在线社交网络

> 参考书籍：
>
> 1. 方滨兴等. 在线社交网络分析[J]. 2014.
> 2. 《The Crowd》（乌合之众）：当个人变成群体中的一员时，他们将失去其个体身份
> 3. 大连接.[美] 尼古拉斯·克里斯塔基斯（Nicholas A. Christakis）/ [美] 詹姆斯·富勒（James H. Fowler）
>
> 相关比赛
>
> - **TREC比赛** 文本检索会议

0x00 绪论

- **社交网络**（social network）是由众多节点构成的一种社会结构，节点通常是指个人或组织，而节点间关系代表着各种社会关系。 ——维基百科
- **在线社交网络**是一种在信息网络上由社会个体集合及个体之间的连接关系构成的社会性结构，包含关系结构、网络群体与网络信息三个要素。
- 内容安全大事件
  - 阿拉伯之春
  - 美国大选（特朗普打败希拉里）
- **分析要求**：“快”、“准”、“全”

结构

> 针对社交网络节点海量、结构复杂性和多维演化性等特点，研究社交网络建模方法，共性特征分析方法，以及社交网络(虚拟社区）的发现方法及其演化规律。
>
> 相关会议：
>
> - World Wide Web Conference 国际互联网大会
> - ACM Internet Measurement Conference 互联网测量会议
> - Physical Review Letters 物理学的核心中的核心期刊
> - SIGKDD 前身 KDD（Knowledge Discovery and Data Mining，知识发现与数据挖掘）

社交网络结构特征分析及建模**

通过用节点表示人，边表示人际交往关系，可将社交网络形式化描述为G={V，E}。

1. **统计特征**

- **度分布**：网络中心度为k的节点占比用p(k)即度分布函数来表示，若其正比于幂指数函数，则网络是服从幂律分布的无标度网络。累积度分布函数，即度不小于k的节点的概率分布，其也服从幂律分布。
  - **累积分布函数(Cumulative Distribution Function，CDF)**，又叫分布函数，是概率密度函数的积分,$F_X(x) = P(X \leq x)$；互补累积分布函数（complementary cumulative distribution function、CCDF），是对连续函数所有大于a的值，其出现概率的和，$F(a) = P(x > a)$。
- 平均路径长度：定义为网络中任意两个节点间最短路径的平均值，也叫做网络的平均距离或网络的特征路径长度。

    - 社交网络中任意两个用户i 与用户j 的距离dij定义为这两个用户间所有通路中所经过边数最少的一条的度，也叫做**最短路径长度。网络直径**定义为网络中所有最短路径中长度最大值。
- 网络密度（Density） 用于刻画网络中节点间相互连边的密集程度，定义：网络中实际存在的边数与可容纳的边数上限的比值。通常情况下，大规模网络的密度要比小规模网络的密度小.当网络为全连通时， d=1 ； 当网络中不存在连边时，d=0

$$d(G) = \frac{2L}{N(N-1)} \tag{1}$$

- **聚集系数**（Clustering Coefficient）用于描述网络中与同一节点相连的节点间也互为相邻节点的程度，对于节点 vi，其聚集系数 Ci表示与它相邻的节点间也存在连接的平均概率。即一个人所有朋友之间彼此也互为朋友的概率。网络平均聚集系数定义为网络中所有节点聚集系数的平均值。

$$C_i = \frac{2e_i}{k_i(k_i-1)} \tag{2}$$

- 介数（Betweeness）用来描述网络中节点承载最短路径数的能力。节点（或边）的介数等于网络中所有最短路径中经过该节点（或边）的概率之和，描述了节点在网络中的影响力与中心性程度

2. 特性分析

- 小世界现象。人类社会是一个具有较短路径长度特点的小世界型网络。
- **无标度特性**（度分布服从幂律分布）节点度分布不存在有限衡量分布范围的特征标度。大多数节点只有少量连边，少数节点有大量连边。体现了**异质性。**
    - 幂律（Power Laws）
        - **Zipf定律**（每个单词出现的频率与它的名次的常数次幂存在简单的反比关系）与**Pareto定律**（20%的人口占据了80%的社会财富）都是简单的幂函数,我们称之为幂律分布。网页被点击次数的幂律分布其幂指数在 0.60-1.03之间,而网站访问量的幂律分布其幂指数则接近1。统计物理学家习惯于把服从幂律分布的现象称为无标度现象，即系统中个体的尺度相差悬殊,缺乏一个优选的规模。“**长尾理论**”即是幂律的口语化表达。
        - 幂律分布表现为一条斜率为幂指数的负数的直线,这一线性关系是判断给定的实例中随机变量是否满足幂律的依据。
    - 服从幂律分布的变量x具有**重尾现象**，当x 趋向于正无穷时，X 取到 X>x的概率是指数分布的低阶无穷小
    - 在双对数坐标系上，是一条斜率为-k的直线
- **同配性** 反映了网络中度相近节点间相互关联的程度。度相关性表示一个节点的度与其邻居节点度之间的相关性（"门当户对"）。网络建立初期，一般具有同配特征，随着用户群体规模的不断扩大，许多网络表现出由同配演化为异配的现象。传统交友型社交网络同配系数通常为正；发布订阅网络（如YouTube、新浪微博等）的同配系数可能为负
    - 邻居平均度，节点 v 的邻居平均度定义为，即相连接节点度比值的累积

