

社交网络中用户区域影响力评估算法研究

肖 宇, 许 炜, 张 晨, 何丹丹

(华中科技大学 电信系互联网技术与工程研究中心, 湖北 武汉 430074)

摘 要: 以人人网为例对在线社交网络的分析, 从区域信息传播的角度出发, 研究社交网络中, 信息传播的微观过程. 通过真实测量用户的信息传播行为, 完成用户信息传播网络的构建和测量. 发现区域信息传播网络中少量核心节点覆盖了大部分的网路传播行为. 针对这些核心节点, 文中提出了一种基于节点传播意愿和传播能力综合考察的节点传播影响力识别算法 InfluenceRank, 并通过与多种相关算法进行比对, 验证了算法的有效性.

关键词: 社交网络; 区域影响力; 节点发现; 信息传播

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2012)07-0058-06

Influence Accessment Algorithms of Regional Influential Nodes in Online Social Networks

XIAO Yu, XU Wei, ZHANG Chen, HE Dan-dan

(Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: Based on the online social network analysis, from a regional perspective, we research micro-process of information diffusion. By the measuring the real behavior of the user's sharing and reading, we construct the network of information diffusion. We found that a small group of nodes can cover the most of the communication behavior of the information diffusion. Then, we proposed a recognition algorithm InfluenceRank to find them, and verify the effectiveness.

Key words: online social network; regional influence; node discovery; information diffusion

1 引言

在线社交网络(Online Social Networks, OSN)是以计算机网络为媒介, 以社会关系为基础, 模拟或重建现实社会的人际关系网络. 它是当前互联网最热门的网络应用, 深刻地改变了互联网信息传播的规则和方式. 国外有代表性的站点有 Facebook、Linkdin 等, 国内相对应就是人人网、开心网^[1-2].

本文围绕人人网中某高校内信息传播样本数据为对象进行研究. 人人网(校内网)是中国最大和最具影响力的 OSN. 高校用户具有活跃性高、聚集性强, 信息频繁等典型特点. 本文通过有效测量和分析社交网络中用户信息传播行为数据, 力图还原信息

真实传播过程, 以实证和定量的方式研究某特定区域网络用户的信息传播影响力问题, 识别和发现真正在传播中起到核心作用的用户, 真实刻画用户在信息传播中所起到的作用.

2 相关研究

近年来国内外大量的专家和学者对对在线社交网络中节点传播影响力问题进行了研究, 总体分为两大类: 一是从实测数据出发研究节点信息传播能力问题; 二是通过预设传播模型, 研究信息传播效果预测问题. 国外代表性的有 Agarwal 利用网络中的入度、出度、发文数、评论数等多种属性综合评估博客网络中有影响力的博客, 发现有影响力的博客不

收稿日期: 2011-10-30; **修回日期:** 2011-11-25

基金项目: “十二五”科技支撑计划重点项目(2011BAK08B00)

一定是非常活跃的博客^[3]. Meeyoung 研究了 Twitter 中用户影响力问题,通过入度、回复数、提及数等不同分析指标来定量评估微博的影响力,发现在不同的应用场景下,不同的指标都有其适应性^[4]. 同样的 Jianshu Weng 将关注点放在识别话题相关联的传播影响力节点,认为节点影响力与其在具体话题传播中所起到的作用相关. 而节点的综合影响力是其在所有话题中影响力分量的集合,并提出了 twitterRank 算法,发现该算法的识别效果较好^[5]. Masahiro 在两种主流传播模型基础之上,IC 和 IR 模型仿真环境下研究用户节点影响力问题,并利用贪婪算法,模拟信息传播的过程,最终将节点影响力问题转换为如何选择合适的种子节点,使传播影响力最大问题^[6]. 国内学者针对用户传播影响力的研究,相对来说偏向于定性研究,定量研究相对较少,偏重于对传播模型、传播预测、舆情追踪和控制方面的研究,对微观节点影响力的研究较少,比较值得注意的是刘耀庭提出 UserRank 的算法实现对用户传播影响力的评估,有一定的借鉴和启发意义^[7].

3 社交网络信息传播网络测量

本文研究了一个特定区域的人群在社交网站中信息传播行为. 从微观角度,捕获每个用户的传播行为,以具体的传播数据为线索,将所有的传播行为进行关联,获得区域内完整的传播链. 多条链路叠加形成传播网络,然后采用社会网络分析法对其进行相关测量和分析.

