网络用户浏览路径分析

网站分析、用户行为

北京 | 上海 | 广州 | 深圳 | 东京 | 硅谷 | 香港

肖嘉敏：jiaminshaw@gmail.com
张　翔：birdzhangxiang@gmail.com

Blog: www.shaw.cu.cc

2012-11-04

目录

准确度量，持续改进

—网站分析驱动目标达成

网站中发生了什么事情

- 用户从**哪里来**？**哪里进入**网站？又是**哪里离开**网站？
- 用户在网站中**寻找什么**？网站中的哪些内容、哪些页面最受用户欢迎？

网站中各部分功能是否正常

- 页面**布局合理**吗？网站**导航清晰**吗？各项网站**功能正常**吗？
- 页面内容是否合适？转化路径是否合理？

网站的改善与优化

- 哪些来源渠道的用户更有价值？网站哪方面存在问题，需要改进?
- 网站的运营策略是否有效？如何采取进一步行动优化?

我们正在做的站内用户行为分析

首页分析	站内搜索	浏览路径	单品访问	下单流程
首页流量、首页作为着陆页的比例。 分析首页点击分布、下一页访问路径，包括各去向页面占比。 顶部频道和左右侧导航点击；公告或排行的点击。 站内广告的点击和转化。	搜索频率，搜索内容排行，下单访客的搜索，对搜索结果的满意度。 类目搜索解析 关键词搜索解析， 搜索来源(首页搜索，频道搜索，内页搜索)	发现主流路径及其中的循环浏览，判断其合理性。 发现路径中的主要流失出口，判断是否存在主流路径上。 区分不同人群的访问路径	购物车与下单按钮的点击 访问商品的分类：大中小行业，品牌 确认下单商品	流程中各步骤的访客流失情况； 重新回到流程的途径； 离开流程的访客数量和去向 观察购物流程执行效率的变化趋势

目录

浏览路径 ⟹ 时间序列

通过序列分析探索用户上网模式
利用日志文件对用户上网模式进行分类

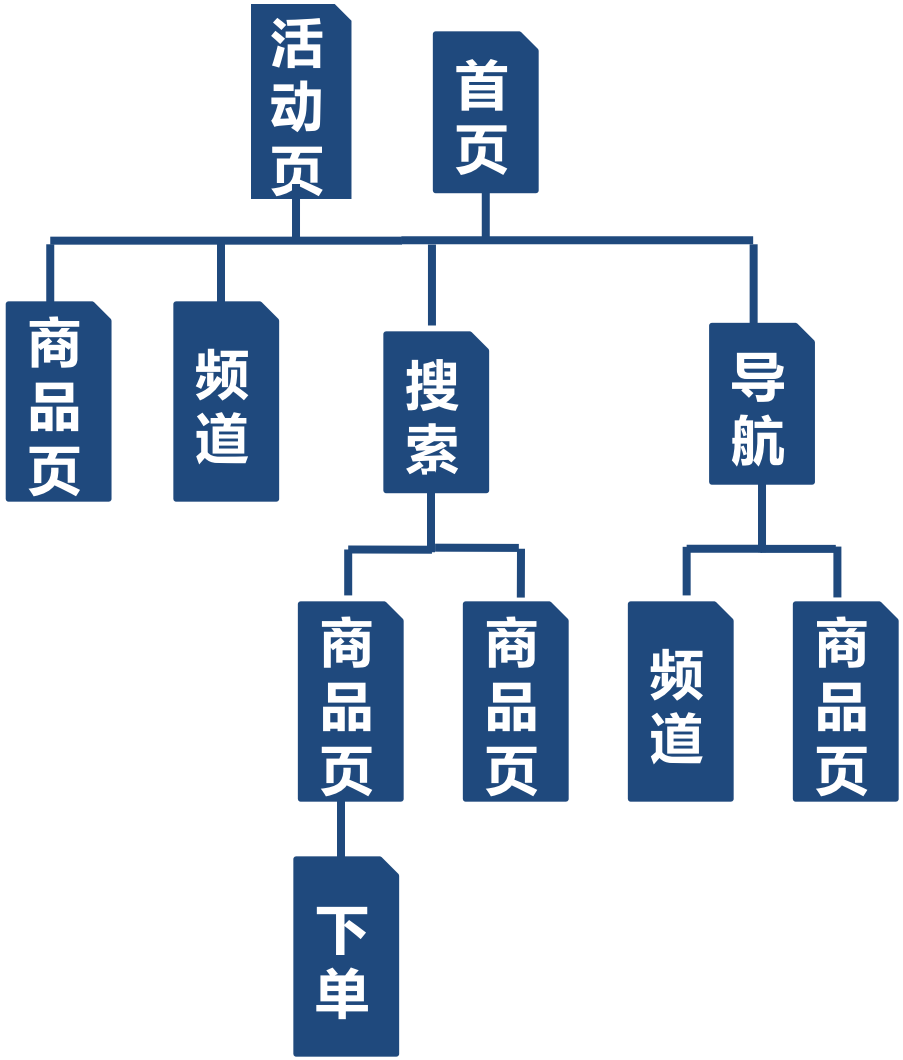
序列分析维度

- 如何从大量序列数据当中**提取感兴趣的特征序列**
- 如何**计算序列数据指标**，量化特征，进一步对序列数据进行统计分析
- 序列数据**可视化**
- 序列数据**相似性的度量**，以此基础继续一些探索性数据分析
- 识别具有**共同模式**的子集

URL归类

URL归类原则

- 按不同频道(垂直类)
- 按网站功能(电商)
- 按URL规则(门户)



