社交网络分析

Social network analysis

1 PART 01
起源

2 PART 02
特性与模型

3 PART 03
技术亮点

4 PART 04
Demo

章智斌、穆方舟、何金鹏、甘正宇、胡志鹏

社交网络的起源

这个名词是1954年由J. A. Barnes 首先使用 ("Human Relations" , 在章节 Class and Committees in a Norwegian Island Parish 内)。

社交网络的发展历程

1954年，J. A. Barnes首次提出社交网络，由此出现社交网络一词。随后出现了社交网络发展的三次浪潮。

01

第一次浪潮：

1971年，由ARPA（Advanced Research Projects Agency）项目科学家发出了世界上第一封电子邮件。电子邮件的问世拉开了社交网络发展的序幕。一直到上世纪90年代，社交网络都处于缓慢发展时期，期间陆续有社交平台问世。

社交网络的发展历程

1954年，J. A. Barnes首次提出社交网络，由此出现社交网络一词。随后出现了社交网络发展的三次浪潮。

02

第二次浪潮：

21世纪初，社交网络进入了新的发展时期。2002年，Friendster的出现，开创了商业社交网站的先河，也是全球首个用户规模达到100万的社交网站。2003年，MySpace出世，社交网络的发展再一次被带动。短短几年时间，社交网络发展有了一个全新的面貌。

社交网络的发展历程

1954年，J. A. Barnes首次提出社交网络，由此出现社交网络一词。随后出现了社交网络发展的三次浪潮。

03

第三次浪潮：

04年，Facebook诞生。Facebook发展迅猛，几年时间里跻身全球最受欢迎的社交网站之列，现在成为了全球最大的社交平台。

05年，YouTube成立，其用户量早已突破10亿。2017年，YouTube进入了BrandZ最具价值品牌100强。

06年，Twitter成立。作为微博客的社交应用平台，Twitter在用户之间受到很大的欢迎，现在全球的用户量超过5亿。

07年，Tumblr成立，轻博客这种全新社交形态由此出现。

六度分割原理

理论指出：你和任何一个陌生人之间所间隔的人不会超过六个，也就是说，最多通过六个人你就能够认识任何一个陌生人。

Social Networking Services将现实中的人际关系搬到了互联网上，让世界上的任何一个人都能联络彼此。

Social Media Mining is the process of representing, analyzing, and extracting meaningful patterns from social media data

数据分析/挖掘的结果往往能较好地解释
what/where/when/how many/how often
但较难回答why

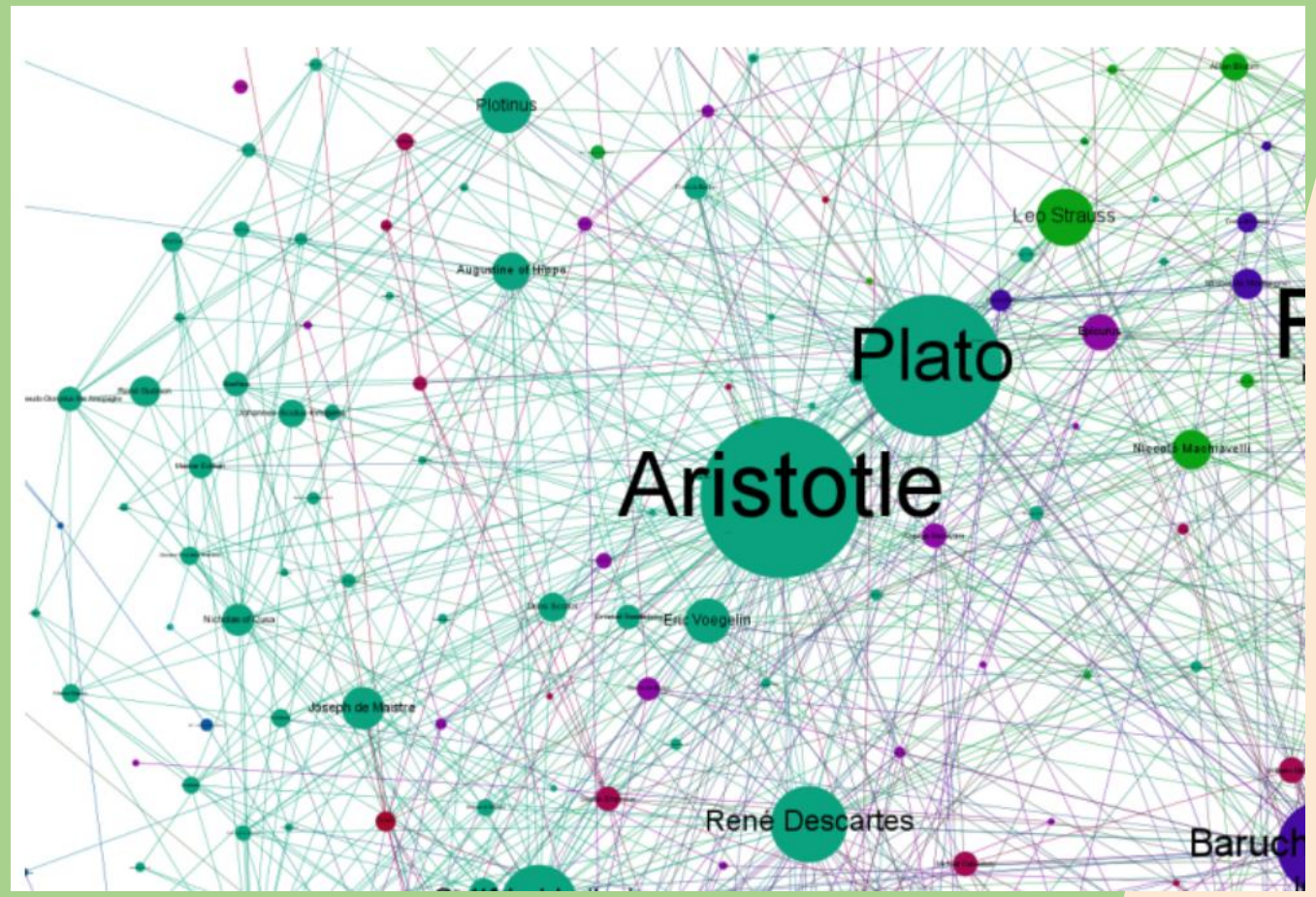
我们面对的是社会学的问题，我们要用交叉的思想去寻找解决方案，并利用数学和信息学的手段来解决它

