

网络中心度计算方法研究综述

吴思竹^{1,2}, 张智雄¹

¹ (中国科学院国家科学图书馆, 北京 100190)

² (中国科学院研究生院, 北京 100049)

摘要: 以网络中心度计算方法为研究对象, 首先, 归纳出网络中心度的概括性定义。而后, 从中心度算法研究、中心度算法分类研究、中心度算法应用研究等多个方面对该领域的国外主要研究发展情况进行深入归纳、总结及分析。最后, 对中心度算法的研究发展方向进行展望。

关键词: 中心度, 网络分析, 中介中心度

中图分类号: G350

Review on Research of Network Centrality Algorithm

Wu Sizhu^{1,2}, Zhang Zhixiong¹

¹(National Science Library, Chinese Academy of Sciences, Beijing 100190)

²(Graduate University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100049)

Abstract: This paper concerns on centrality algorithm, and gives a general definition of centrality first. Then, it makes a review on foreign researches of this area from many aspects: research on centrality algorithm、research on category of centrality algorithm、research on application of centrality. Finally, it discusses the development of the research on centrality algorithm.

Keyword: centrality, network analysis, Betweenness Centrality

中心度是应用于网络分析的一个重要度量指标, 用于测量网络中“元素”的重要性, 这里的“元素”是一种泛指, 包括网络中的节点、边、社团及整个网络。本文对中心度的研究主要集中在节点中心度上。

本文的篇章结构组织为首先解释并明确中心度的概念, 回顾当前的节点中心度的主要算法, 对现有算法研究、分类研究及应用研究进行总结和分析, 最后对中心度算法研究情况予以总结及展望。

1 中心度的概念

中心度研究能够识别网络中的重要节点, 节点的重要程度由网络的拓扑属性、结构特点及节点在网络中的具体位置决定。Mike Gotta 指出“中心度的概念, 简单来说识别网络中具有高度连接的活动者”。这个定义有些片面, 而维基百科中也没有给出明确的定义, 给出的是“在图论和网络分析中, 对一个节点的多种中心度测量, 这些测量主要是决定图中一个节点的相对重要性。”的一种解释。

之前的研究中并没有给出关于中心度的严格定义, 本文认为它是关于节点重要性的度量指标。这种重要性, 根据不同网络和结构特点及关系表现为节点的影响力、权威性(重要思想、知识或判断决策的源头)、流行度、控制力(如传输、流量的控制能力), 便利性(位置上的优势、易于访问)或某种特殊意义, 也可以表现为节点的脆弱性和易受攻击性。不同的中心度算法含义不同, 此处给出的是一个总结性质的定义。

2 中心度算法分类及研究进展

2.1 中心度算法分类研究

已有不少学者对中心度算法的分类进行了研究。Noah^[1]在描述中心度测量的理论基础时将中心度划分为三种类型，包括总体效果中心度、接近效果中心度和调解中心度。Dirk Koschützki 等^[2]将中心度分为四种类型，包括基于可达性、基于流量、基于活力值和基于反馈的四类中心度。Zhang^[3]将中心度算法划分为五类，包括基于点度、基于最短路径、基于流、基于随机行走和基于反馈的中心度。从其他角度考虑中心度的划分，还可以分为基于全局的中心度算法、基于局部的中心度算法，或根据具体算法分为绝对中心度和相对中心度算法，根据网络动态性，分为动态中心度算法和静态中心度算法。另外，根据网络属性、权重和方向又可以分为基于有向网络的算法和基于无向图网络的算法、基于加权网络和基于无权网络的算法及几者组合的分类方式。

2.2 主要的中心度算法研究

根据 2.1 的分类，本文基于生成原理将中心度算法分类，划分为基于连接的、基于最短路径的、基于流的、基于随机行走的和基于反馈的中心度算法五种。

(1) 基于连接的中心度算法

度中心度算法 (Degree) 用于测度网络中一个节点与其他节点直接连接数的总和。在有向网络中，度又分为入度和出度。Hildrun^[4]基于合著网络比较有权和无权的度中心度算法，提出查找一种函数，使元素权重均等分布数等于元素的个数。得到偏离相等分布越大，函数值越小的包含权重连接关系的复杂度中心度算法。Benjamin^[5]聚焦于在网络规模、边密度、边强度和方向性等变量变化条件下的度中心度研究。由于度中心度算法反映的是静态的局部联系，在反映重要节点时具有局限性，研究中通常与其它网络属性测度指标结合应用。

(2) 基于最短路径的中心度算法

基于最短路径的中心度算法包括接近中心度 (Closeness)、中介中心度 (Betweenness)、Harmonic 中心度、Eccentricity 和 Centroid 等。这一类中心度基于网络中节点间的最短距离。Yannick^[6]对 Harmonic 中心度和接近中心度进行比较，指出 Harmonic 中心度可以作为接近中心度的替代算法，并将其扩展到无向图当中。Centroid 也可看作是接近中心度的变种，它测量的是一个节点比其他节点在位置上的优势程度。中介中心度是基于最短路径的经典算法，能够用于揭示网络中具有连接桥作用的节点，从而发现网络连接中的关键点或脆弱点。由于该算法的重要性，诸多学者关注该算法的研究，对其进行改进与优化。其中，Ulrik Brandes^[7]较系统地总结了中介中心度的变种算法及其适用范围。虽然基于最短路径的算法是中心度测量的重要方法之一，但是基于最短路径的原理也成为这类算法的局限，在实际网络中并不是所有重要节点都通过最短路径。