用户日志转化序列数据对象

- Sample ID
- Datetime
- Domain
- Duration
- service

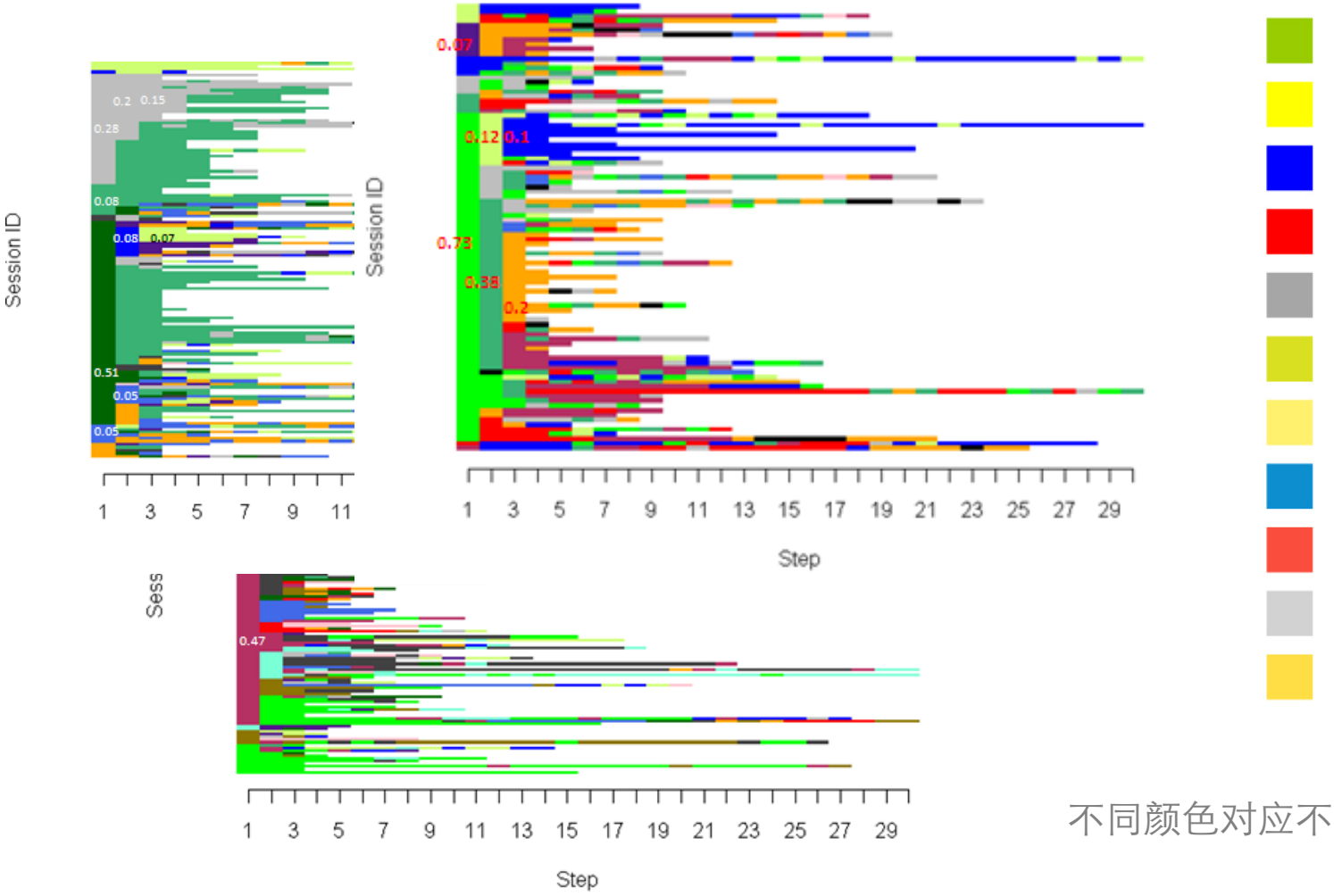
Code	Conversion	Example
STS	from/to	*Id* 18 19 20 21 22 23 24 25 26 27 101 S S S M M MC MC MC MC D 102 S S S MC MC MC MC MC MC MC
SPS	from/to	*Id* 1 2 3 4 101 (S,3) (M,2) (MC,4) (D,1) 102 (S,3) (MC,7)
DSS	to	*Id* 1 2 3 4 101 S M MC D 102 S MC
SPELL	from	*Id* *Index* *From* *To* *State* 101 1 18 20 S (single) 101 2 21 22 M (married) 101 3 23 26 MC (married with children) 101 4 27 27 D (divorced) 102 1 18 20 S (single) 102 2 21 27 MC (married with children)

```
> print(te, format='STS')
    Sequence
1167 0-0-0-0-0-0-0-0-0-3-6-6-6-6-6-6
514  0-1-1-1-1-1-1-1-1-1-1-3-6-6-6-6
```

```
> print(te, format='SPS')
    Sequence
[1] (0,9)-(3,1)-(6,6)
[2] (0,1)-(1,10)-(3,1)-(6,4)
>
```

序列数据可视化(站内)