特性

Characteristic

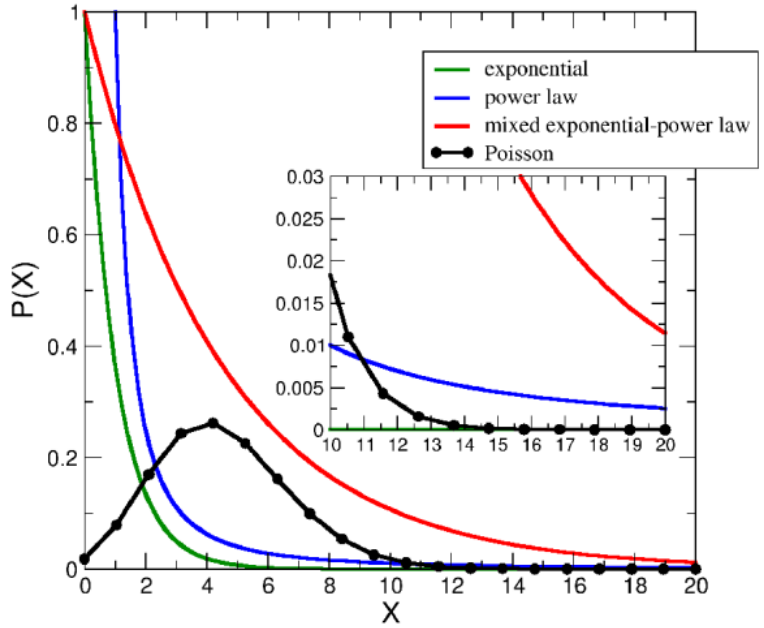
统计特性 | 度的幂律分布

Power law distribution of degree



特性
Characteristic

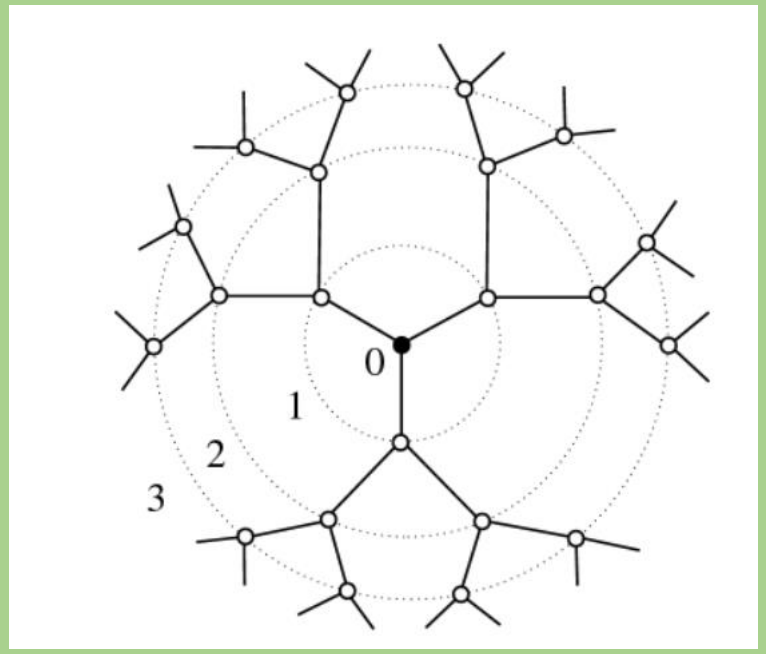
统计特性 | 度的幂律分布
Power law distribution of degree



poisson: $p(k) = \frac{\lambda^k}{k!} e^{-\lambda}$, exponent: $p(x) = Ce^{-\lambda x}$, power law: $p(x) = Cx^{-\alpha}$

