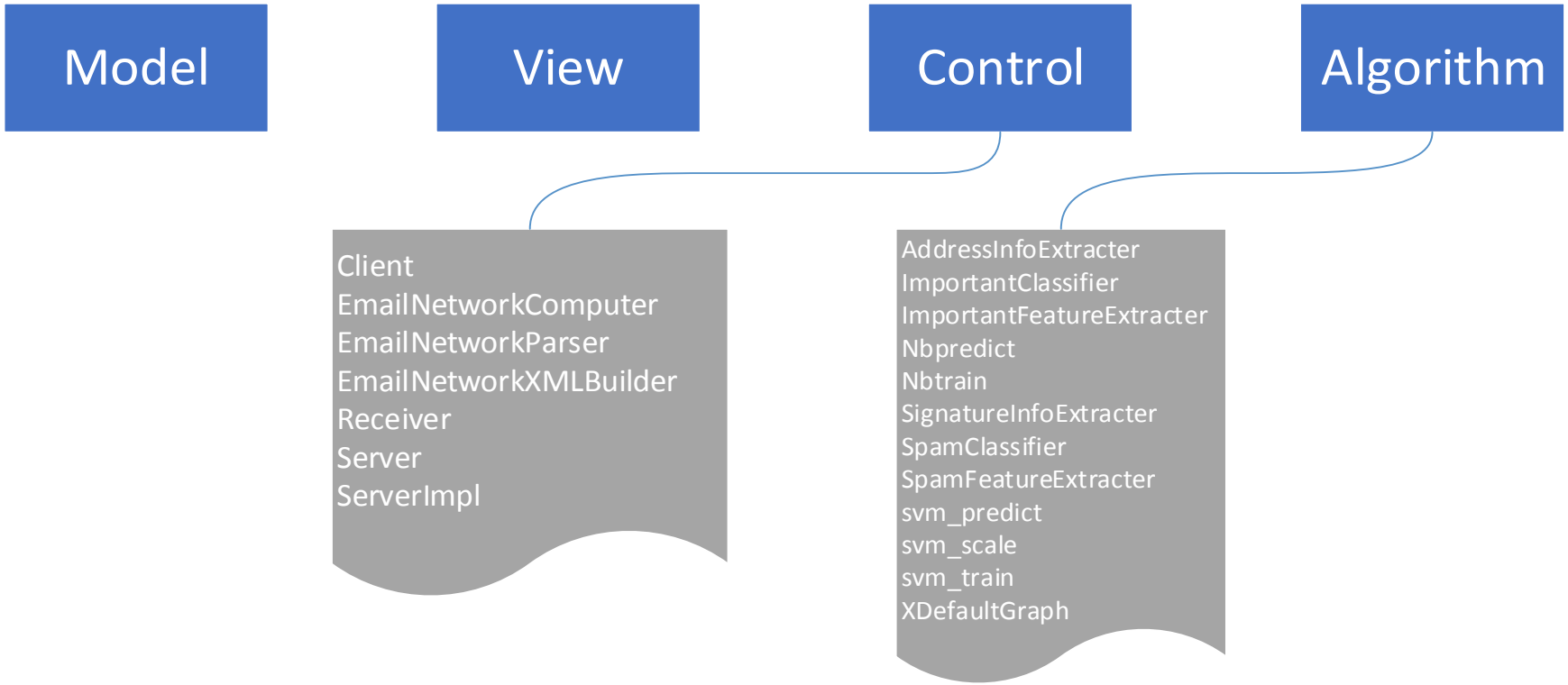
代码重构

- MVC架构


Model
View
Control
Algorithm
Address
Contact
Email
EmailLabel
FolderModelC1
FolderModelC3
FolderModelC5
FolderModelI1
FolderModelN2
FolderTreeNode
StoreTreeNode
XGraph
ComputedAddress
DateSubject
EmailImpl
FolderModel
FolderModelC2
FolderModelC4
FolderModelC6
FolderModelN1
FolderModelN3
Record
Xedge
ComponentFrame
FolderViewer
FolderViewerC2
FolderViewerC4
FolderViewerC6
FolderViewerN1
FolderViewerN3
MainViewer
MultipartViewer
EmailNetworkViewer
FolderViewerC1
FolderViewerC3
FolderViewerC5
FolderViewerI1
FolderViewerN2
LoginViewer
MessageViewer
TextViewer