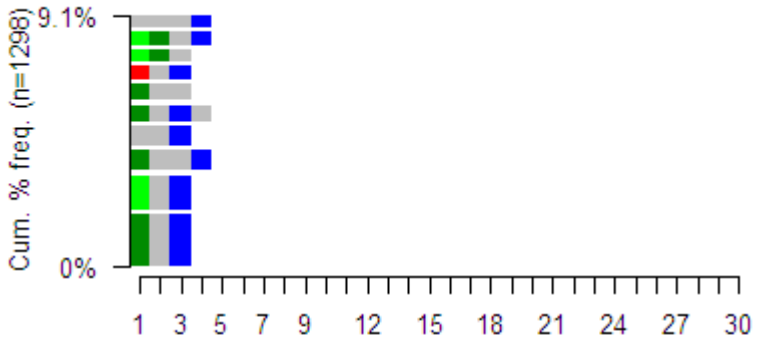
站内浏览路径汇总



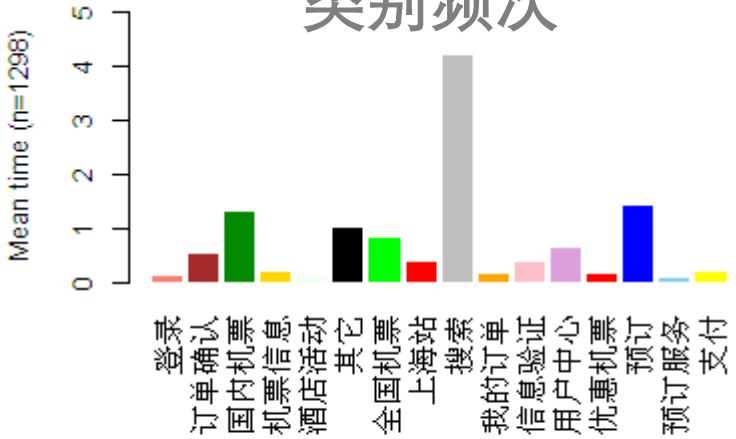
不同颜色对应不同URL类别

序列数据可视化(站内)

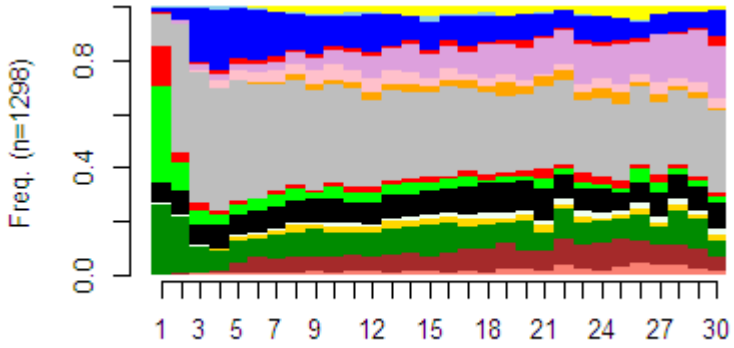
序列频次

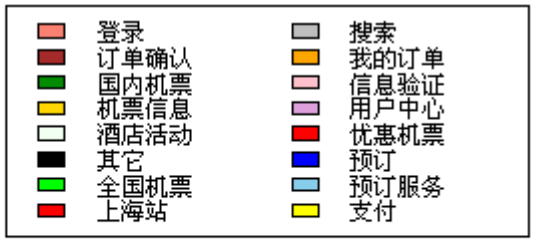


类别频次



类别分布



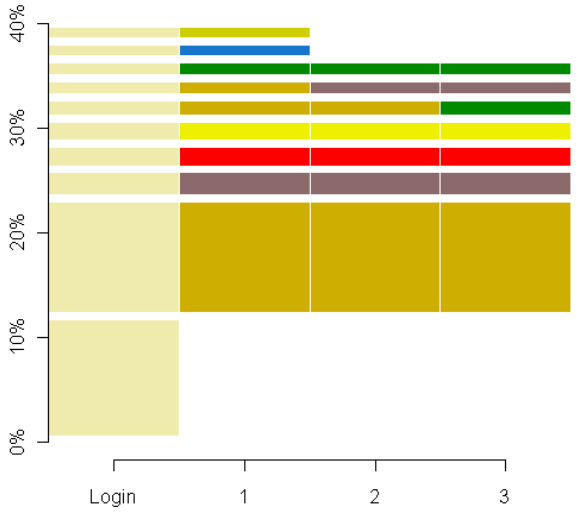


source: http://news.iresearch.cn/Zt/173489.shtml

序列数据可视化(站外)

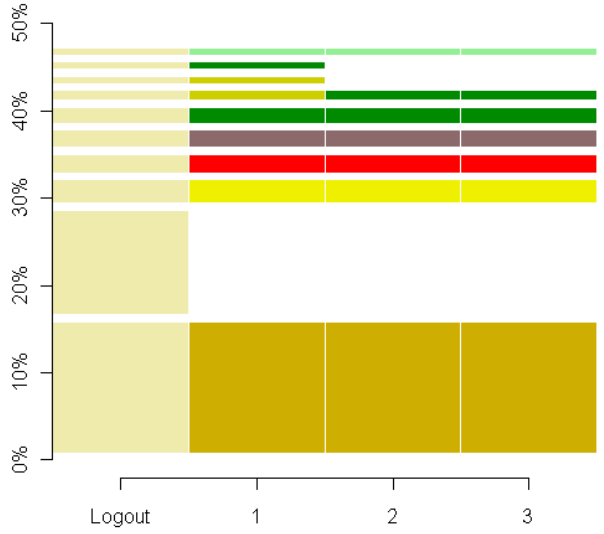
站外来源路径分析

- 进站前来源点
- 进站前精准路径
- 进站前模糊路径



站外去向路径分析

- 出站后去向点
- 进站前精准路径
- 出站后模糊路径



序列数据相似(异)性度量

评价序列相异性的两个重要方法：

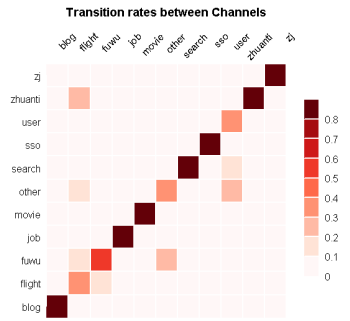
1. 计算它们之间匹配总量；
   Simple Hamming, Longest common prefix, Longest common suffix, Longest common subsequence