特性
Characteristic

统计特性 | 较小的直径
Smaller diameter



假设每个节点连接 z 个邻居

$$z^d = N,\ d = \log N / \log z$$

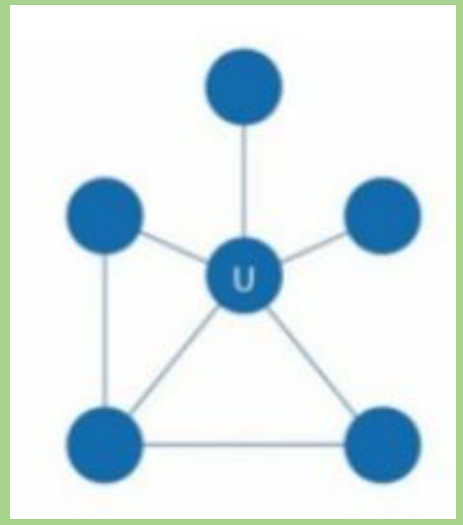
$$N \approx 6.7 \text{ bln},\ z = 50 \text{ friends},\ d \approx 5.8.$$

特性
Characteristic

统计特性 | 聚类系数
Clustering coefficient

聚类系数用于描述网络中与同一节点相连的节点也互为相邻节点的程度

$$CC(u) = \frac{2R_u}{k_u(k_u - 1)}$$



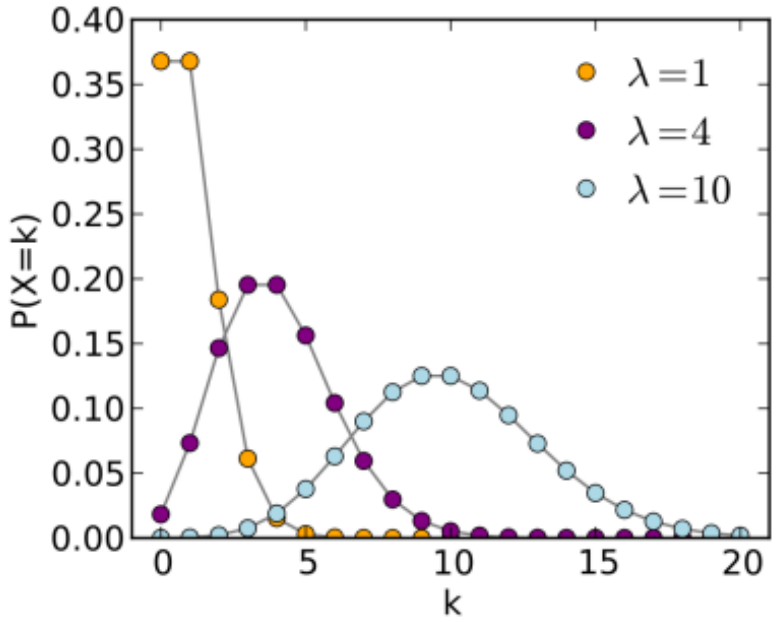
模型

Model

基础模型 | 随机模型

stochastic model

模型G(n,p)中，随机连接节点构成一个图。
图中每个连边彼此独立，连接的概率为p。



$$P(k_i = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad \lambda = pn$$

模型

Model

基础模型 | BA模型

BA model

（1）t=0:初始状态是n0个节点

（2）网络成长：在每一个时刻t={1，2，3，…}，加一个有m（m<n0）条边的节点

（3）择优连接：新的节点连接到已存在的节点i的概率取决于节点i的度，即

$$\Pi(k_i) = \frac{k_i}{\sum_i k_i}$$

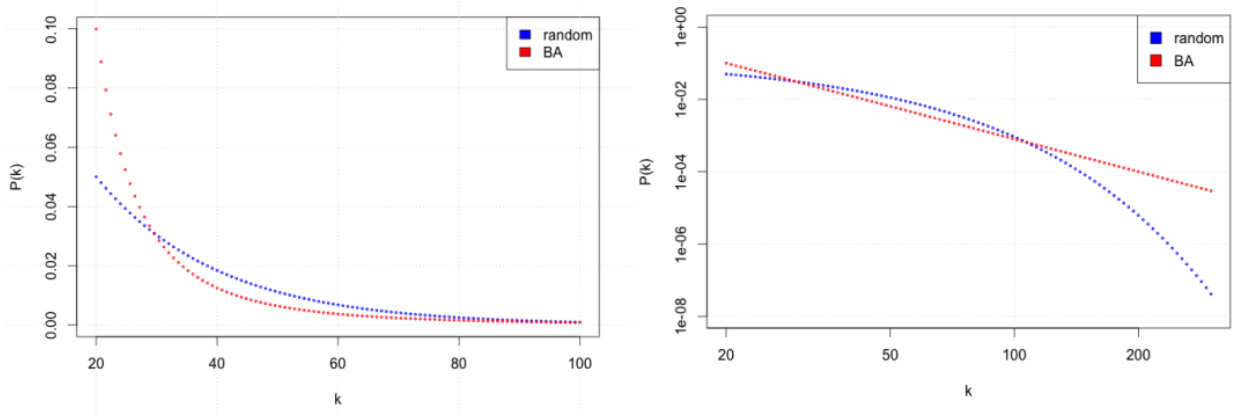
（4）重复（2）（3）直至网络达到规模

模型

M o d e l

基础模型 | BA模型

BA model



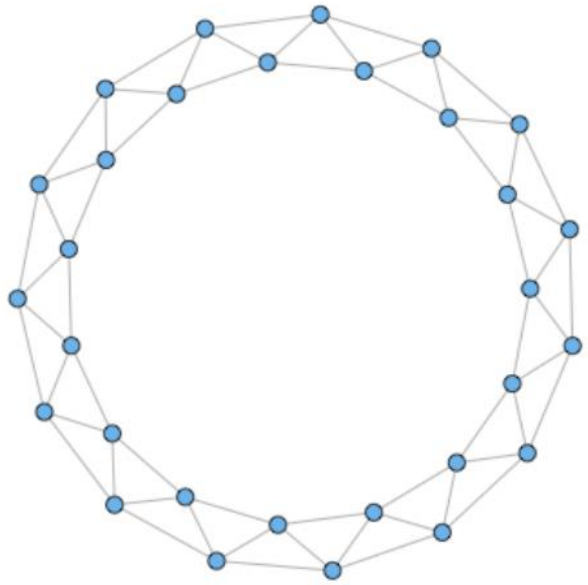
$$BA: P(k) = \frac{2m^2}{k^3}, \qquad RG: P(k) = \frac{e}{m} e^{-\frac{k}{m}}$$