$$k_{nn,i} = \frac{1}{k_i}\sum_j a_{ij}k_j \tag{3}$$

    - 同配系数r
- 互惠性 衡量网络中两个节点形成**相互双向连接**的程度。交友类在线社交网络的互惠系数通常比较高；微博类在线社交网络的互惠性较差
- 强联系：对应朋友(Friend)关系， 彼此之间具有高度的互动 。弱联系：对应认识(Acquaintance)关系 。 弱连接则较能够在不同的团体间传递非重复性的讯息，使得网络中的成员能够增加修正原先观点的机会。 在社交网络中，关系密切的朋友之间的连接成为强连接，关系疏远的朋友之间的连接成为弱连接。

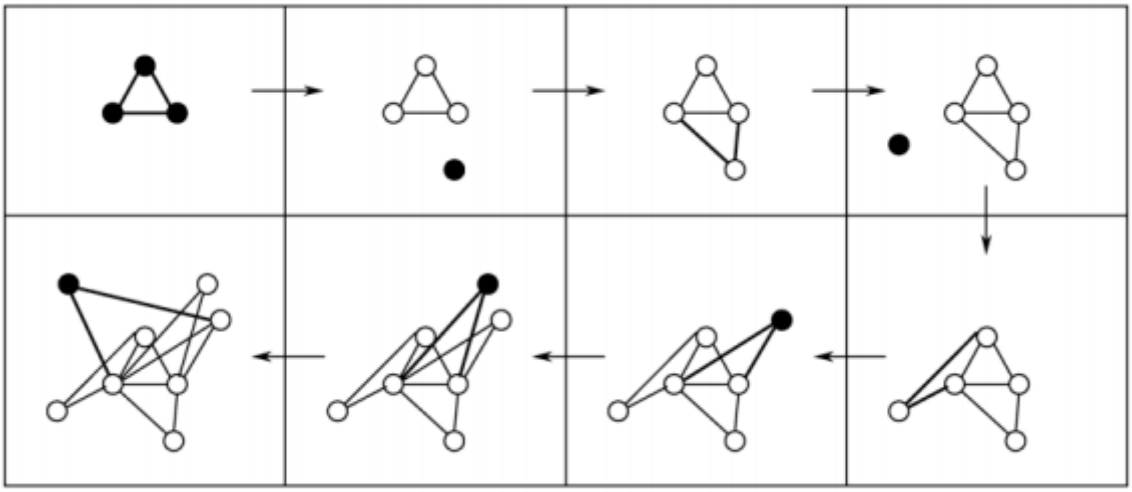
3. 生成模型

- WS模型，可生成小世界网络，NW模型对其小幅改动以维护网络的连通性

小世界现象揭示了客观世界许多复杂网络具有的特征，即较大的平均聚集系数和较短的平均最短路径长度

- **算法描述**
  - **（1）以一个含有n个节点，每个节点度为2k的环形栅格网络为初始网络，网络中的每个节点与其位置上最邻近的2k个节点相连，其中k是大于零的整数（通常k的值较小）；**
  - **（2）指定一个概率p，并对初始网络中的每条边，以概率p对该边进行重连（重连时随机选择一个节点对该边所连接的一个节点进行替换），新的连接保证不出现自连接和重复连接。**

- BA模型，通过"**择优依附**"(preferential attachment)机制模拟网络中资源的富集效应，表现幂律分布特点



虚拟社区发现技术与方法**

> 社区(Community):任何基于协作关系的有机组织形式，存在着具有局部紧密连接特性的节点集合

弱连接假设：研究发现，强连接出现在虚拟社区内部而弱连接出现在虚拟社区之间。

三元闭包特性：虚拟社区内部存在大量的三元闭包结构（三角形）。

定义

1. 基于子图的局部定义

社区结构是复杂网络节点集合的若干子集，每个子集内部的节点之间的连接相对紧密，而不同子集的节点之间的连边相对稀疏

2. 基于网络模块度的全局定义

网络中连接两个同类型的节点的边（同一社区内部的边）的比例，减去在同样（社区）结构下任意连接这两个节点的边的比例的期望值

3. 基于节点相似度的定义

社区内部的节点都是相似的，社区间的节点相似性低

衡量社区发现算法准确度的数字评价指标

- 模块度：通过比较现有网络与基准网络(随机网络)在相同社区划分下的连接密度差来衡量网络社区划分的优劣

  假设 A 是复杂网络的邻接矩阵，k_v 表示节点 v 的度数，即 $k_v=\sum_w A_{vw}$ 。在对应的基准网络中，一条边 (v,w) 存在的概率为 $\frac{k_v k_w}{2m}$，其中，m 表示网络图 A 中连边的数目。模块度的完整数学表达公式如下：