3.1 数据获取

能否准确区分出区域外和区域内用户是刻画区域传播网络的一个关键因素. 本研究通过在校园网网络出口处,将所有校园用户访问人人网的网络行为数据进行抓取. 同时为了保护用户的隐私,所有研究数据中,凡涉及到具体用户信息,只采用抽象 ID 进行标识. 对于已经采集到数据,抽取完信息传播相关数据后,全部删除. 本数据获取方式保证了数据的准确性和完备性,对研究社交网络用户传播行为奠定了良好的基础.

人人网中,个体可以申请互为好友关系,好友之间才能并能享信息. 这种行为被作为信息传播的一种典型方式. 而且考虑到某些好友由于没有上线或者其他缘故,未能看到好友分享信息,也视为信息传递失败.

考虑到用户之间分享信息的多样性,包括文本消息、图片、视频等各种多媒体数据,并且由于各种

数据类型的差异性,本研究只抽取视频传播信息作为研究样本. 同时还需要考虑信息传播的有效性问题. 如果由用户 A 分享,用户 B 看到相关信息,就认为信息传播成功. 如果用户 C 虽然是用户 A 的好友,但是由于其未上线或其他原因未能及时看到相关信息,认为该信息传播是失败的. 这和传统的信息传播相关研究中认为用户之间存在通路,就可以传播成功的假设是截然不同. 本文完全以真实用户行为来观察信息传播行为. 也可以称之为“有效传播”测量. 某信息从区域外到区域内的传播网络模型如图 1 所示,其中黑色节点为区域外用户,白色节点为区域内用户并且成功看到该信息,灰色节点为失效节点或者活跃性底的节点,未能看到该信息.

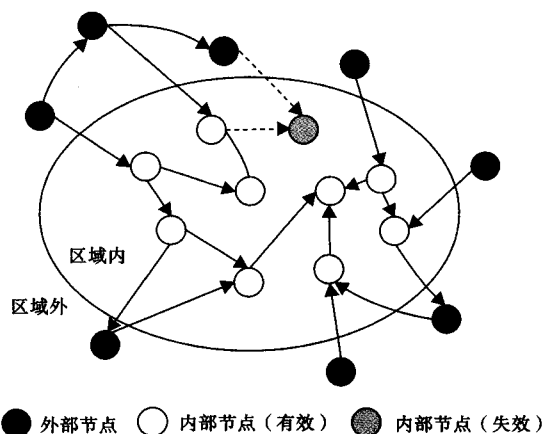


图1 社交网站中信息在特定区域中传播网络

本文对 2011 年 1 月某校园内用户在人人网中进行视频分享的数据进行统计. 如表 1 所示.

表1 获取数据统计

统计项目	值
监测时间(T)	2011 年 1 月
分享总视频数(N)	21415 个
日平均分享视频数(D)	713 个
有分享行为用户数(G)	3809 人
有播放行为用户数(W)	3850 人
区域内上线人数(K)	7659 人
涉及区域外人员(C)	44134 人
总计涉及人员(S)	51793 人

以上数据统计中,用户数均为独立用户数,例如一个用户有多次数据分享行为,人数只计一次. 一个月该区域有 7659 独立人次对人人网进行了访问,充分说明人人网在该区域的影响力非常高. 值得注

意的是,大量信息从区域外传播到区域内,涉及外部用户数量竟达到 44134 人,充分说明信息来源广泛存在.对于这些区域外人员的分析不是本文的研究重点,因此后文中不做深入分析.

3.2 实证分析

根据图论定义,信息传播网络实际可以看成有向图 G ,网络上边权 W 定义为节点之间信息传播的次数,也可以理解为该通路被使用的次数. $G=(V, W)$. V 和 W 代表了图 G 的子集.

节点集 $V=\{v_1, v_2, v_3, \dots, v_n\}$

带权重边集 $W=\{w_1, w_2, w_3, \dots, w_k\}$

节点 V 中的点代表发帖用户(vertex 或 node),边集 W 中的边(edge 或 link)代表节点之间边权. V 中元素个数和 E 中元素个数分别成为网络的阶(order)和边数(size). W_{ij} 表示用户 v_i 和用户 v_j 之间存在交流关系. W 值大小,代表用户之间的亲密程度.该网络实测数据如图 2 所示.