经过计算，BA模型的聚集系数C为

$$C \sim N^{-0.75}$$

模型
M o d e l

基础模型 | WS模型
WS model

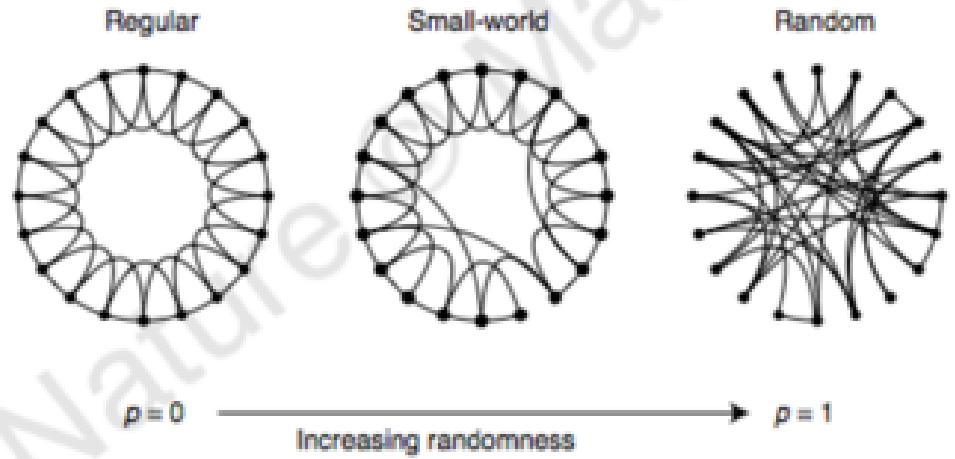


Clustering coefficient $C = 1/2$
Graph diameter $d = 8$

模型
M o d e l

基础模型 | WS模型
WS model

（1）初始网络是一个有n个节点的环形网格状网络
（2）指定概率p，并对初始网络中的每条边，以概率p进行重连，重连时随机选择一个节点进行替换

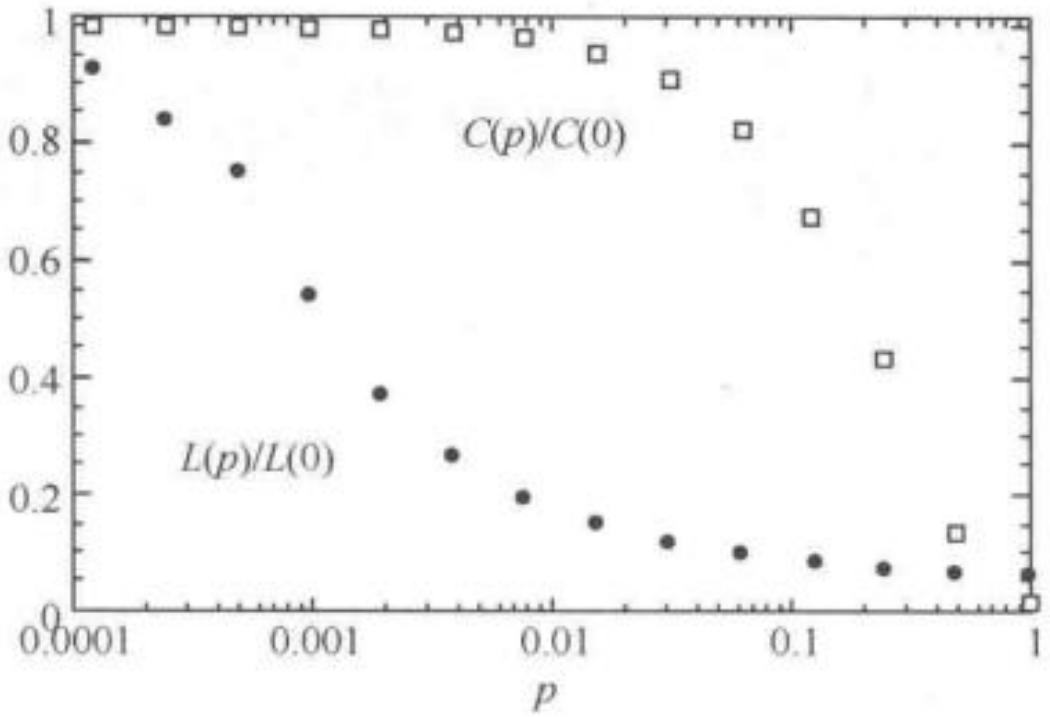


模型

M o d e l

基础模型 | WS模型

WS model



总结

Summary