2. 两条序列相互转化的成本。
   Optimal matching, Hamming, Dynamic Hamming

序列数据相似性度量

转换比率 transition rates

$$p(s_j \mid s_i) = \frac{\sum_{t=1}^{L-1} n_{t,t+1}(s_i, s_j)}{\sum_{t=1}^{L-1} n_t(s_i)}$$



$n_t(s_i)$ 当t不是最后一个位置时， s_i状态的个数; $n_{t,t+1}(s_i, s_j)$为 t位置为s_i状态, t+1位置为s_j状态的个数。

替换成本 substitution-cost

$$2 - p(s_i \mid s_j) - p(s_j \mid s_i)$$

序列数据相似性度量

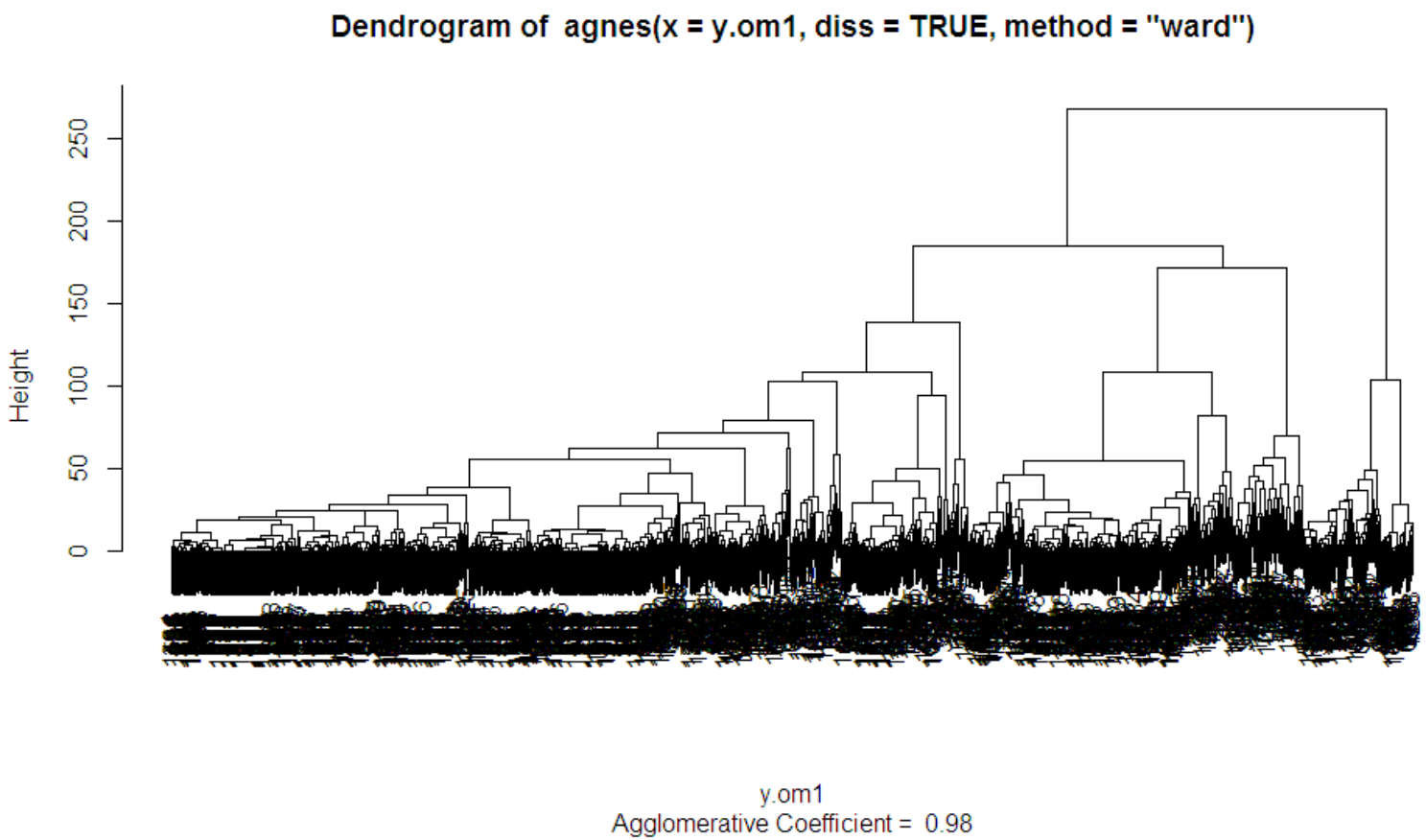
编辑距离 Edit distance

Optimal matching(OM), Generalized Hamming (HAM) and dynamic Hamming (DHD)

搜索引擎识别拼写错误，并提示正确的写法：从一个字符变到另一个字符主要有三种方式：替换一个字符、增加一个字符和删除一个字符，把这三种操作都看做一次字符的修改，两个单词的Edit Distance就是从一个单词变成另一个单词需要的最少字符修改次数。