  $$Q=\frac{1}{2m}\sum_{vw}\left[A_{vw}-\frac{k_v k_w}{2m}\right]\delta(c_v,c_w) \qquad (3\text{-}1)$$

  其中，c_v 表示节点 v 所属的社区。如果 $i=j$，$\delta(i,j)=1$；反之，$\delta(i,j)=0$。该公式的数学意义为，网络中同一社区内部的边的比例与在同样社区结构下基准网络内部边的比例的期望值之差。模块度值越高，则复杂网络中社区划分的结果越好。

  - 模块度的最大值对应的社区划分结构并不一定是最佳的社区划分结果
  - 在很多情况下，存在一个潜在的最小社区尺度，任何尺度小于该值的社区结构均会对模块度的优化造成负面影响。模块度优化方法存在分辨率问题
- NMI（归一化互信息/规范化互信息） NMI利用信息熵来衡量算法划分的社区结构和预先已知的社区结构之间的差异。NMI值越大，则表明获得的社区结构划分越好，当该值达到最大值1时，说明算法发现的社区结构与已知社区结构完全一致。

  NMI 是基于混合矩阵（Confusion Matrix）N 来计算的数字指标。给定两个社区划分 $a=(a_1,a_2...a_n)$，$b=(b_1,b_2...b_n)$，其中 a_p，b_p $(p=1,2,\cdots,n)$ 表示第 p 个节点在两个划分中的社区编号，n 表示网络节点数量。NMI 计算公式如下：

  $$\mathrm{NMI}=\frac{-2\sum_{i,j}N_{ij}\ln\left(\frac{N_{ij}n}{N_{i.}N_{.j}}\right)}{\sum_i N_{i.}\ln\left(\frac{N_{i.}}{n}\right)+\sum_j N_{j.}\ln\left(\frac{N_{.j}}{n}\right)} \qquad (3\text{-}3)$$

  式中，$N_{i.}$ 表示矩阵 N 中第 i 行元素的总和，$N_{.j}$ 表示矩阵 N 中第 j 列元素的总和。

- **互信息（MI）** $$I(X;Y)=\sum_{y\in Y}\sum_{x\in X}p(x,y)\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right),$$
- **标准化互信息（NMI）**
  - **用熵做分母将MI值调整到0与1之间** $$U(X,Y)=2R=2\frac{I(X;Y)}{H(X)+H(Y)}$$