	Random	BA model	WS model	Empirical networks
P(k)	$\frac{\lambda^k e^{-\lambda}}{k!}$	k^{-3}	poisson like	power law
C	$\langle k \rangle / N$	$N^{-0.75}$	const	large
$\langle L \rangle$	$\frac{\log(N)}{\log(\langle k \rangle)}$	$\frac{\log(N)}{\log\log(N)}$	$\log(N)$	small

社交平台及拓展

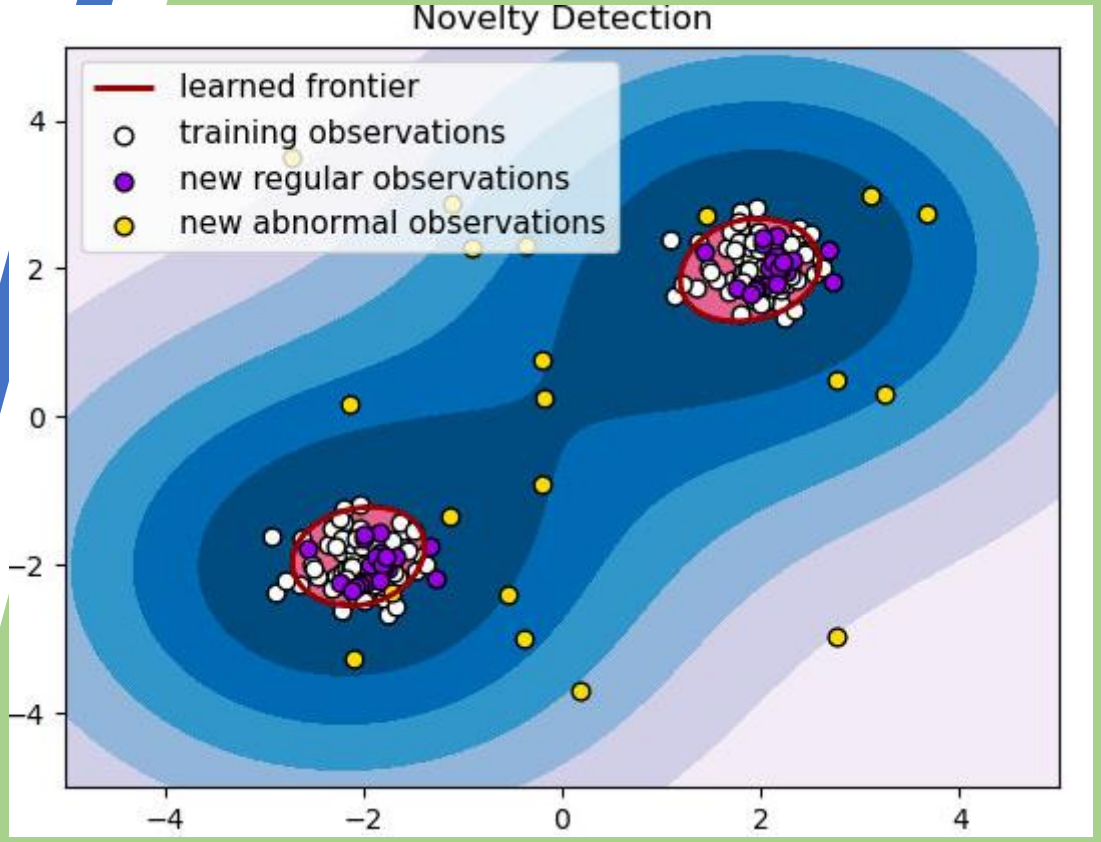
双向好友:好友形成闭环,需要对方认证,从而保护隐私。但双方认证成本较大,容易引发骚扰现 象的发生,容易在建立关系时断裂和失控。