代码重构

- MVC架构

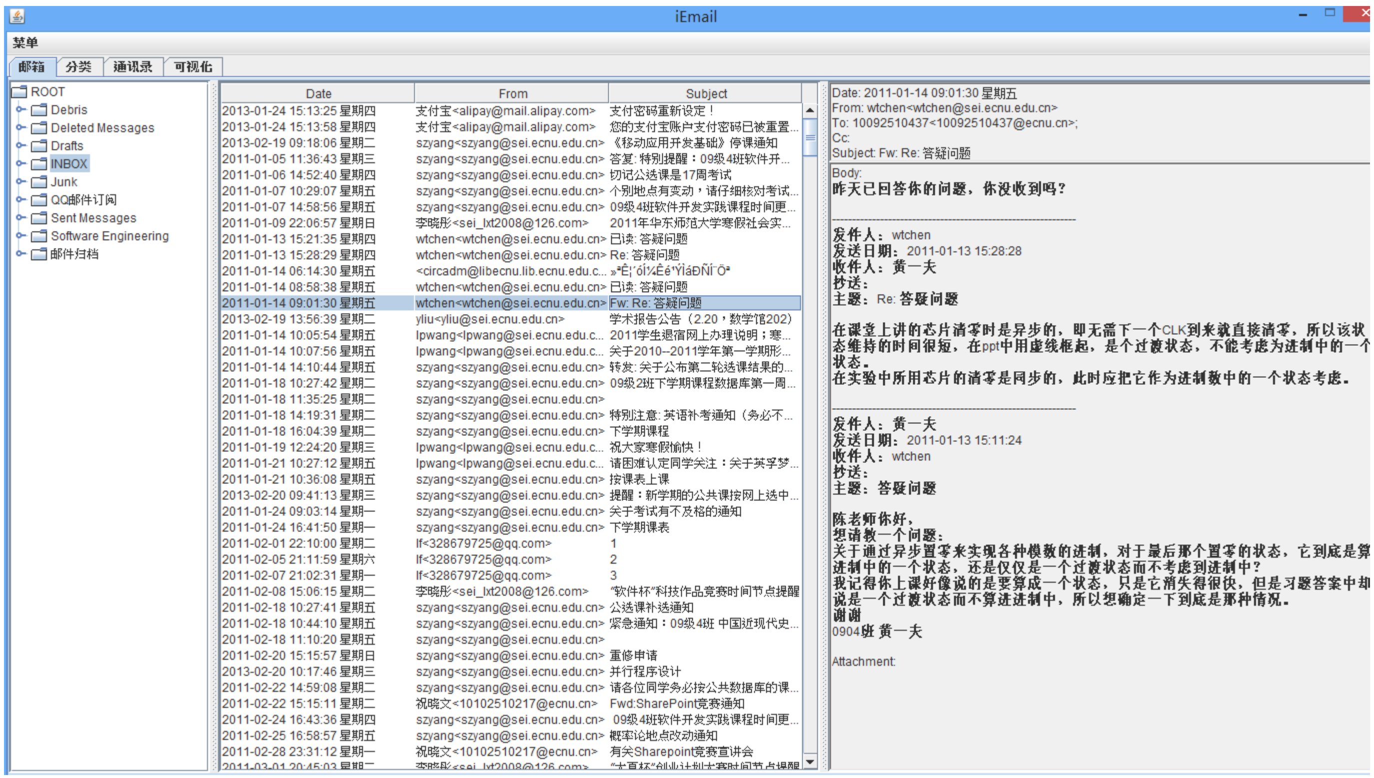

Model
View
Control
Algorithm
Client
EmailNetworkComputer
EmailNetworkParser
EmailNetworkXMLBuilder
Receiver
Server
ServerImpl
AddressInfoExtracter
ImportantClassifier
ImportantFeatureExtracter
Nbpredict
Nbtrain
SignatureInfoExtracter
SpamClassifier
SpamFeatureExtracter
svm_predict
svm_scale
svm_train
XDefaultGraph