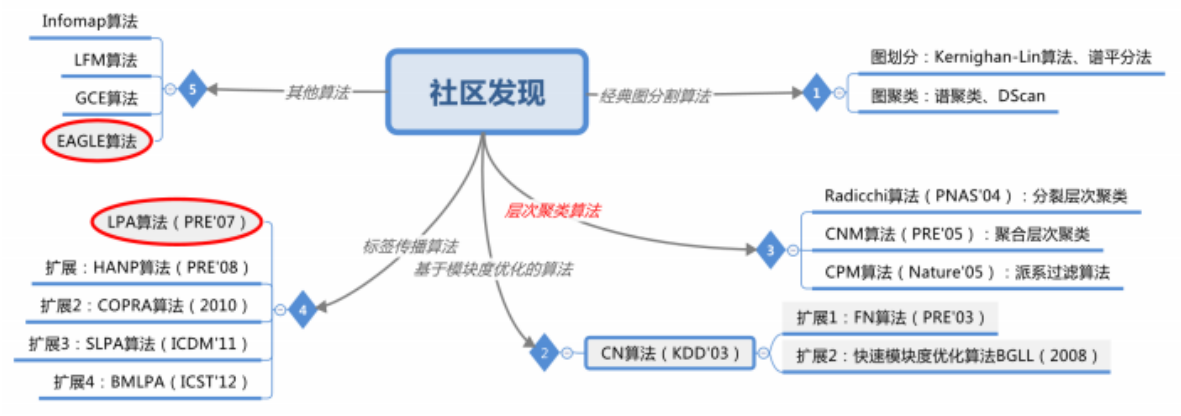
其中，H(X)和H(Y)分别为X和Y的熵

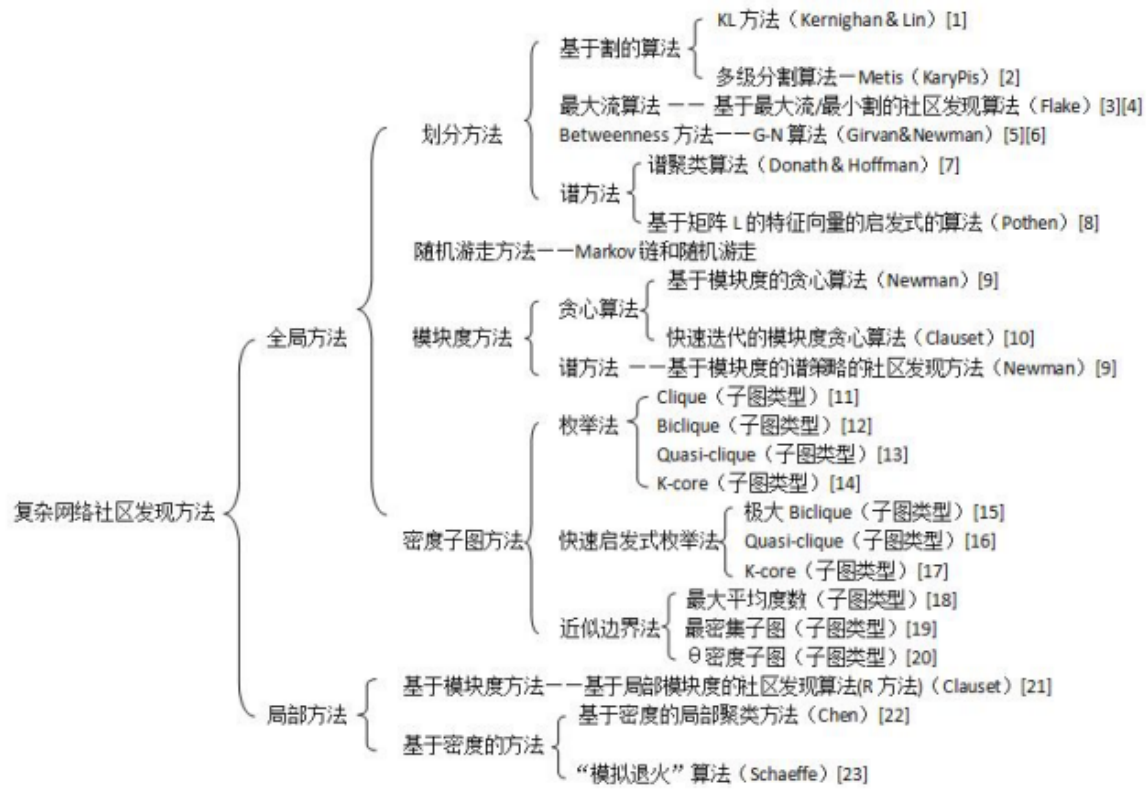
$$H(X)=\sum_{i=1}^{n}p(x_i)\,I(x_i)=\sum_{i=1}^{n}p(x_i)\log_b\frac{1}{p(x_i)}=-\sum_{i=1}^{n}p(x_i)\log_b p(x_i),$$

- Rand index 表示在两个划分中都属于同一个社区或者都属于不同社区的节点对的数量与所有可能的节点对数量的比值。与 Rand index 原理相似的一个数字指标是 Jaccard index，表示在两种划分中都被划入同一子集的元素对的数目占划分到同一子集与没有划分到同一子集的比例。

常用的典型数据集

- 实际网络基准图：从实际社交网络中抽象提取出来，具有明显的社区结构，但往往受到各种因素的影响，并非完全符合社区的规律。
- 人工网络基准图：基于复杂网络中节点度的幂律分布特性和小世界原理等拓扑性质所构建，如GN人工网络和LFR人工网络（参数多但更灵活更符合真实拓扑）





静态计算发现算法

> 计算全部节点的某种划分组合是否满足全局优化目标

模块度（modularity）最优化

> 通过最大化模块度Q来获得网络最优的社区划分，困难：网络所有可能的划分数量是巨大的，从中找到最优是NP难问题

- 经典贪心算法FN（又称贪婪算法）：局部最优解，去掉网络中所有的边，每次加入一条边，计算社区模块度，选择模块度增长最大或者减小最少的。总的时间复杂度是 O ((m+n)n)
    - 最优化模块度的启发式算法：模拟退火法、极值优化算法

- 快速模块度优化：贪心算法AGAIN、层次性贪心算法（基于模块度增量最大化标准决定合并社区，构建新的网络，并不断重复），不需要提前设定网络的社区数，适合超大规模网络中的社区发现