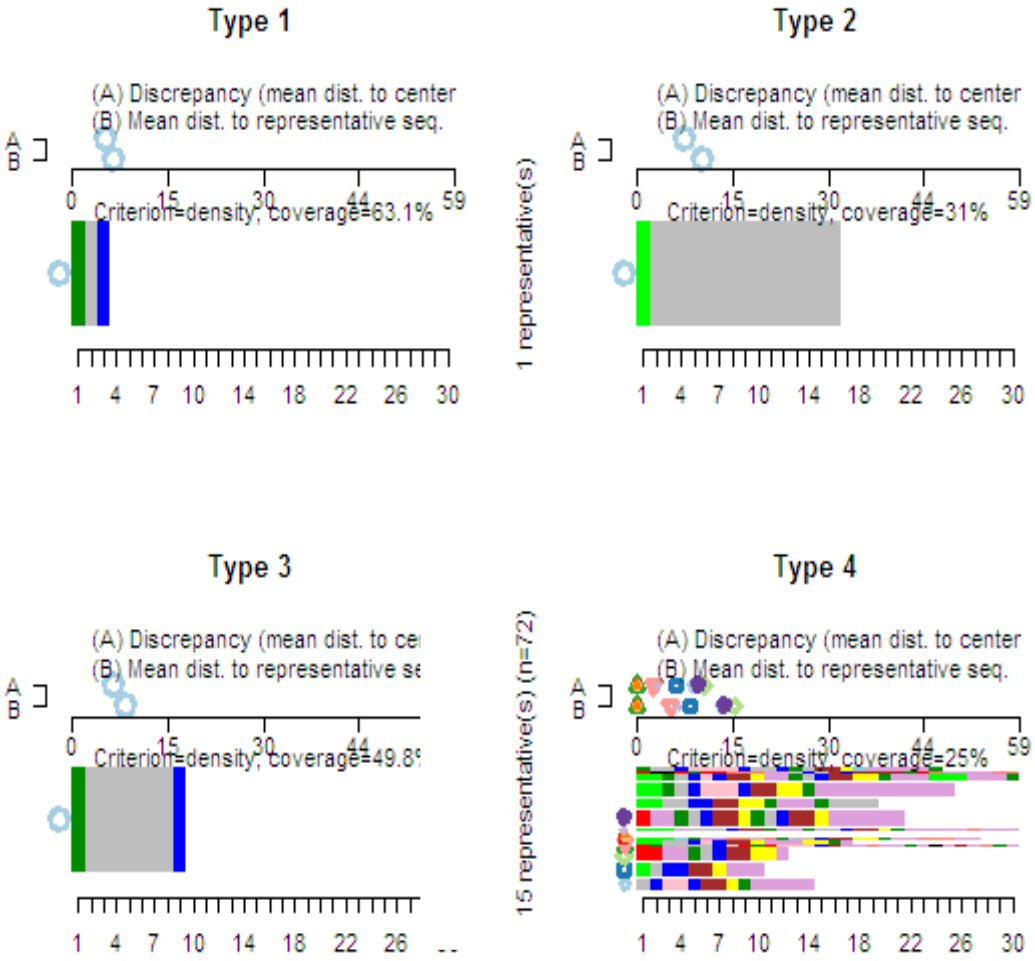
基于相似(异)性矩阵的序列分析

聚类分析



基于相似(异)性矩阵的序列分析

提取代表序列



一般的处理方法：

1. 提取出现频次最高的序列；
2. 提取中心度最高的序列

搜索代表序列方法

- **根据代表性得分对序列进行排序**
  - sequence frequency, neighborhood density, mean state frequency, centrality, sequence likelihood.

- **设定阈值删除冗余序列**
  - The redundancy threshold is set as a percentage (10% by default) of the maximum theoretical dissimilarity D_{max} between two sequences and the representative set will thus not contain any pair of sequences that are nearer each other than this threshold.

目录

同样适合R编程的经典语句

"我们当中大多数人接受的教育是，在编程时，要把一个任务细分成多个更小的步骤，**按一定的顺序执行程序**，进行想要计算。但是，如果也按这种思想来处理SQL编程，那么最终只能得到**平庸的结果**"

-----Microsoft SQL Server 2008技术内幕:T-SQL查询

Excel, sql, R切换自如

Excel 透视表

交互式报表，可快速合并和比较大量数据。旋转其行和列以看到源数据的不同汇总，而且可显示感兴趣区域的明细数据

sql 语句

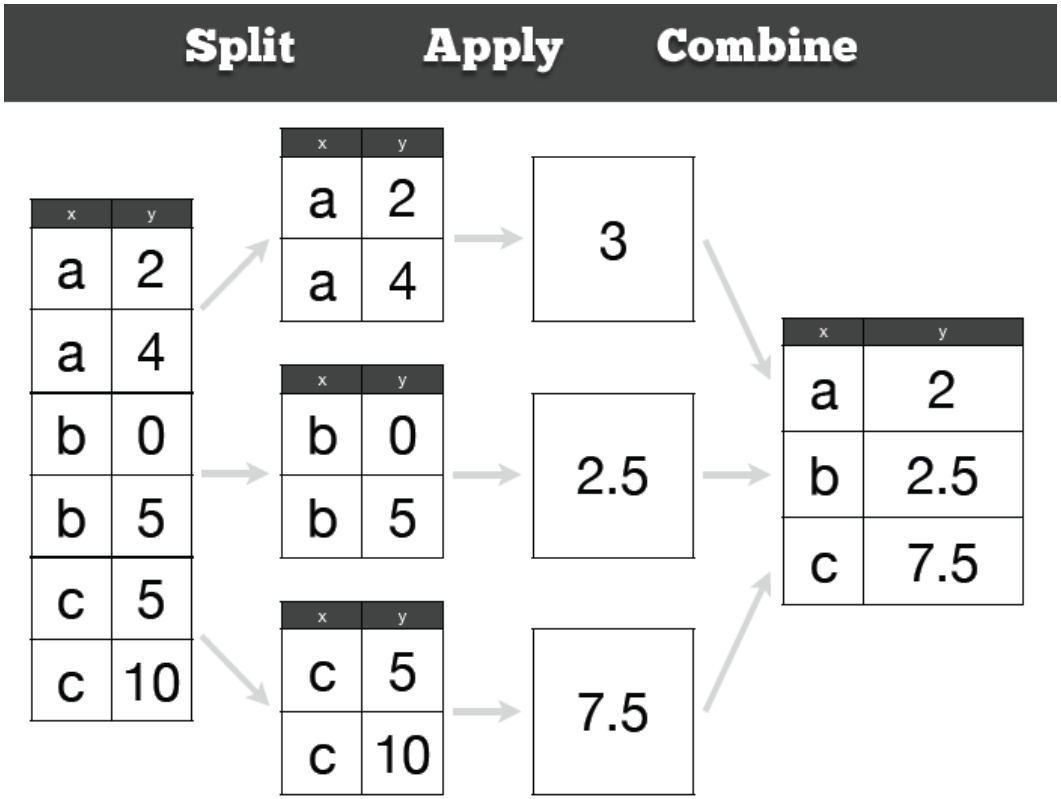
select语句中可以通过group by将行划分成较小的组，然后使用聚集函数返回每一个小组的汇总信息