单向好友:好友形成U型,无需对方认证,降低使用成本,关系建立顺畅。但较之于双向好友关系 不对等,马太效应巨大,隐私问题困扰。

反向好友:以上两者的变种,被关注者获得主动权,但发出邀请成本较大,故而限制好友数量,建 立关系成本较大,也会在建立关系的过程中断裂,但较之双向关系可能性较低,隐私性极强。

弹性好友:以上三者的改进形态,也是我认为的最优秀的设计，使用成本为0，好友关系建立成本为 0,关系断裂可能性极低,关系建立顺畅。缺陷是隐私保护不足。

跨社交网络的同一用户识别算法

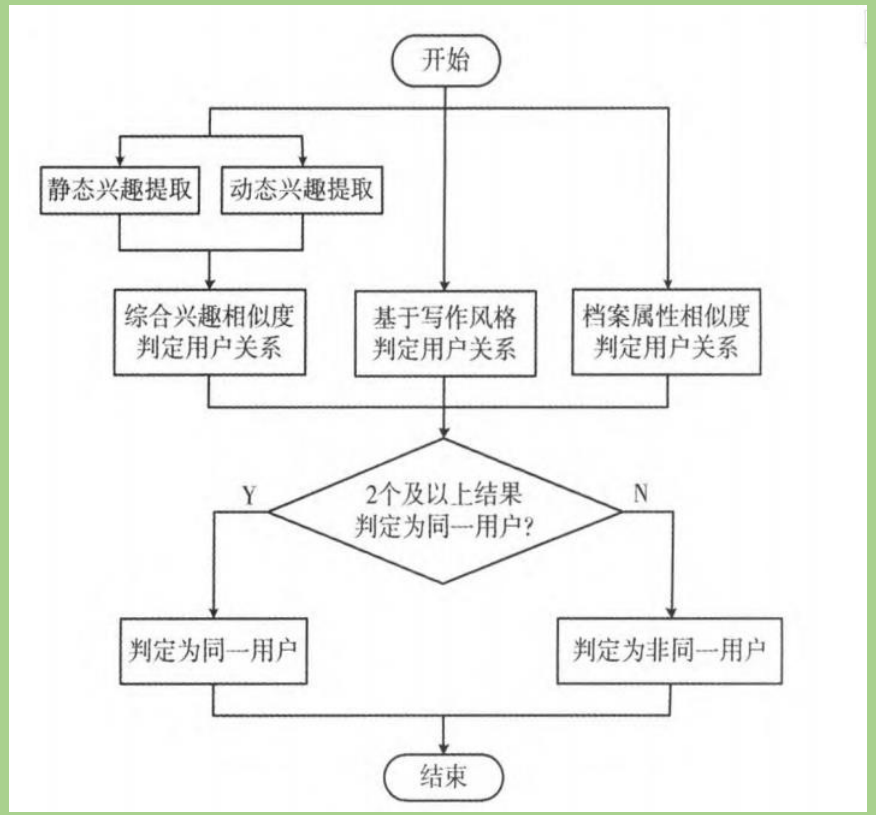


静态兴趣提取

$$sim_{static}(V_s, V_d) = \frac{|V_s.static \cap V_d.static|}{|V_s.static \cup V_d.static|}$$

动态兴趣提取

Single—pass聚类算法



用户筛选器设计

$$\mathrm{sim}_{lev}(V_s.\mathrm{name}, V_d.\mathrm{name}) = 1 - \frac{d_{lev}(V_s.\mathrm{name}, V_d.\mathrm{name})}{\max(\mathrm{len}(V_s.\mathrm{name}), \mathrm{len}(V_d.\mathrm{name}))}$$

其中，Vs．name表示源社交网络用户名，Vs．name表示目标社交网络用户名，len(Vs．name)、len(Vd．name)表示用户名的编辑距离。利用计算得到的用户名相似度筛选另一网络中的所有目标用户Vd。

LDA模型提取用户兴趣

用户在社交平台中发表的单条文本字数较少。难以直接挖掘主题，故先利用聚类算法，将同一用户的文本内容按话题分为多个簇，将同一簇内的文本合并为一个文档。首先利用Single-pass聚类算法对源用户Vs的候选目标用户V`d发表的文本内容分别进行聚类，得到Vs的簇文本集

$$W_u^* = \{w_{u1}^*, w_{u2}^*, \cdots, w_{uQ}^*\}$$

和聚类个数Ns与另一 网络中的候选目标用户V`d的簇文本集，然后将源用户V。的簇文本集W。与候选目标用户V。的簇文本集形i合并，通过Single-pass聚类算法得到聚类个数K，作为LDA模型的参数提取文本一主题矩阵，并引入基于时间的遗忘因子

$$\gamma = e^{-h \times (\text{nowTime} - \text{lastTime})}$$

其中，h是调节系数，用来调节衰减的速度；nowTime是当前系统时间，lastTime是每一个聚类簇对应的最近一次更新时间。可得源用户Vs与候选目标用户V`d的遗忘因子与文本一主题矩阵的对应关系如下