目录

- 研究意义
- 邮件预处理
- 邮件分类
- 邮件通讯录管理
- 邮件网络分析与可视化
- 代码重构
- **界面设计**

界面设计

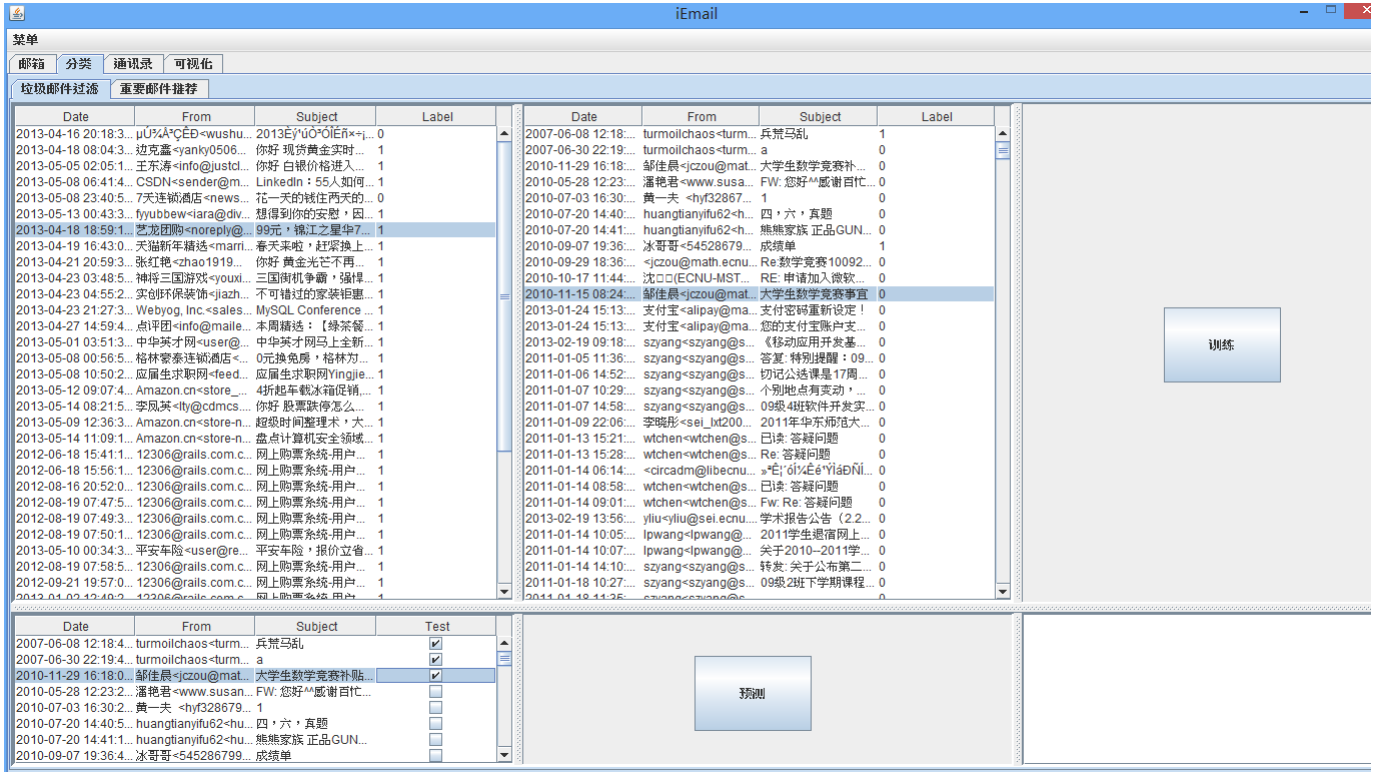
- 邮箱界面



iEmail
菜单
邮箱
分类
通讯录
可视化
ROOT
Debris
Deleted Messages
Drafts
INBOX
Junk
QQ邮件订阅
Sent Messages
Software Engineering
邮件归档
Date
From
Subject
2013-01-24 15:13:25 星期四 支付宝<alipay@mail.alipay.com> 支付密码重新设定！
2013-01-24 15:13:58 星期四 支付宝<alipay@mail.alipay.com> 您的支付宝账户支付密码已被重置...
2013-02-19 09:18:06 星期二 szyang<szyang@sei.ecnu.edu.cn> 《移动应用开发基础》停课通知
2011-01-05 11:36:43 星期三 szyang<szyang@sei.ecnu.edu.cn> 答复: 特别提醒：09级4班软件开...
2011-01-06 14:52:40 星期四 szyang<szyang@sei.ecnu.edu.cn> 切记公选课是17周考试
2011-01-07 10:29:07 星期五 szyang<szyang@sei.ecnu.edu.cn> 个别地点有变动，请仔细核对考试...
2011-01-07 14:58:56 星期五 szyang<szyang@sei.ecnu.edu.cn> 09级4班软件开发实践课程时间更...
2011-01-09 22:06:57 星期日 李晓彤<sei_lxt2008@126.com> 2011年华东师范大学寒假社会实...
2011-01-13 15:21:35 星期四 wtchen<wtchen@sei.ecnu.edu.cn> 已读: 答疑问题
2011-01-13 15:28:29 星期四 wtchen<wtchen@sei.ecnu.edu.cn> Re: 答疑问题
2011-01-14 06:14:30 星期五 <circadm@libecnu.lib.ecnu.edu.c...
2011-01-14 08:58:38 星期五 wtchen<wtchen@sei.ecnu.edu.cn> 已读: 答疑问题
2011-01-14 09:01:30 星期五 wtchen<wtchen@sei.ecnu.edu.cn> Fw: Re: 答疑问题
2013-02-19 13:56:39 星期二 yliu<yliu@sei.ecnu.edu.cn> 学术报告公告（2.20，数学馆202）
2011-01-14 10:05:54 星期五 lpwang<lpwang@sei.ecnu.edu.c... 2011学生退宿网上办理说明；寒...
2011-01-14 10:07:56 星期五 lpwang<lpwang@sei.ecnu.edu.c... 关于2010--2011学年第一学期形...
2011-01-14 14:10:44 星期五 szyang<szyang@sei.ecnu.edu.cn> 转发: 关于公布第二轮选课结果的...
2011-01-18 10:27:42 星期二 szyang<szyang@sei.ecnu.edu.cn> 09级2班下学期课程数据库第一周...
2011-01-18 11:35:25 星期二 szyang<szyang@sei.ecnu.edu.cn>
2011-01-18 14:19:31 星期二 szyang<szyang@sei.ecnu.edu.cn> 特别注意: 英语补考通知（务必不...
2011-01-18 16:04:39 星期二 szyang<szyang@sei.ecnu.edu.cn> 下学期课程
2011-01-19 12:24:20 星期三 lpwang<lpwang@sei.ecnu.edu.c... 祝大家寒假愉快！