R

apply系列, plyr, sqldf等扩展包

切片计算

Many problems involve splitting up a large data structure, operating on each piece and joining the results back together:**split-apply-combine**



Hadley Wickham, **Visualisation and data manipulation in R @ ebay**
http://courses.had.co.nz/11-ebay/

访次&上网出入口的设计

排序 -> 移位-> 比较

rank	ID	Time	domain
1	1	18:49:24	a
2	1	18:57:25	b
3	1	18:57:29	b
4	1	19:57:47	a
5	1	19:58:22	a
6	2	18:59:13	c
7	2	18:59:18	a
8	2	19:00:01	a
9	2	21:00:09	a
10	3	19:00:17	c

rank	ID	Time	domain
2	1	18:57:25	b
3	1	18:57:29	b
4	1	19:57:47	a
5	1	19:58:22	a
6	2	18:59:13	c
7	2	18:59:18	a
8	2	19:00:01	a
9	2	21:00:09	a
10	3	19:00:17	c
1	1	18:49:24	a

R 涉及apply系列函数；sql涉及分页查询，全连接

访问次数 用户访问该网站比前一次访问的时间间隔超过30分钟，访问次数加1次，在30分钟之内连续访问该网站页面，只算1次。

上网出入口 用户一天当中上网的起点各终点网站, 最大连续不在线时间间隔。

目录

聊天记录分析

表1 发言量排行榜

Rank	Names	Freq
1	肖嘉敏(61792715)	1281
2	李源栋(276868740)	592
3	阿铁(355665588)	404
4	钱海燕(278310114)	190
5	王静(52392252)	185
6	王昭林(158242136)	174
7	江(278310998)	133
8	周和根(278746367)	128
9	包军(281209168)	91
10	邓海梅(4197562)	87

表2 活跃天数排行榜

Rank	Names	Freq
1	肖嘉敏(61792715)	143
2	李源栋(276868740)	49
3	钱海燕(278310114)	23
4	缪静(277844672)	20
5	周和根(278746367)	20
6	阿铁(355665588)	17
7	江(278310998)	16
8	王昭林(158242136)	14
9	邓海梅(4197562)	13
10	文德权(120849170)	11

表3 发起群聊次数排行

Rank	Names	Freq
1	肖嘉敏(61792715)	72
2	李源栋(276868740)	16
3	阿铁(355665588)	6
4	钱海燕(278310114)	6
5	周和根(278746367)	6
6	王强(276964812)	4
7	文德权(120849170)	4
8	夏老师(493594996)	4
9	陈珍珠(240014170)	3
10	邓海梅(4197562)	3

表4 结束群聊次数排行

Rank	Names	Freq
1	肖嘉敏(61792715)	53
2	李源栋(276868740)	26
3	江(278310998)	12
4	缪静(277844672)	10
5	钱海燕(278310114)	7
6	王昭林(158242136)	5
7	周和根(278746367)	5
8	包军(281209168)	3
9	刘志鹏(250429640)	3
10	梅林茂(275587572)	3