$$
\begin{array}{cc} & \begin{matrix} T_1 & T_2 & \cdots & T_K \end{matrix} \\ \begin{matrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_{N_s} \end{matrix} & \begin{matrix} w_1^* \\ w_2^* \\ \vdots \\ w_{N_s}^* \end{matrix} \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{22} & \cdots & p_{2K} \\ \vdots & \vdots & & \vdots \\ p_{N,1} & p_{N,2} & \cdots & p_{N,K} \end{bmatrix} \end{array}
$$

$$
\begin{array}{cc} & \begin{matrix} T_1 & T_2 & \cdots & T_K \end{matrix} \\ \begin{matrix} \gamma_{N_s+1} \\ \gamma_{N_s+2} \\ \vdots \\ \gamma_M \end{matrix} & \begin{matrix} w_{N_s+1}^* \\ w_{N_s+2}^* \\ \vdots \\ w_M^* \end{matrix} \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{22} & \cdots & p_{2K} \\ \vdots & \vdots & & \vdots \\ p_{M1} & p_{M2} & \cdots & p_{MK} \end{bmatrix} \end{array}
$$

$$
P(T_k)=\frac{\sum_{i=1}^{N}[\gamma_i \times p_{ik}]}{\sum_{k=1}^{K}\sum_{i=1}^{N}[\gamma_i \times p_{ik}]} \quad k=1,2,\cdots,K
$$

$$
\mathrm{sim}_{\mathrm{dynamic}}(V_s, V_d')=1-D(\theta_s \| \theta_d')
$$

$$
\mathrm{sim}_{\mathrm{ints}}(V_s, V_d')=\alpha \cdot \mathrm{sim}_{\mathrm{static}}(V_s, V_d')+(1-\alpha)\mathrm{sim}_{\mathrm{dynamic}}(V_s, V_d')
$$

总结

这个算法利用用户名设计筛选器，降低算法计算开销，并通过改进后的TF—IDF算法设计过滤器滤除噪声文本，提高用户动态兴趣提取准确性，实现了在微博和豆瓣平台之间的同一用户识别。

模型

Model

社交网络挖掘和搜索 | 概述

introduction

社交网络搜索和挖掘是社交网络研究的重要部分，它们在推荐算法，舆情分析与控制，甚至在网络安全等方面都有应用。而传统的数据挖掘的和搜索方法已经不能很好的适用。

模型

Model

社交网络挖掘 | 综述

summary

目前在社交网络分析与挖掘方向主要的研究有社交网络结构建模、信息传播研究、社区发现、情感分析及事件监测等。目前在社交网络分析与挖掘方向主要的研究有社交网络结构建模、信息传播研究、社区发现、情感分析及事件监测等。

社交网络结构建模:基于图：ER随机图模型到小世界、无标度模型、六度分割。基于非图：基于Agent的模型和基于贝叶斯网络多式联运的行为模型。

模型

M o d e l

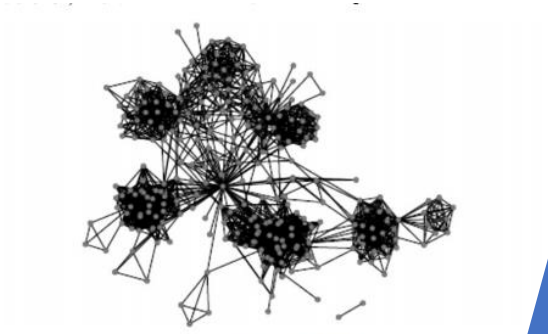


图 1　典型的社区图

$$\mathrm{sim}(D_1, D_2) = \cos\theta = \frac{\sum_{i=1}^{n} (A_i \times B_i)}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

式中：$\mathrm{sim}(D_1, D_2)$表示相似度函数，D_1 和 D_2 表示文档内容，而A_i和B_i表示两个 n 维向量。

社交网络挖掘 | 综述

summary

社交网络结构建模:基于图：ER随机图模型到小世界、无标度模型、六度分割。基于非图：基于Agent的模型和基于贝叶斯网络多式联运的行为模型。

社交网络信息传播：解释模型：　INFOPATH
预测模型：基于图：IC,LT,T-BaSIC 基于非图:OIS-SIRS,SIS,Seinr

社区发现（左图）：目前的研究方法通常集中在图论的相关 算法，例如图分割、图聚类、图的修剪等方法。

情感分析：目前学术界基于情感分析的研究方法主要集中在社交网络文本的情感词方法。

社交网络事件监测:目标是对社交内容中的事件和热点话题的自动识别和已知话题的持续跟踪

模型

M o d e l

社交网络搜索 | 综述

summary

在社交网络搜索的关键技术上，目前搜索技术的研究热主要集中在传统的搜索引擎技术和社交网络搜索技术。传统技术包括PageRank，HITS算法等。在社交网络搜索技术，有FaceBook开发了备受欢迎的知识图谱搜索技术等。