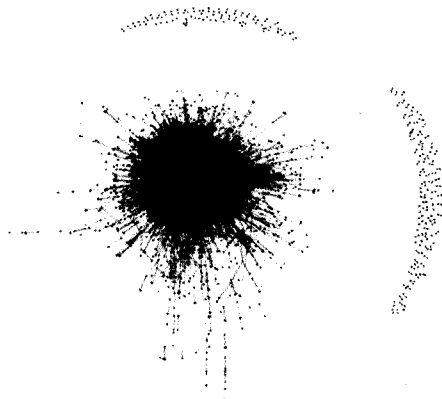


图 2 校内传播网络结构拓扑图

通过对上图观察发现,网络中存在最大联通子图和少量散列节点,说明绝大部分校园用户在信息传播网中出现聚集状态,而少量节点信息传播路径短,影响力非常有限.表 2 为社交网络分析法对网络特征进行提取.

通过对上表数据分析发现,最大联通子图人数占总监测人数的 88.5%,同时最大联通子图的聚集系数达到 0.171,最短平均路径达到 5.29,说明每个节点可以在不多于 6 个人的联系范围内,与所有节点进行联系.充分说明用户节点之间存在紧密的联系,用户的活跃度非常高.

为了进一步研究是否存在一些核心节点在信息传播中起到主导和核心作用.本研究以传播行为的次数,对所有节点进行排序,获得节点传播次数序

列,统计 Top K 用户节点的传播行为占总传播行为的覆盖率.统计结果如图 3 所示.

表 2 传播网络测量特征值

统计项目	特征值
节点数(N)	7659
边数(L)	21206
节点平均入度(I)	3.12
节点平均出度(O)	3.12
最大联通子图节点数(M)	6780
最大联通子图边数(E)	19744
最大联通子图聚集系数(C)	0.171
最大联通子图最短平均路径(P)	5.29

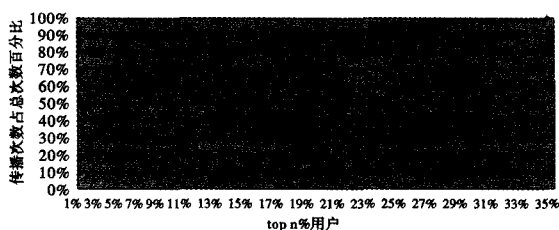


图 3 校内用户传播行为统计

该统计结果表明, top 20% 的节点覆盖了 90% 的传播行为,同时考虑到 53% 的节点只“看”不“传”,充分说明了校园内信息的传播是存在核心节点的.这些核心节点在传播网络中处于核心地位,具有较高的入度和出度.在此测量基础上,下文将提出有效算法,进一步精确识别出这些用户节点.

4 传播影响力节点识别算法

4.1 传播影响力评估模型

用户节点的传播影响力和很多因素有关,包括自身传播信息的频度、好友的数量、好友的质量(自身的传播能力)、好友的活跃程度等.以上影响因素可归纳为以下两点:

(1) 个体传播意愿:从个体视角出发,是否意愿进行信息的传播和分享.单位时间内用户传播频度可以近似度量用户传播意愿.

(2) 社会传播能力:从全局视角出发,用户的传播能力,和其所处的社交网络环境有关,是一个系统和整体的变量.用户的传播能力是其好友数量、质量、活跃度等因素的综合反映.

根据以上分析,本文创新性提出用户传播影响力 P 的度量模型:

$$P = W \times A \quad (1)$$

其中 W (Will)是个体的信息传播意愿, A (Ability)是个体的信息传播能力. 该公式表明, 只有个体的传播意愿和传播能力相匹配的时, 才能充分发挥出自身传播影响力.

4.2 算法描述

根据公式(1), 准确评估节点用户的传播影响力需要对 W 和 A 这两个关键因子进行数值度量. 其中传播意愿 W 计算公式如下:

$$W_i = F_i / T \quad (2)$$

其中 F_i 是用户 i 在 T 时间段内信息传播的次数. 可通过实测数据获取. 单位时间内传播次数越多, 可以认为用户的传播意愿越高.