2011-01-21 10:27:12 星期五 lpwang<lpwang@sei.ecnu.edu.c... 请困难认定同学关注：关于英孚梦...
2011-01-21 10:36:08 星期五 szyang<szyang@sei.ecnu.edu.cn> 按课表上课
2013-02-20 09:41:13 星期三 szyang<szyang@sei.ecnu.edu.cn> 提醒：新学期的公共课按网上选中...
2011-01-24 09:03:14 星期一 szyang<szyang@sei.ecnu.edu.cn> 关于考试有不及格的通知
2011-01-24 16:41:50 星期一 szyang<szyang@sei.ecnu.edu.cn> 下学期课表
2011-02-01 22:10:00 星期二 lf<328679725@qq.com> 1
2011-02-05 21:11:59 星期六 lf<328679725@qq.com> 2
2011-02-07 21:02:31 星期一 lf<328679725@qq.com> 3
2011-02-08 15:06:15 星期二 李晓彤<sei_lxt2008@126.com> "软件杯"科技作品竞赛时间节点提醒
2011-02-18 10:27:41 星期五 szyang<szyang@sei.ecnu.edu.cn> 公选课补选通知
2011-02-18 10:44:10 星期五 szyang<szyang@sei.ecnu.edu.cn> 紧急通知：09级4班 中国近现代史...
2011-02-18 11:10:20 星期五 szyang<szyang@sei.ecnu.edu.cn>
2011-02-20 15:15:57 星期日 szyang<szyang@sei.ecnu.edu.cn> 重修申请
2013-02-20 10:17:46 星期三 szyang<szyang@sei.ecnu.edu.cn> 并行程序设计
2011-02-22 14:59:08 星期二 szyang<szyang@sei.ecnu.edu.cn> 请各位同学务必按公共数据库的课...
2011-02-22 15:15:11 星期二 祝晓文<10102510217@ecnu.cn> Fwd:SharePoint竞赛通知
2011-02-24 16:43:36 星期四 szyang<szyang@sei.ecnu.edu.cn> 09级4班软件开发实践课程时间更...
2011-02-25 16:58:57 星期五 szyang<szyang@sei.ecnu.edu.cn> 概率论地点改动通知
2011-02-28 23:31:12 星期一 祝晓文<10102510217@ecnu.cn> 有关Sharepoint竞赛宣讲会
Date: 2011-01-14 09:01:30 星期五
From: wtchen<wtchen@sei.ecnu.edu.cn>
To: 10092510437<10092510437@ecnu.cn>;
Cc:
Subject: Fw: Re: 答疑问题
Body:
昨天已回答你的问题，你没收到吗？
发件人：wtchen
发送日期：2011-01-13 15:28:28
收件人：黄一夫
抄送：
主题：Re: 答疑问题
在课堂上讲的芯片清零时是异步的，即无需下一个CLK到来就直接清零，所以该状态维持的时间很短，在ppt中用虚线框起，是个过渡状态，不能考虑为进制中的一个状态。
在实验中所用芯片的清零是同步的，此时应把它作为进制数中的一个状态考虑。
发件人：黄一夫
发送日期：2011-01-13 15:11:24
收件人：wtchen
抄送：
主题：答疑问题
陈老师你好，
想请教一个问题：
关于通过异步置零来实现各种模数的进制，对于最后那个置零的状态，它到底是算进制中的一个状态，还是仅仅是一个过渡状态而不考虑到进制中？
我记得你上课好像说的是要算成一个状态，只是它消失得很快，但是习题答案中却说是一个过渡状态而不算进进制中，所以想确定一下到底是那种情况。
谢谢
0904班 黄一夫
Attachment:

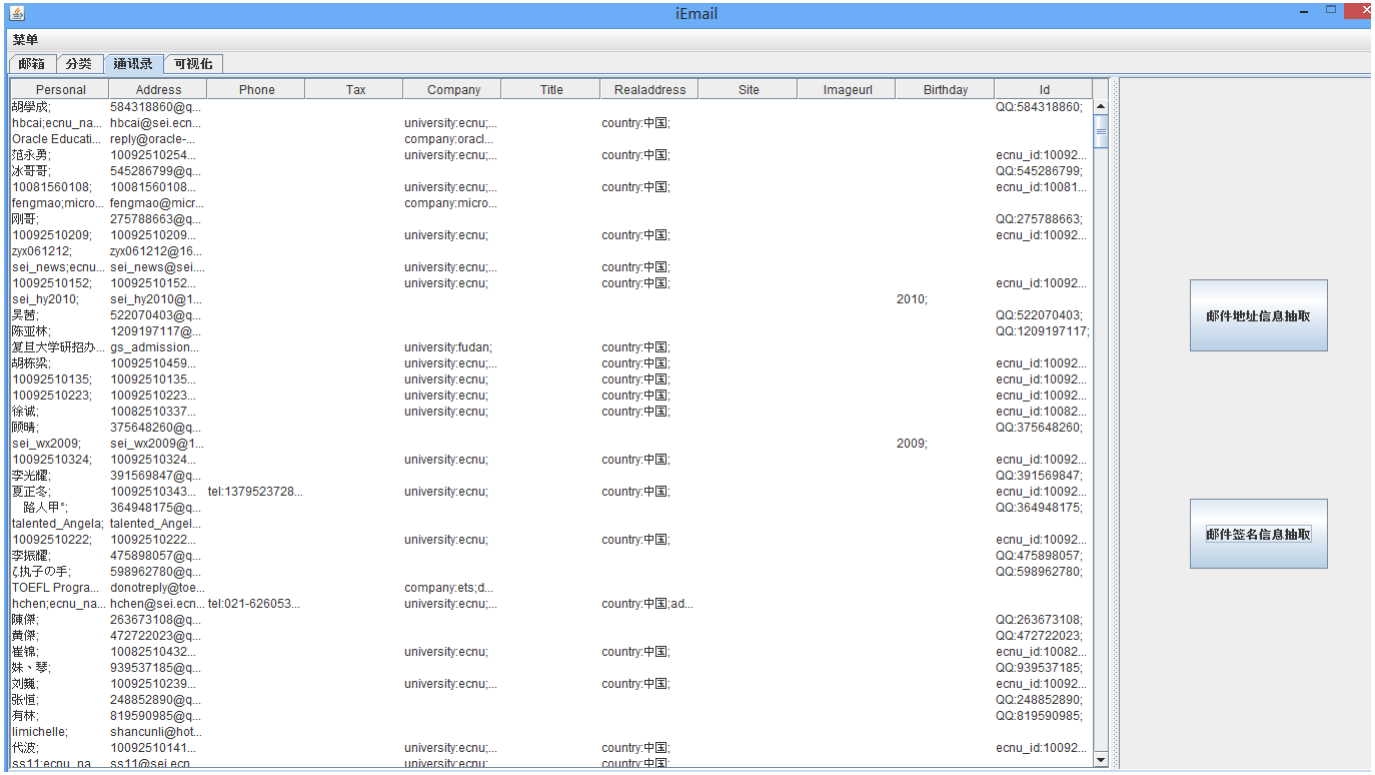
界面设计

- 分类界面


iEmail
菜单
邮箱
分类
通讯录
可视化
垃圾邮件过滤
重要邮件推荐
Date
From
Subject
Label
Test
训练
预测

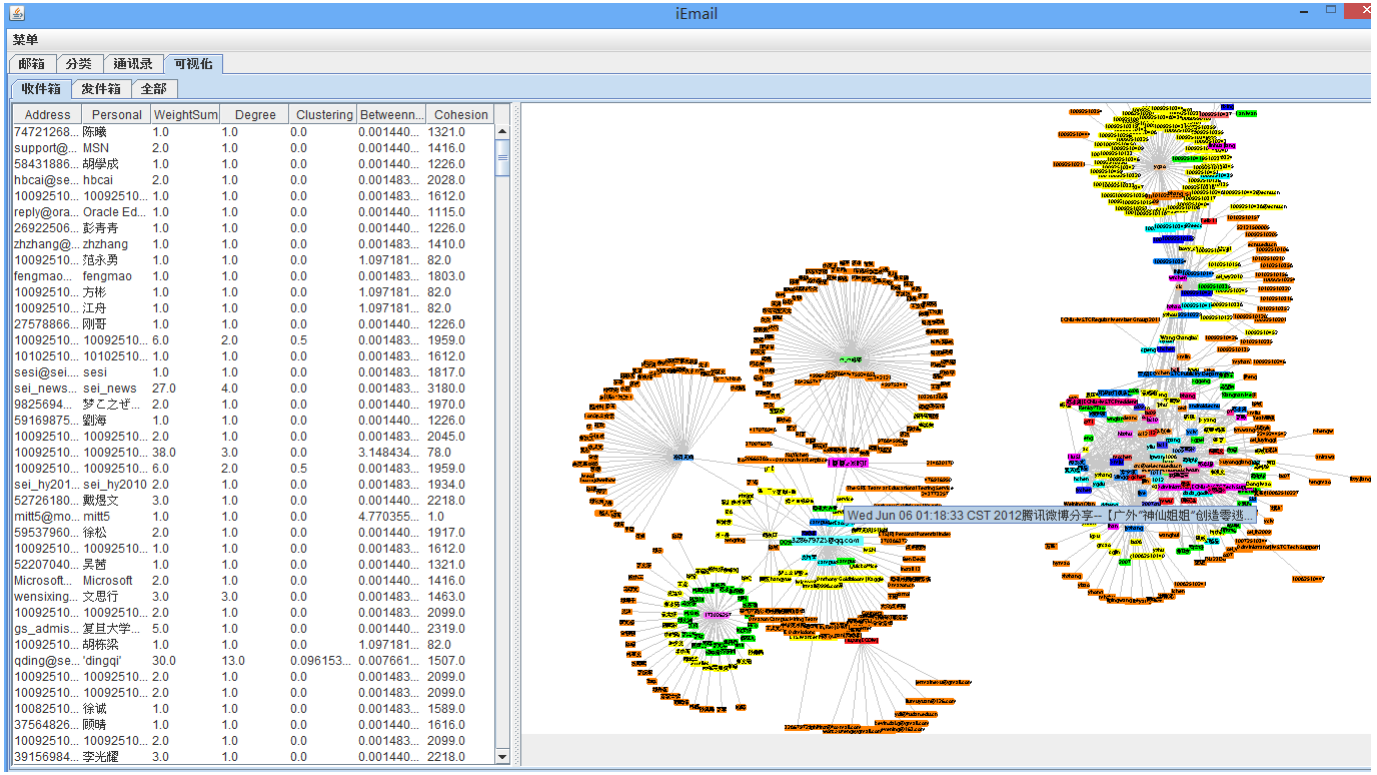
界面设计

- 通讯录界面


iEmail
菜单
邮箱
分类
通讯录
可视化
Personal
Address
Phone
Tax
Company
Title
Realaddress
Site
Imageurl
Birthday
Id
邮件地址信息抽取
邮件签名信息抽取