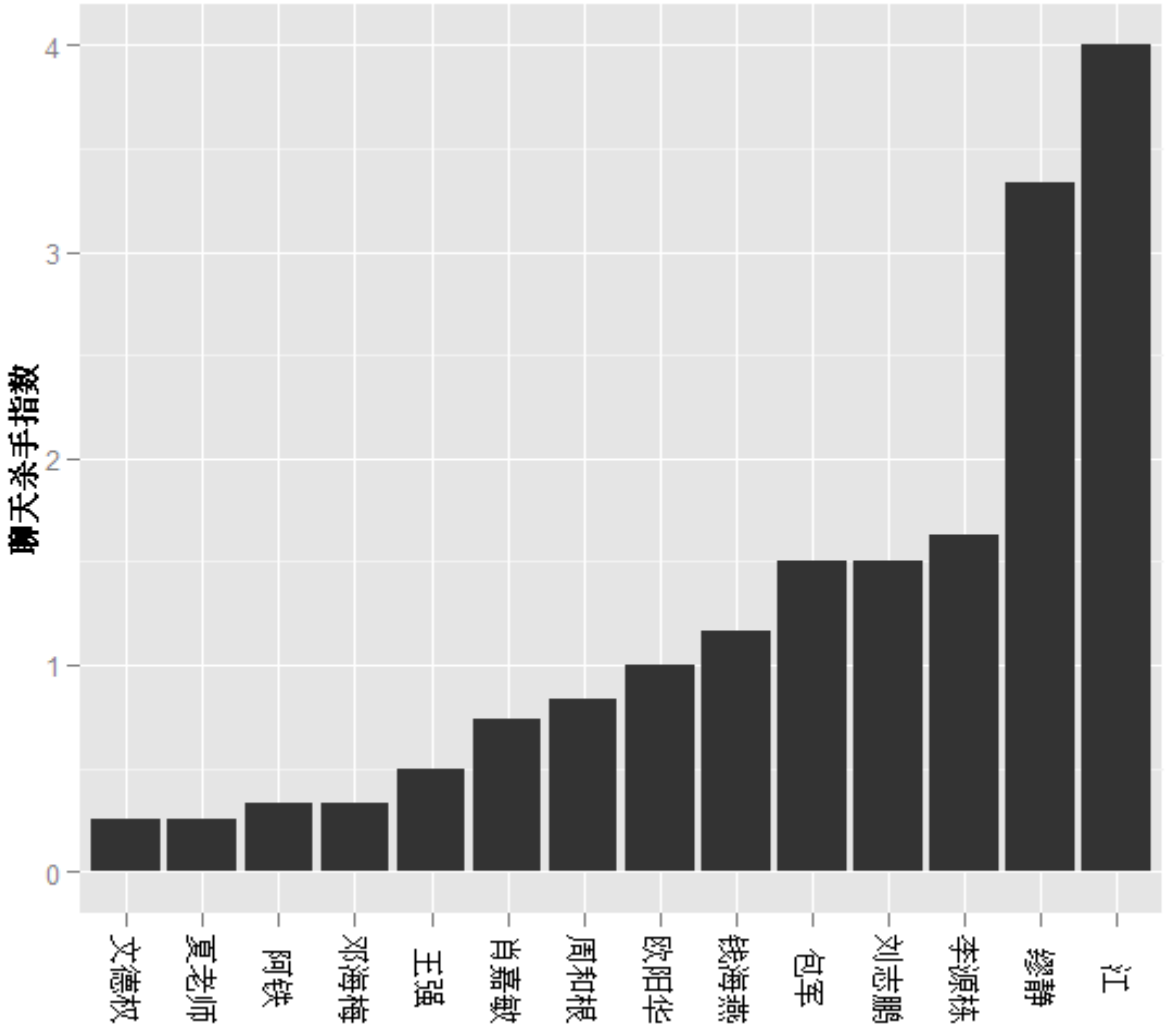
数据源：结合多人聊天记录文件

发言量：QQ消息数

活跃天数：参与群聊的天数

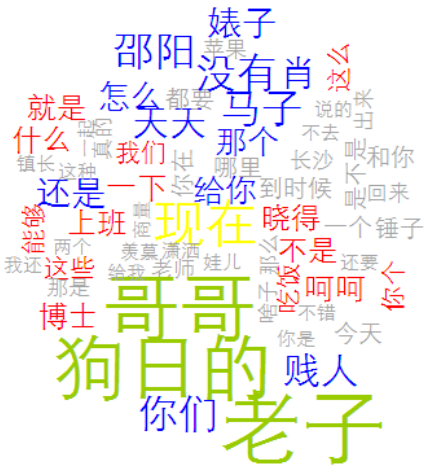
群聊话题次数：将所有数据按时间排序，间隔超过30分钟认为是一个新话题的开始。每次群聊话题必须有两个及以上的人参与。