多目标优化算法

基于概率模型的算法

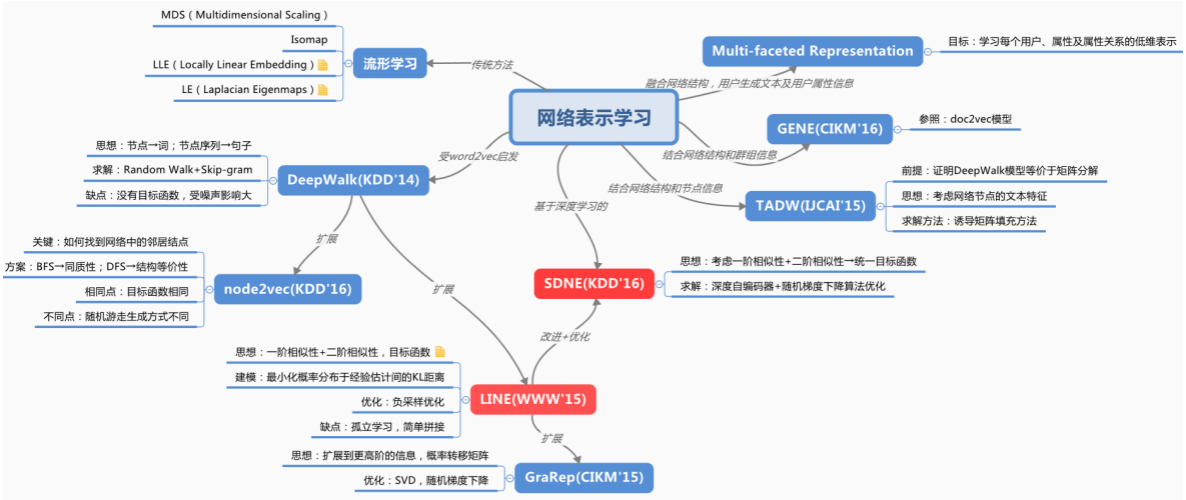
信息编码的算法

> 信息编码方法根据最小描述长度（MDL）原理，用尽可能短的码字编码尽可能多的信息，压缩原来的信息容量。可以使用在线社交网络的虚拟社区结构对网络上的信息流进行压缩编码描述。

- InfoMap
  - 使用随机游走（Random Walk）作为网络上信息传播的代理，网络上的随机游走会产生相应的数据流；使用霍夫曼编码，分配较短的码字给随机游走经常访问的节点。
  - 每次合并更新后计算出节点编码平均描述长度，平均描述长度取得最小值时为最佳划分

基于网络表示学习

网络表示学习指通过算法自动学习网络节点特征表示的一系列机器学习方法，一般将实现高维数据的低维映射



DeepWalk：基于当前流行的神经网络语言模型word2vec；在词向量学习任务中，输入是文本语料；在网络表示学习任务中，输入是网络。Deepwalk把节点作为一种人造语言的单词，通过在网络中进行随机游走，获得随机游走路径，把随机游走路径作为句子。

LINE：提高大型网络中的适用性，不仅关注节点之间的一阶相似性，即两点之间是否直接相连，并且考虑了其二阶相似性(即拥有许多共同的邻节点)

动态计算发现算法

> 从局部节点出发，依据一定规则更新局部节点状态，并逐步推演出所有节点的全局最终划分结果

派系过滤算法——发现重叠社区（用户通常可以属于多个群体或者参与多个话题）在线社交网络中，用户有聚集成团的倾向，基于团的重叠社区结构发现算法比较适合在线社交网络中虚拟社区的发现

CPM算法

（1）邻接 k 团：两个不同的 k 团如果共享 $k-1$ 个节点，则它们是邻接的。

（2）k 团链：一系列连续的邻接 k 团的集合称为 k 团链。

（3）k 团连通性：如果两个 k 团是同一个 k 团链上的一部分，则这两个 k 团是 k 团连通的。

（4）k 团社区：网络上的 k 团连通部分，即通过一系列邻接 k 团相互连通的所有的 k 团的集合。

最大团：在一个节点集合中，不是其他更大完全子图的子集的团

该算法的步骤如下。

输入：团规模的门限值 k。

（1）找出网络中所有的最大团。

（2）建立团—团重叠矩阵 $\boldsymbol{O}$。$\boldsymbol{O}$ 是一个 $n_c \times n_c$ 的对称矩阵，n_c 表示网络中最大团的数量。矩阵的元素 O_{ij} 是最大团 i 和最大团 j 之间共享的节点的数量，O_{ii} 是最大团 i 的规模。