用户的传播能力 A 是一个系统的全局性变量, 著名的 PageRank 算法^[8]是衡量网络中节点重要程度的经典算法. PageRank 的核心思想是每个节点的 PR 值根据反向链接的数量, 均匀的“流”向所有的关系节点. 每个节点的 PR 值为所有邻居好友对其贡献 PR 值的总和. 在传播网络中, 文中借鉴 PageRank 算法的思想, 提出 InfluenceRank (IR) 算法来评估节点用户的传播能力 A , 定义如下:

$$IR(u) = (1 - d) + d * \sum_{v \in B_u} IR(v) * C(v) \quad (3)$$

$IR(u)$ 节点 u 的 InfluenceRank 值, B_u 是指向 u 的节点的集合, T_v 是 v 指向的节点的集合, 其中 $C(v)$ 是 v 节点的传播能力分配给 u 的比例因子. d 是阻尼系数, 可设定在 $(0, 1)$ 之间, 本文中选取 0.85. 将所有节点的 InfluenceRank 初始值设为 0.1, 通过迭代到收敛为止, 可以得到所有用户的 $IR(u)$.

$C(v)$ 是传播力的分配因子, 本质上与好友的传播能力有关, 我们提出假设, 如果一个人的传播能力越大, 其好友指向他的可能性就越大. 比较合理的方式是根据每条边的指向概率方向占总的概率方向的比例来“分摊”影响力, $C(v)$ 定义如下:

$$C(v) = \frac{P_{v \rightarrow u}}{\sum_{(v, p) \in R} P_{v \rightarrow p}} \quad (4)$$

其中 $P_{v \rightarrow u}$ 是 v 指向 u 的方向概率, 取决于双方的传播能力, 而双方传播能力最重要的因素就是好友的数量, 由于好友的活跃性不同, 根据上文分析, 采用好友用户等级之和来表征有效的好友数量, 同时好友的传播能力强弱也得到了体现, 定义如下:

$$P_{v \rightarrow u} = \frac{N_u}{N_u + N_v} \quad (5)$$

$$N_u = \sum_{v \in B_u} R_v \quad (6)$$

其中 N_u 为节点 u 的“有效好友数量”, R_v 表示所有节点 u 指向的好友的用户等级.

物理上, 用户 i 的信息传播能力 A_i 取决于以下四个关键因素:

- (1) 邻节点的数量: 好友的多寡.
- (2) 邻节点的质量: 好友的信息传播能力 A_j .
- (3) 指向邻节点链接的数量: 影响好友信息传播能力的分配的因子.
- (4) 邻节点的活跃程度: 好友是否经常在线, 经常传播信息. 例如一个从不上线的好友, 对信息传播来说毫无意义. 每个节点的活跃程度可以近似采用“用户等级”来度量, “用户等级”是人人网中特有的用户积分体系, 和用户上线频次和连续上线时间有关, 是一个较为准确的评估用户活跃性的指标.

本算法充分考虑了以上因素. 通过公式(3)的反复迭代, 最后每个节点的 IR 值趋于稳定和收敛. 相关 IR 值即用户的传播能力 A_i 和传播意愿 W_i 带入到公式(1)中, 即获得每个节点的真实传播影响力度量值 P_i .

4.3 评价指标

本文采用传播实际影响人次 P 作为用户的传播能力的真实评测指标. 该数据是通过用户的微博信息在区域内被真实传播的过程中所能影响到的人次的实测结为依据, 将每个微博用户所发出的所有微博的实测影响人数进行迭加获得每个微博用户的真实的传播影响力. 值得注意的是, 如果用户 i 的某条微博在转发过程中多次传播到用户 j , 这种情况需要进行合并, 只计 1 人次. 定义如下:

$$P_i = \sum_{j=\{1, \dots, k\}} M_{C_j} \quad (7)$$

其中 M_{C_j} 是用户 i 发布的第 j 条微博传播覆盖人群. 如果一条微博经多次转发, 多次转发过程中所传播到的所有人群都计为该微博的影响范围.

5 结果与分析

为了方便比较, 本研究采用三种传统方法得到的用户影响力排名作为相关对比衡量指标.

(1) 入度 (Indegree): 指向本节点的邻居数之和. 一个人的好友数量多少一定程度上决定了信息在该用户节点可能的扩散面.