界面设计

- 可视化界面


iEmail
菜单
邮箱
分类
通讯录
可视化
收件箱
发件箱
全部
Address Personal WeightSum Degree Clustering Betweenn... Cohesion
74721268... 陈曦 1.0 1.0 0.0 0.001440... 1321.0
support@... MSN 2.0 1.0 0.0 0.001440... 1416.0
58431886... 胡學成 1.0 1.0 0.0 0.001440... 1226.0
hbcai@se... hbcai 2.0 1.0 0.0 0.001483... 2028.0
10092510... 10092510... 1.0 1.0 0.0 0.001483... 1612.0
reply@ora... Oracle Ed... 1.0 1.0 0.0 0.001440... 1115.0
26922506... 彭青青 1.0 1.0 0.0 0.001440... 1226.0
zhzhang@... zhzhang 1.0 1.0 0.0 0.001483... 1410.0
10092510... 范永勇 1.0 1.0 0.0 1.097181... 82.0
fengmao... fengmao 1.0 1.0 0.0 0.001483... 1803.0
10092510... 方彬 1.0 1.0 0.0 1.097181... 82.0
10092510... 江舟 1.0 1.0 0.0 1.097181... 82.0
27578866... 刚哥 1.0 1.0 0.0 0.001440... 1226.0
10092510... 10092510... 6.0 2.0 0.5 0.001483... 1959.0
10102510... 10102510... 1.0 1.0 0.0 0.001483... 1612.0
sesi@sei.... sesi 1.0 1.0 0.0 0.001483... 1817.0
sei_news... sei_news 27.0 4.0 0.0 0.001483... 3180.0
9825694... 梦乙之ゼ... 2.0 1.0 0.0 0.001440... 1970.0
59169875... 劉海 1.0 1.0 0.0 0.001440... 1226.0
10092510... 10092510... 2.0 1.0 0.0 0.001483... 2045.0
10092510... 10092510... 38.0 3.0 0.0 3.148434... 78.0
10092510... 10092510... 6.0 2.0 0.5 0.001483... 1959.0
sei_hy201... sei_hy2010 2.0 1.0 0.0 0.001483... 1934.0
52726180... 戴煜文 3.0 1.0 0.0 0.001440... 2218.0
mitt5@mo... mitt5 1.0 1.0 0.0 4.770355... 1.0
59537960... 徐松 2.0 1.0 0.0 0.001440... 1917.0
10092510... 10092510... 1.0 1.0 0.0 0.001483... 1612.0
52207040... 吴茜 1.0 1.0 0.0 0.001440... 1321.0
Microsoft... Microsoft 2.0 1.0 0.0 0.001440... 1416.0
wensixing... 文思行 3.0 3.0 0.0 0.001483... 1463.0
10092510... 10092510... 2.0 1.0 0.0 0.001483... 2099.0
gs_admis... 复旦大学... 5.0 1.0 0.0 0.001440... 2319.0
10092510... 胡栋梁 1.0 1.0 0.0 1.097181... 82.0
qding@se... 'dingqi' 30.0 13.0 0.096153... 0.007661... 1507.0
10092510... 10092510... 2.0 1.0 0.0 0.001483... 2099.0
10092510... 10092510... 2.0 1.0 0.0 0.001483... 2099.0
10082510... 徐诚 1.0 1.0 0.0 0.001483... 1589.0
37564826... 顾晴 1.0 1.0 0.0 0.001440... 1616.0
10092510... 10092510... 2.0 1.0 0.0 0.001483... 2099.0
39156984... 李光耀 3.0 1.0 0.0 0.001440... 2218.0
Wed Jun 06 01:18:33 CST 2012腾讯微博分享--【广外"神(仙姐姐"创造零迷...

華東師範大學 软件学院 software engineering institute

Thank You!