（3）将矩阵 $\boldsymbol{O}$ 中非对角线上小于 $k-1$ 的元素和对角线上小于 k 的元素置为 0，将剩下的非零元素置为 1。

（4）对处理后的矩阵 $\boldsymbol{O}$ 进行成分分析，找出连通的部分作为最终的 k 团社区。

- CPMw：主要处理加权在线社交网络（加权图）。当k团的强度(所有边权重的几何平均值)大于设定的阈值I时，则将这个k团划分到社区中，否则舍弃这个k团。不仅考虑节点间是否有关系，还考虑这种关系的强度
- 快速派系过滤法SCP：第一阶段按顺序加边到图中并发现K团，第二阶段检查发现的k团之间的重叠程度，逐步形成最终的社区划分。

基于相似度的聚合方法：将节点映射到n维空间，借助节点距离判断

标签传播算法：用已标记节点的标签预测未标记的标签

局部拓展优化LFM：从种子节点迭代拓展找到最大健康度的子图

虚拟社区演化分析技术

1. 虚拟社区演化的**累积效应**

- fringe加入虚拟社区的可能性具有累积效应，但也存在“收益递减（diminishing returns）”的现象
- 用户在虚拟社区内朋友之间的连接强度越大，用户加入虚拟社区的概率越高
- 当社区内三角形密度很高时，虚拟社区增长速度反而变慢

2. 虚拟社区演化的结构多样性。考察用户邻居节点所属的连通分量的个数（不同的连接结构）对用户产生注册行为的影响
3. 虚拟社区演化的结构稳定性

社区演化发现方法分类

- 基于相邻时刻相似度直接比较的演化虚拟社区发现
  - 夹角余弦分别定义了节点间相似性及社区间距离
- 基于演化聚类分析的演化虚拟社区的发现
- 基于拉普拉斯动力学方法的演化虚拟社区发现
- 基于派系过滤算法的演化虚拟社区发现
- 基于节点行为趋势分析的演化虚拟社区发现

行为

> 针对社交网络中群体交互强实时、影响力动态演化等特性，研究群体行为形成机理、情感建模方法、群体交互影响度量，网络群体的产生、发展、消亡规律。
>
> 网络社会群体，是指网络个体就某个事件在某个虚拟空间聚合或集中，相互影响、作用、依赖而形成的网络个体集合

用户行为分析**

> 社交网络用户行为：用户在对自身需求、社会影响和社交网络技术进行综合评估的基础上做出的使用社交网络服务的意愿，以及由此引起的各种使用活动的总和。

1. 一般活动

用户活动角度

用户在线活动的主要类别以及各活动之间的转移规律

时间角度

花费的时间规律以及持续时间

2. 内容创建行为
3. 内容消费行为

被动视角：介绍用户的浏览行为规律

主动视角：介绍用户的信息获取行为规律

不同的社交网络结构模式队转发概率的影响

- 社交网络的亲密邻居的转发概率要比稀疏邻居的转发概率大，并且亲密邻居之间的交互频率越高，其转发概率也就越大
- 用户发帖数目越多，虽然其被转发次数也会增加，但每条消息的转发概率将会下降

社交网络情感分析**

> 情感分析(Sentiment Analysis) 又称意见挖掘（Opinion mining），是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程

根据**处理文本的粒度**不同，可以将情感分析分为篇章级、语句级和词语级等多个研究层次

- **篇章级**情感分析将整篇文档作为情感分析对象，挖掘文章对于一个事件或者产品的整体情感倾向性
- **语句级**情感分析将语句作为独立的情感分析对象，首先判断该语句是客观描述句还是主观观点句，然后针对主观观点句判断语句的情感极性。
- **词语级**情感分析主要针对词语进行情感极性判别，其主要用途在于构造情感词典。忽略上下文的影响，难于区别相同词语在不同的上下文环境中不同的情感极性。

（意见）一般采用四元组 g, s, h, t 来表示，其中 g表示情感对象或目标， s 表示情感倾向性， h 表示观点持有者，t 表示时间。

（基于实体的意见）采用五元组 e, a, s, h, t 来表示，其中 e 表示实体，a 表示实体的不同**属性**， s 表示情感倾向性，h 表示观点持有者，t 表示时间。

文本情感分析

- 基于语义规则的情感分析技术

    - 基于情感词典：根据已标注的情感词典获取评价词的情感极性
    - 基于语义规则：建立在情感词典之上，通过语义规则计算抽取出的评价词与情感词典的距离，统计所有的词的SO-PMI，最后进行情感极性的评判
        - **情感倾向点互信息算法**SO-PMI算法：依靠已有的情感词典或领域词典进行情感倾向性计算。基本思想是：选用一组褒义词（Pwords）跟一组贬义词（Nwords）作为基准词。若把一个词语word跟Pwords的点间互信息减去word跟Nwords的点间互信息会得到一个差值，就可以根据该差值判断词语word的情感倾向。SO-PMI（word1）> 0；为正面倾向。

$$SOPMI(word) = \sum_{Pword \in Pwords} PMI(word, Pword) = \sum_{Nword \in Nwords} PMI(word, Nword) \quad (4)$$