(2) 用户被访问量 (Visit Count): 网站提供的用户的个人主页被其他用户的访问量. 访问量越大, 说明这个人发布的信息越受欢迎. 是人人网官方

的用户影响力指标。

(3) 标准 PageRank(Global PageRank): 不考虑用户传播意愿和活跃性, 对全样本数据直接执行标准 PageRank 算法。

为了进一步对比和分析算法在真实影响人次这一评价指标下对用户影响力的识别准确程度, 通过相关是算法获得用户影响力的排名, 然后通过对排名的 Top k% 用户影响人次覆盖率进行对比, 并与真实用户传播影响人次降次序列进行对比, 考察各算法结果用户影响力倒序序列之间的相对关系。结果如图 4 所示。

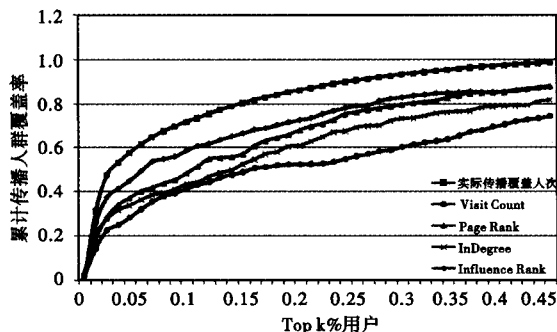


图4 四种算法 Top k% 用户传播影响人次覆盖率统计

通过对图 4 的分析可以看出, 本文提出的 InfluenceRank 算法在累计传播人次这一指标下表现最好, 也就意味着该算法所获取的用户传播影响力序列最接近用户的真实人群传播影响力。而传统的 PageRank 算法的表现略差于 InfluenceRank 算法, 这说明加入了用户意愿和活跃性因子后确实能够有效提高算法的准确性。同时发现入度 (InDegree) 的作为一种评价指标的效果并不太差, 这确实是一个有意义的发现, 这说明社交网络是一个“熟人社会”, 由于这一网络中用户之间彼此熟悉, 他们之间会选择彼此熟悉的信息或者互动方式增强彼此的感情。因此这种强连接的小圈子之间的互动非常密切。因此一个社会交往面大的用户, 自然接触到新鲜信息的机会就多, 因此一个人的邻居数量往往决定了这个人的社会交往关系, 好友数量的越多和用户的活跃度也很有关系, 需要不断的主动扩大好友圈子和增加与好友的互动才能不断增加邻居的数量。因此通过入度来判断一个人的传播影响力也有一定合理性。整体而言表现最差是用户主页的访问量 (Visit Count), 其原因也不难理解, 因为一个人的访问量大多可能是小圈子里面的用户之间的密切沟通造成了, 但是这种小圈子里的密切沟通并不能产生太多的对外的信息的传播和扩散。因此一个用户的访

问量并不能代表这个用户的信息传播的能力。

将本文提出的信息传播影响力评估算法结果与用户实际传播影响人次 P_{real} 进行数值对比, 通过定量的分析两个序列之间的匹配度。通过用户实际传播影响人次 P_{real} 定义用户的期望值 E_P :

$$E_P = \sqrt{P_{\text{real}}} \quad (8)$$

定义 ρ 为用户 P 值与期望值 E_P 的比值, 即:

$$\rho = \frac{P}{E_P} \quad (9)$$

对 ρ 值的分布进行统计, 观察用户的传播影响力 P 值与期望值之间的差异性分布。如图 5 所示。

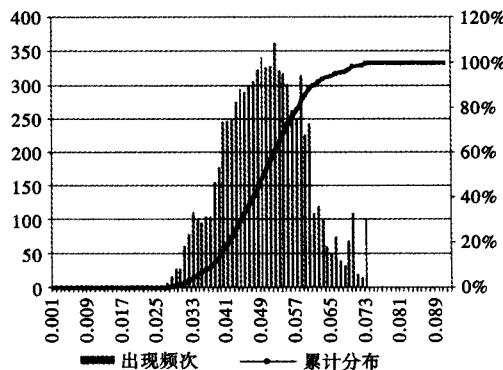


图5 用户传播影响力 P 与期望值对比

从图 5 中可以看出, ρ 值以 0.5 为中心向两边递减, 绝大部分都分布在 0.038~0.058 之间, 这部分节点占到总节点数的 80%, 这充分说明本文提出的用户传播影响力识别算法与用户真实访问量非常接近, 从而也证明了算法的有效性。

