Citation Graph
引文图分析

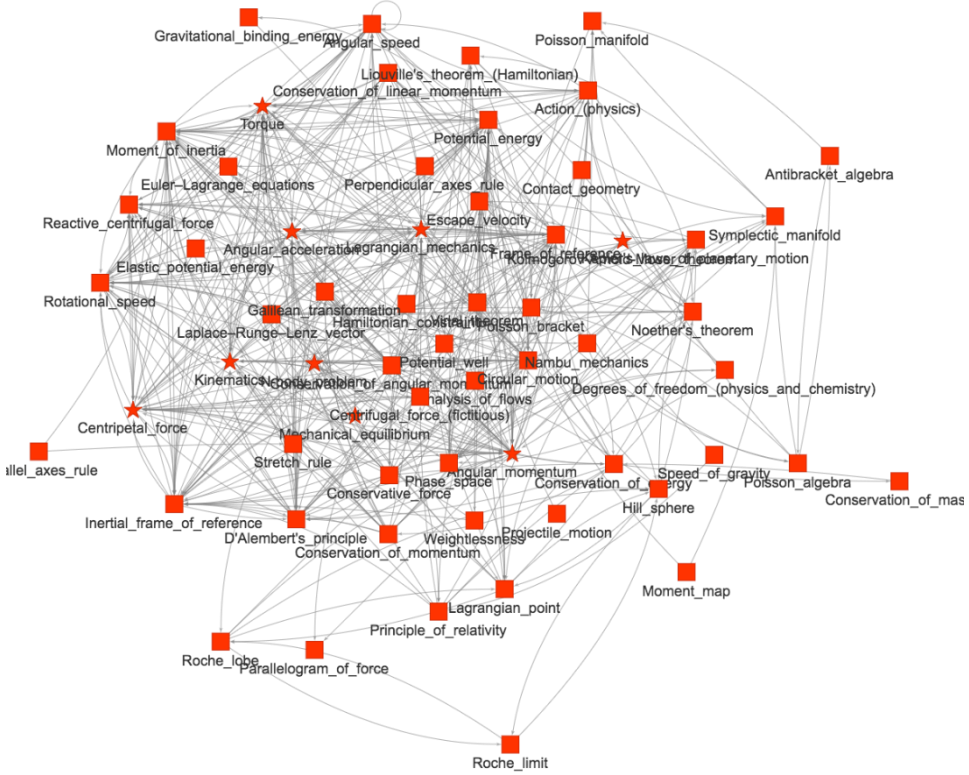
任泽华 3121154002

2022年5月6日

CONTENTS

一、引文图简介

在信息科学和文献计量学中，引文图（或引文网络）是描述文章集合中的引文的**有向图**。图中的每个顶点代表集合中的一篇文章，并且每条边都从**一篇文章指向它引用的另一篇文章**（抑或相反，取决于定义）。属于社会网络范畴又有所不同：引用**关系确定**，不可随意添加或删除。引用关系在时间上具有**单向性**，只能是后期的文献引用前期的文献。



代表着某领域的研究现状及未来发展趋势，广泛应用于计算影响因子、研究学科相关、论文推荐与知识发现等。

二、引文图性质

Yuan An等[1]在2004年对引文图性质进行了研究，他们在ResearchIndex爬取了三个不同领域（神经网络、自动机和软件工程）的论文，分别探究了这三个网络及其联合引文图的性质。

1、度分布

引文图度数遵循**幂律分布**，即对某些 $\gamma > 1$，度数为 i 的论文比例与 $1/i^{\gamma}$ 成正比。 γ 值在三个网络及其联合图中都几乎相等，入度为1.71 出度为2.32。此结论**可推广**到所有引文图。

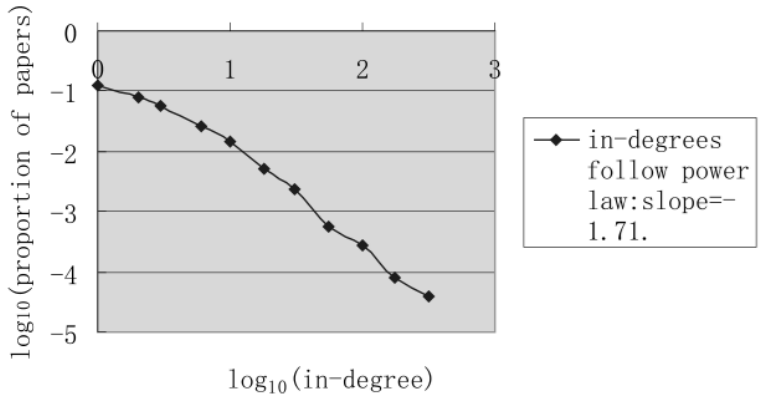


Fig. 3. The in-degree distribution in the union citation graph in computer science literature subscribes to the power law with exponent = 1.71.

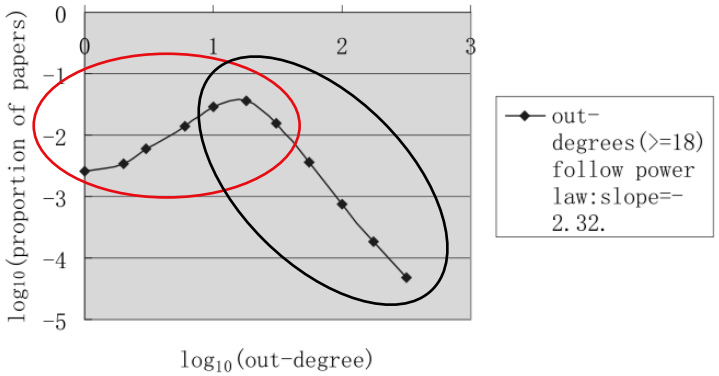


Fig. 4. The out-degree distribution in the union citation graph.