        - **PMI相关性** 使用PMI **（Pointwise Mutual Information）**衡量两个变量之间的相关性， 如果x和y无关，p(x,y)=p(x)p(y)；如果x和y越相关，p(x,y)和p(x)p(y)的比就越大。log取自信息论中对概率的量化转换（对数结果为负，一般要再乘以-1，当然取绝对值也是一样的）。

$$PMI(x;y) = \log \frac{p(x,y)}{p(x)p(y)} = \frac{logp(x|y)}{p(x)} = \frac{logp(y|x)}{p(y)} \quad (5)$$

    - 基于整体词典进行意见挖掘的方法：每个句子中存在多个特征与多个情感词，则每个特征的的情感极性为情感极性与情感和特征的距离的比值：

$$score(f) = \sum \frac{w_i * SO}{dis(w_i, f)} \quad (6)$$

    - 基于WordNet语义距离收集情感词的方法：情感词的同义词或者反义词也是情感词语
- 基于监督学习的情感分析技术
    - 朴素贝叶斯分类
    - 支持向量机
- 基于话题模型的情感分析技术
    - 最大熵LDA模型
    - HMM-LDA：话题
    - 半监督学习

要关注情感分析的**动态变化**

个体影响力分析及技术**

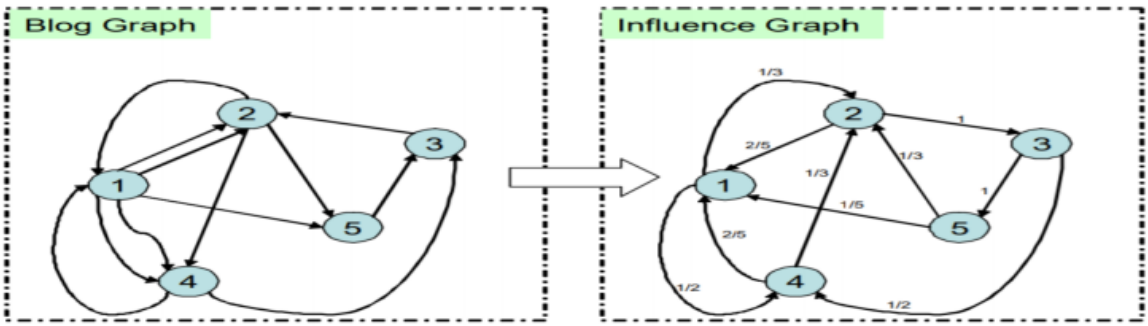
1. 影响强度计算

- 基于共同邻居数目的影响强度计算: 用社会网络中两节点的共同邻居数目来计算节点间的影响强度，共同邻居数目越多，则影响强度越高。如果节点A与B之间拥有大量的共同邻居，则认为A与B为强关系（Strong Tie），否则认为A与B为弱关系（Weak Tie）。杰卡德相似系数：

$$S(A,B) = \frac{n_A \cap n_B}{n_A \bigcup n_B} \quad (7)$$

- 基于转载频度的影响强度计算

2006年， Akshay Java与Pranam Kolari利用影响力图用于刻画上述关系，弧的方向表示影响力来源，权重代表影响力强度，用$deg^{out}(u)$表示节点的出度， $c_{u,v}$表示从节点u到节点v之间的平行边条数，其计算公式为 $$w_{u,v} = \frac{c_{u,v}}{\deg^{out}(u)}$$



- 基于行为的影响强度计算：分析行为的分布规律和因果关系

2. 影响力个体发现

- 基于网络结构的个体影响力计算
  - 点度中心度（Degree Centrality）：与该节点直接相连的节点个数
  - 利用斯皮尔曼相关系数度量影响力个体的相关性：三种行为属性网络得到结果的相关性
  - 接近中心度（Closeness）：个体与社会网络中所有其他节点的捷径距离（最短路径）之和，可用来用分析个体通过社会网络对其他个体的**间接影响力**
  - 中间中心度（Betweenness）：指的是社会网络中节点处于其它节点最短路径上的能力（“中介”作用）
  - HITS算法