将用户按照访问量进行排序, 得到用户访问量序列, 在图 6 中按照降序从左到右依次列出其对应的 P 值。观察发现本研究获取的用户传播影响力 P 与用户访问量 VC 的吻合度很高, 用户的信息传播影响力 P 值随着访问量的减少而相应的减少。同时, 个别用户的值发生的偏差较大, 特别是访问量较大的节点。从图中判断, 这些用户所占的比例较小, 因此可以得出定性的结论: 本文中提出的用户影响力识别算法还是比较准确的。

6 结束语

本文以人人网为例, 深入研究了社交网络中信息在用户间分享和传播过程。通过对某高校的实测数据发现, 少量的传播节点覆盖了主要的传播行为。本文提出传播影响力识别算法。综合考虑个体传播意愿和社会传播能力这两个因素, 定量的对每个个体的传播影响力进行度量, 并通过与用户访问量进

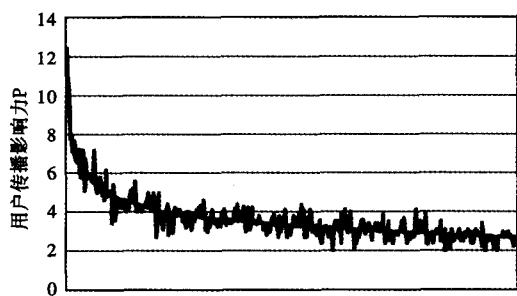


图6 准确度对比图

行对比,有效证明了算法的准确性.后续将进一步研究区域内外信息传播的方式和特点以及核心节点在信息传播中所起到的作用.

参考文献:

- [1] 周荣庭,方冰. Web2.0 网站新闻传播的特性比较与趋势 [J]. 网络传播,2009(11):60-61.
- [2] 薛可,余阳明. 人际传播学[M]. 上海: 同济大学出版社,2007.
- [3] Nitin Agarwal, Huan Liu. Identifying the influential bloggers in a community [C]// Proceedings of the international conference on Web search and web data mining. USA: Seattle, 2008:207-217.
- [4] Cha M, Haddadi H. Benevenuto, measuring user influence in twitter: the million follower fallacy [C]//

Proc. of International AAAI Conference on Weblogs and Social Media (ICWSM). Washington, 2010:125-134.

- [5] Jianshu Weng, Ee-Peng Lim, Jing Jiang, et al. TwitterRank: finding topic-sensitive influential twitterers [C]// Proceedings of the third ACM international conference on Web search and data mining. USA: New York, 2010:69-86.
- [6] Kimura M, Saito K, Nakano R. Extracting influential nodes for information diffusion on a social network [C]// Proceedings of the 22nd AAAI Conference on Artificial Intelligence. Canada Vancouver, 2007. AAAI, 1371-1376.
- [7] 刘耀庭. 社交网络结构研究 [D]. 杭州:浙江大学, 2008.
- [8] Page L, Brin S, Motwani R. The pagerank citation ranking: Bringing order to the web [R]. Stanford University, Technical report, 1998.

作者简介:

肖宇 男,(1979—),博士研究生.研究方向为复杂网络.
 许炜 男,(1977—),副教授.研究方向为网络行为分析.
 张晨 男,(1988—),硕士研究生.研究方向为社会网络、数据挖掘.
 何丹丹 女,(1987—),硕士研究生.研究方向为社会网络.

(上接第57页)

- [6] 胡艳,胡瑞林. 基于小波模极大值和形态学的图像边缘检测算法[J]. 计算机应用与软件,2010,27(9):75-77.
- [7] 卢萌,赵合计. 一种改进的小波自适应边缘检测算法[J]. 计算机应用研究,2011,28(1):285-287.
- [8] 袁修贵,龚正,孟正中. 基于边缘方向性的小波边缘检测算法[J]. 计算机工程与科学,2010,32(3):49-51.

作者简介:

韩慧妍 女,(1980—),博士研究生.研究方向为信号与信息处理.
 韩雯 女,(1964—),教授,博士生导师.研究方向为信号与信息处理.