[1] An, Yuan; Janssen, Jeannette; Milios, Evangelos E. (2004), "Characterizing and Mining the Citation Graph of the Computer Science Literature", Knowledge and Information Systems, 6 (6): 664–678, doi:10.1007/s10115-003-0128-3, S2CID 348227.

二、引文图性质

2、图直径

引文图最大弱连通分量直径满足“**小世界性**”，其值约为18。表示图中大部分节点都有联系且大部分结点可以从任一其他点经少数几步就可到达。此结论可推广到所有引文图。

对于有向图（强连通分量）情况完全不同，任何一对节点之间存在有向路径的概率仅为2%，最大强连通分量直径也在30左右。说明论文之间**极少存在互引**的情况。

由于存在离散点，故将网络中最大弱联通分量的直径作为全图直径（下条性质保证90%以上论文都在这个弱联通分支内）。

实际的社会、生态、等网络都是小世界网络，在这样的系统里，信息传递速度快。

Table 1. The diameters of citation graphs built from different topics as well as union citation graph. Topic: N.N.: Neural Networks, S.E.: Software Engineering.

	graph size	directed diameter	undirected diameter
citation graph–N.N.	23,371	24	18
citation graph–Automata	28,168	33	19
citation graph–S.E.	19,018	22	16
union citation graph	57,239	37	19
average		29	18

二、引文图性质

3、连通性

形成最大**弱连通分量 (WCC)包含了90% 的节点**，这个连通分量中**68.5% 的节点没有传入边**（未被他人引用）。在巨型 WCC 内部，大约 **58% 的节点形成了一个大型双连通分量**（BCC），其余40%节点几乎都属于平凡双连通分量（一个节点）。

与我们预期相反，引文图中**存在三个较大的强连通分量**（SCC）。

双连通：两点间边数 > 1。

双连通分量（BCC）：一个极大子图，其中每对顶点都是双连通的。

虽然存在少量**大度数节点**（权威论文），但是**去掉它们并不影响整体的连通性**。引文图非常有弹性，不依赖中心和权威的存在。（具有小世界性的无标度网络）

	graph size	size of largest WCC	percentage of largest WCC	size of second largest WCC
citation graph–N.N.	23,371	18,603	79.6%	21
citation graph–Automata	28,168	25,922	92%	20
citation graph–S.E.	19,018	16,723	87.9%	12
union citation graph	57,239	50,228	87.8%	21

	graph size	size of largest SCC	size of second largest SCC	size of third largest SCC
citation graph–N.N.	18,603	144	14	10
citation graph–Automata	25,922	192	29	24
citation graph–S.E.	16,723	17	11	8
union citation graph	50,228	239	155	60

三、引文图划分

文中使用最小割算法对图进行划分，期望按照研究领域对论文进行切割。

工作表明，引文图作为一个整体的连通性使得**不可能用简单的方法**（例如最小割）来提取这样的社区。需要**更复杂的方法**以及用图论术语对社区进行**精确定义**。

1、全局最小割

在全局递归运行最小割算法，结果大部分割只将一个节点与图的其余部分分开。

```
1. procedure Explore_Min_Cut (H = (V, E))
2.   while |H| > 0
3.     compute min-cut C of H;
4.     calculate edge weight over crossing edge set F;
5.     let H1 = (C, E1) be graph induced by C;
6.     let H2 = (V − C, E2) be graph induced by V − C;
7.     Explore_Min_Cut(H1);
8.     Explore_Min_Cut(H2);
9.   end while;
```

2、特定节点最小割

计算两个最不相关论文之间最小割，结果获得了高度不平衡分割。（1:1000）

图 $H = (V, E)$ 的边割 C 是一组边，当它们被移除时图就会断开。边割大小是它的边数。给定边权重函数 $w: E \to R$，最小割是总权重最小的割。

四、后续研究进展

1、引文图相似性

Wangzhong Lu等[2] 2007年提出。利用引文图中节点的本地邻域。描述了两个节点之间基于链接的相似性估计的两种变体。用于在引文图上寻找相似论文。

2、引文图采样与可视化

Lei Shi等[3] 2015年提出。总结大型引文网络中的影响因子，使用基于流和本地化的上下文方法。将其定义为一个优化问题，使用矩阵分解的方法得到了对用户呈现的最优采样。

3、引文图连接预测

Hanwen Liu等[4] 2019年提出。是一种结合时间、关键词和作者信息并优化现有论文引用网络的链接预测方法。返回一组满足用户对关键词需求的论文。。

[2] Lu, Wangzhong; Janssen, J.; Milios, E.; Japkowicz, N.; Zhang, Yongzheng (2007), "Node similarity in the citation graph", Knowledge and Information Systems, 11 (1): 105–129, doi:10.1007/s10115-006-0023-9, S2CID 26234247.
[3]L. Shi, H. Tong, J. Tang and C. Lin, "VEGAS: Visual influEnce GrAph Summarization on Citation Networks," in IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 12, pp. 3417-3431, 1 Dec. 2015, doi: 10.1109/TKDE.2015.2453957.
[4] Liu, H., Kou, H., Yan, C. et al. Link prediction in paper citation network to construct paper correlation graph. J Wireless Com Network 2019, 233 (2019).

谢谢大家！

2022年5月6日