谁是聊天杀手



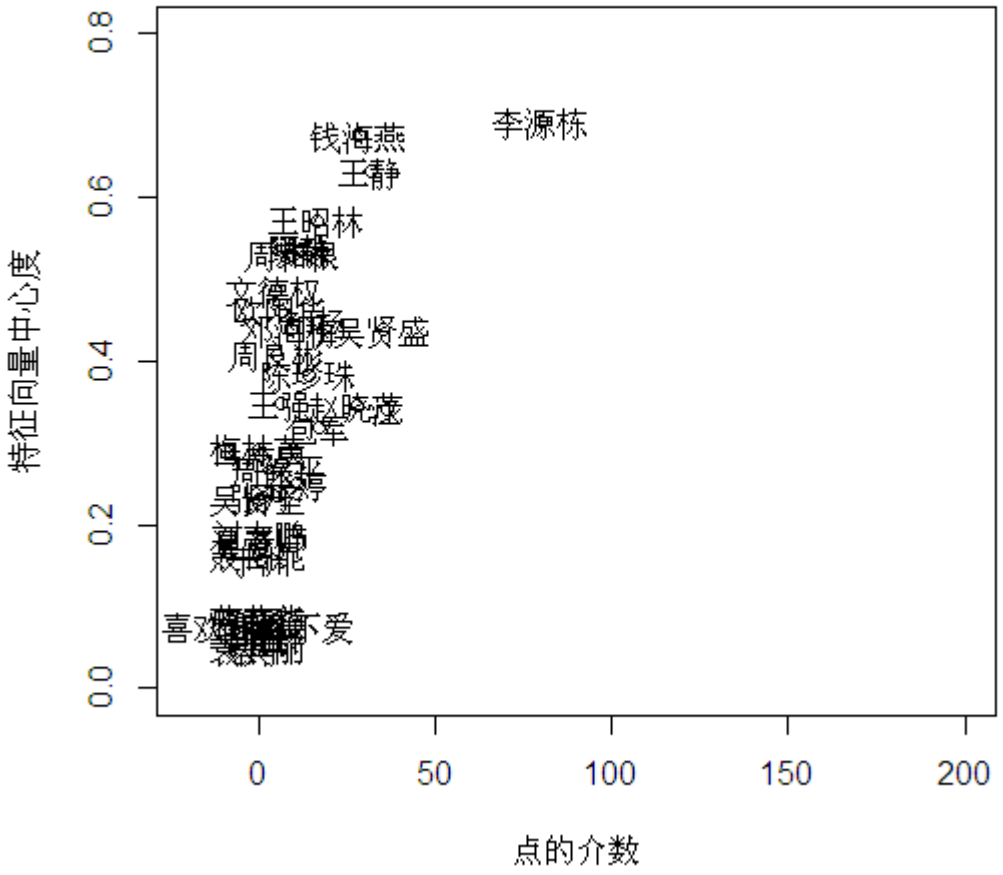
看看自己曾经做过多少次聊天杀手，即话一出，群内至少沉默半个小时以上。

大家都说些什么？



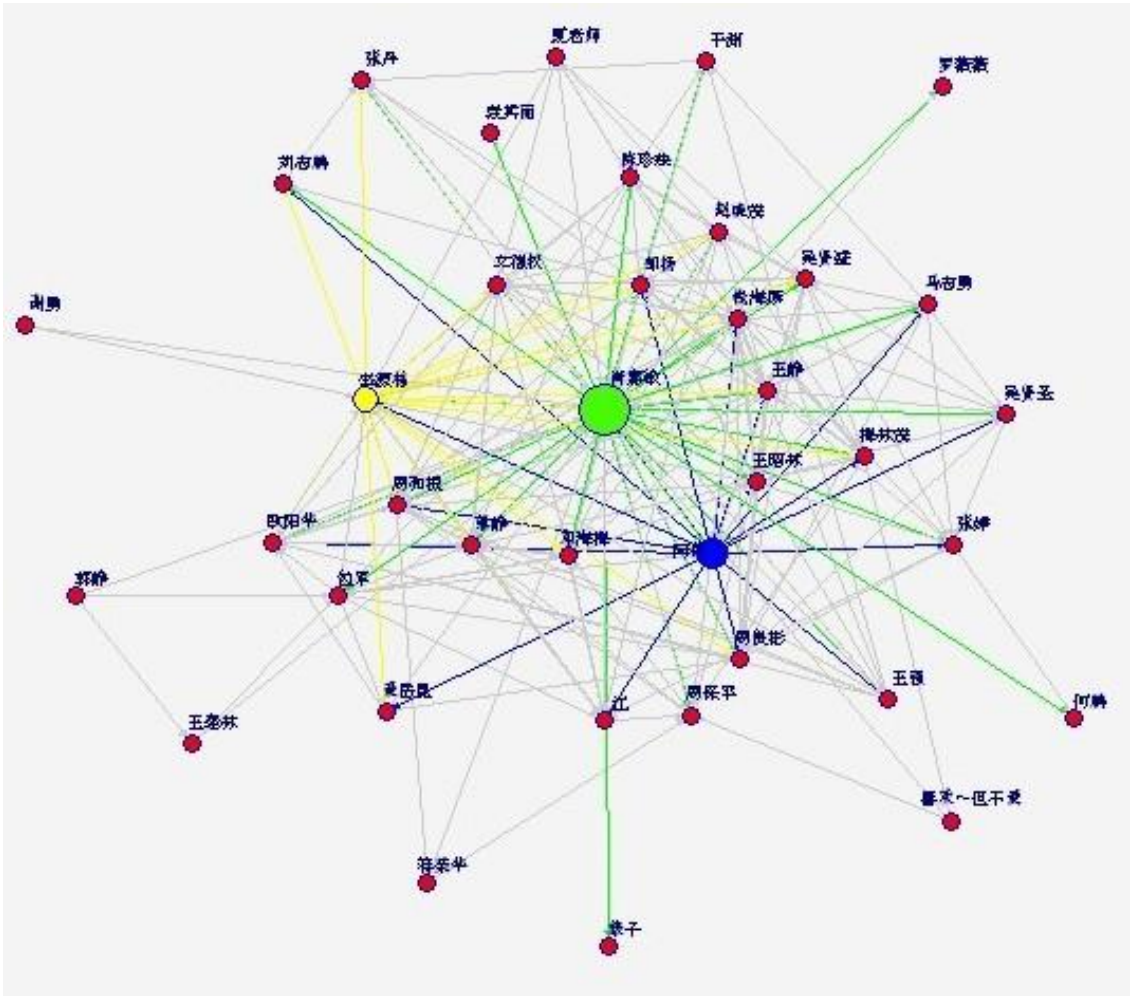
群里大家聊的最多的是 **“我们”、“你们”、“现在”、“怎么样”、“做什么”。**

节点的重要性



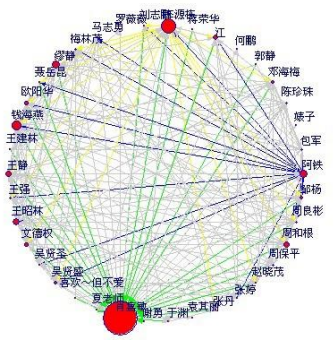
根据网络的 Eigenvector Centrality 和 Betweenness Centrality（即vertex和edge信息）寻找各个节点的重要性，“肖嘉敏”地位有点类似异常值，已删除。

群关系网



参与话题的群成员当作网络中的节点，在一个会话中有互动(参与聊天)即各个成员之间存在关系，也就形成连接不同节点的边。

如图，整个群成员之间主要由“肖嘉敏”，“李源栋”，“钱海燕”等几点重要的点将大家紧密联系在一起。



选择艾瑞 选择可以信任的合作伙伴

iResearch
艾瑞咨询集团

艾瑞十年
2002-2012