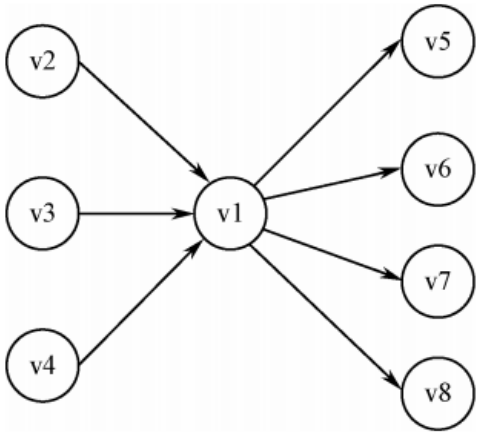
重要性。对网络图中的每个节点 v_i ，令 $a(v_i)$ 为该节点的权威度， $h(v_i)$ 为该节点的中心度，则节点权威度与中心度计算过程如下：

$$a^{(k+1)}(v_i) = \sum_{v_j \in \text{inlink}[v_i]} h^{(k)}(v_j), \quad h^{(k+1)}(v_i) = \sum_{v_j \in \text{outlink}[v_i]} a^{(k+1)}(v_j)$$

右图给出了HITS算法的计算示意图，个体 v_1 的权威值由 v_2, v_3, v_4的中心值决定；个体 v_1 的中心值由 v_5, v_6, v_7, v_8的权威值决定，即

$$a(v_1) = h(v_2) + h(v_3) + h(v_4)$$

$$h(v_1) = a(v_5) + a(v_6) + a(v_7) + a(v_8)$$



  - IP（Influence-Passivity）算法

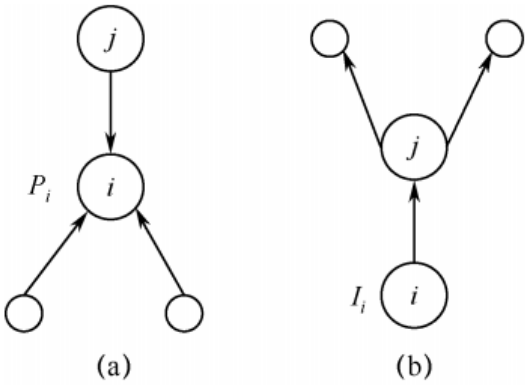
算法的IP（Influence-Passivity）算法。对每条边 $e=(i,j)\in E$，定义接受率

$$u_{ij}=\frac{w_{ij}}{\sum_{k:(k,j)\in E} w_{kj}}$$

表示用户 i 对用户 j的认可度；另外，对每条边 $e=(i,j)\in E$定义拒绝率

$$v_{ji}=\frac{1-w_{ji}}{\sum_{k:(j,k)\in E}(1-w_{jk})}$$

- 个体的影响力得分依靠：
  （1）它影响用户的消极得分；
  （2）好友相对其他用户对影响力的接收比例。
- 个体消极得分依靠：
  （1）它周围用户的影响力得分；
  （2）好友相对其他用户拒绝影响的比例。



- 基于行为的个体影响力计算
  - 四种行为形成的多关系网络
  - 多关系网络随机游走模型

群体聚集及影响机制

> 网络社会群体，是指网络个体就某个事件在某个虚拟空间聚合或集中，相互影响、作用、依赖而形成的网络个体集合

传播

> 针对社交网络中信息多源并发、内容演化等特性、研究信息内涵的表示方法、传播能量度量方法、信息传播规律与演化机理，以及信息传播影响力最大对抗策略。

社交网络信息传播规律**

信息传播影响机制

- 社交网络结构
  - 不同类型的网络传播方式不同：QQ等属于一对一传播，微博等属于一对多传播
  - 连接强度与网络密度
- 网络群体：用户的不同行为特征
- 信息：时效性、多源并发、主题多样性

基于网络结构的传播模型

- 依据信息传播所在的网络结构和邻居节点间的作用进行建模，每个节点只能处于两种状态：活跃或者非活跃状态
- 线性阈值模型、独立级联模型、扩展模型

基于群体状态的传播模型

基于信息特性的传播模型

信息热度预测方法 • 多源信息传播分析 • 信息传播模型 • 信息溯源技术

信息溯源：是指在一个网络上，给定底层网络结构属性、信息传播的模式等，在已知被观测到的传播结果的条件下，确定信息传播的最初源节点。

话题发现与演化

> 社交网络中话题数据是多源、动态、海量的

- **主题模型** • 思想：话题在潜在语义上是相似的 • 代表性的模型有：LSA模型，pLSA模型，**LDA模型**(隐含狄利克雷分布)等
  - LDA的概率公式：$p(w|d) = \sum p(z)p(w|z)p(z|d)$
  - 每个话题z被表示成一个词典 v上的多项式分布θ；每篇文档d对这T个话题有一个文档特定的多项式分布θ
- 基于向量空间的模型 • 思想：基于这样的一种假设，即话题相近的文档在内容上是相似的
- 基于词项关系图的模型： • 思想：词项之间的共现频率在某种程度上反映了词项之间语义关联

影响力最大化计算方法

> 影响力最大化是针对给定的社交网络图，寻找最关键的K个节点，使得这K个节点所产生的传播影响力最大化。例如：在微博上进行消息推送，寻找到关键大V，使得消息的传播影响力最大化