搜索索引：对倒排索引的改进，使之更加适应社交网络。
排序研究：主要是基于传统排序算法的改进和引入情感计算、社会影响力等因素，提高搜索排序算法的精度和准确性。

模型

Model

社交网络挖掘和搜索 | 问题 problems

在线社交网络挖掘领域中的各热点话题都面临着如何研究更有效方法应对在线网络的大规模化、复杂化等带来的效率和质量问题。

在线社交网络搜索技术对特定对象精准搜索的研究还存在不足。实现对用户的社交行为分析、搜索意图理解的智能化、智慧化搜索是一个亟待解决的问题。

如何结合数据挖掘技术实现基于时空特性的社交网络搜索也是一个需要解决的问题。

产品展示

Demonstration

准备工作

获得《三国演义》的部分文本

```
chapters = get_sanguo() #文本列表，每个元素为一章的文本
print(chapters[0][:106])
```

获取结果

第一回　宴桃园豪杰三结义　斩黄巾英雄首立功　滚滚长江东逝水，浪花淘尽英雄。是非成败转头空。　青山依旧在，几度夕阳红。　白发渔樵江渚上，惯看秋月春风。一壶浊酒喜相逢。　古今多少事，都付笑谈中。

产品展示

Demonstration

知识库

部分实体的指称及类别获取

```
entity_mention_dict, entity_type_dict = get_sanguo_entity_dict()
print("刘备的指称有：",entity_mention_dict["刘备"])
print("刘备的类型为",entity_type_dict["刘备"])
print("蜀的类型为",entity_type_dict["蜀"])
print("蜀的指称有",entity_mention_dict["蜀"])
```

获取结果

```
刘备的指称有:['刘备', '刘玄德', '玄德', '使君']
刘备的类型为 人名
蜀的类型为 势力
蜀的指称有 ['蜀', '蜀汉']
```

产品展示

Demonstration

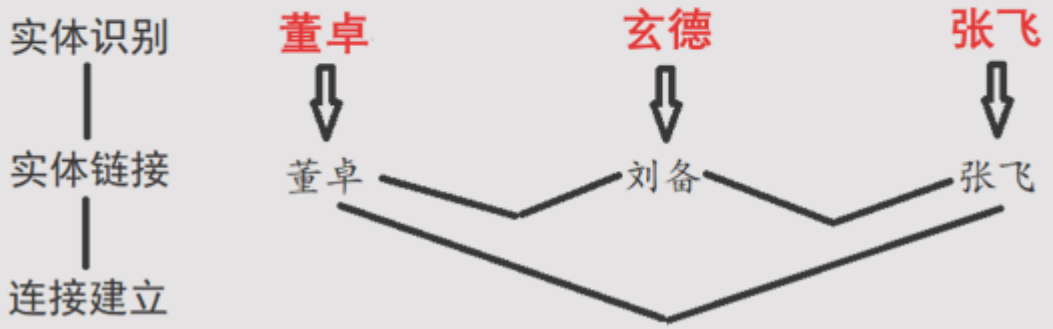
连接

将各个实体进行连接

```
ht = HarvestText()
ht.add_entities(entity_mention_dict, entity_type_dict)
# 加载模型
print(ht.seg("誓毕，拜玄德为兄，关羽次之，张飞为弟。
",standard_name=True))
```

利用邻近共现关系。每当一对实体在两句话内同时出现，就给它们加一条边。

且说董卓字仲颖，陇西临洮人也，官拜河东太守，自来骄傲。当日怠慢了玄德，张飞性发，便欲杀之。



产品展示

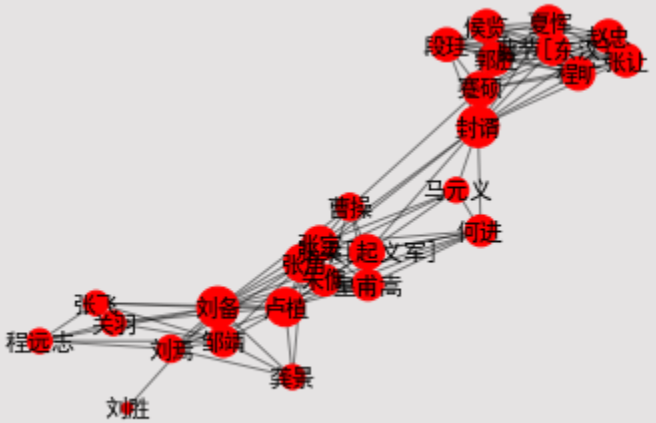
Demonstration

网络绘制

姓名替换，分句

```
doc = chapters[0].replace("操","曹操")      # 将缩写改为全称
ch1_sentences = ht.cut_sentences(doc)      # 分句
doc_ch01 = [ch1_sentences[i]+ch1_sentences[i+1] for i in
range(len(ch1_sentences)-1)]   #获得所有的二连句
ht.set_linking_strategy("freq")
# 对所有人物建立社交网络
G = ht.build_entity_graph(doc_ch01, used_types=["人名"])
```

挑选主要人物画图



Summary

感谢观看