(3) 基于流的中心度算法

基于流的中心度算法引入电网电流流动理论，将网络关系看作是包含电压、电流的电网，基于在网络中电流的流动进行建模。主要包括基于流的中介中心度，基于流的接近中心度及信息中心度。这类中心度算法主要被应用于社会网络中用于探测社团结构，如 Franco^[8]提出通过计算图中相关的局部流中介中心度，利用边的权重建模进行社团的抽取及聚类。

(4) 基于随机行走的中心度算法

基于随机行走的中心度算法包括随机行走中介中心度、随机行走接近中心度和马尔可夫中心度算法。这类算法主要基于随机行走原理，计算在起始节点和目标节点间对中间节点的随机访问次数。基于随机行走中介中心度算法是由 Newman 提出的，用于解决最短路径的局限性问题。S.Lee 等^[9]提出基于偏好随机行走的中心度算法，并将其应用于复杂网络，在度相当大的条件下该算法能够对多种中心度算法进行统一。

(5) 基于反馈的中心度算法

基于反馈的中心度算法，包括 Katz's status、Hubbell、Eigenvector 及著名的 PageRank、HITS 等算法。

研究广泛集中在 PageRank 等网络排序算法上。由于其缺乏对网页主题内容等其他因素的考虑，不少学者对该算法进行了改进和扩展。算法改进注重综合多方面影响因素，包括网页链接、网页内容、用户点击及浏览行为等。目前较突出的研究如主题敏感的 PageRank、

个性化加权 PageRank 算法等能够对已知查询识别出更多相关度更高的网页。

2.3 中心度近似算法研究

由于精确算法需要多次迭代，在时间和空间消耗上较大，特别是在真实大型网络中，为实际计算带来了巨大挑战。因此对近似算法的研究能够在一定程度上解决这方面的问题，近似算法研究关注在迭代次数、最短路径计算、抽样等方面的改进，提高计算效率和性能。

David 等^[10]提出一种基于适应性抽样技术的中介中心度近似算法，大大地降低了对高中心度节点的单源最短路径计算的消耗。Kazuya^[11]等根据接近中心度的精确值和近似值计算提出新的算法。算法主要用于选取接近中心度值最高的节点，而非全部节点从而降低了时间消耗。近似算法能够在一定程度上提高计算效率并保证得到的结果在可接受的误差之内。

2.4 并行中心度算法研究

由于网络、特别是真实网络，更加复杂并且规模庞大，因此对于现有算法的性能和效率提出了挑战。当前的一些算法研究关注于在更精细粒度上的并行方法，对算法进行切分和加和，分布到多台机器上运行，提高运行效率的同时提高运行效果。

Christian 等^[12]提出一种新的并行模式的 PageRank 算法。该算法通过引入网络二维视图，保存主机 ID 作为区分，而后将 PageRank 划分为不相连的部分，应用 Gauß-Seidel 迭代算法进行快速的线性系统求解。该算法与其他并行 PageRank 算法相比，在每次迭代时间上有很大改进。Tan 等^[13]提出一种新的适用于 CREW PRAM 的并行中介中心度算法，应用于大规模网络分析，通过适当的数据处理器映射、新的数边策略和三元数据矩阵结构，通过记录最短路径减少访问共享存储器冲突问题。

3 中心度计算方法应用研究

虽然中心度计算方法主要被应用于复杂网络研究，但是由于它是基于网络的，在其他领域研究中也受到广泛关注。本文对其应用领域进行归纳，可主要体现在复杂网络分析、期刊及论文评价、大众标注标签分析及推荐、网页排名、关键词抽取及文本摘要、语义结构探测及语义消歧、科学前沿及创新探测、重要作者识别等方面。由于网页排名主要涉及 PageRank 算法前文已经总结，此处着重于归纳中心度算法在其他领域的研究情况。

3.1 复杂网络分析

中心度算法主要被应用于复杂网络分析，如社会网络、生物蛋白质网络、电力网络等。Nicola^[14]等研究了复杂网络中的中心度算法，回顾并比较了基于图矩阵拓扑属性的中心度算法，包括 PageRank、Eigenvector 和 HITS，发现一些中心度是相互关联的。Francesco 等^[15]将中心度的拓扑概念应用于解释复杂网络连接的可靠性和安全性，将度中心度、中介中心度、接近中心度、信息中心度拓展为可靠性度中心度、可靠性接近中心度等算法，将其应用到电力传输系统网络中，用于评估网络路径连接元素的重要性。

3.2 期刊及论文评价

随着研究的深入，不少学者将中心度算法应用于期刊引用网络，为科学质量评价提供新的方法。其中 Jevin West 提出一种新的学术期刊评价指标 Eigenfactor，它基于 Eigenvector，利用文献对期刊的引用率对期刊进行排序，其与传统期刊评价指标的区别在于其考虑了整个引用网络的结构，考虑了间接联系及效果。Chena 等^[16]应用 PageRank 算法对 1982 年-2003 年物理评论系列期刊中的所有出版物进行了重要性评估。发现每个出版物的 PageRank 值和引用数相关。利用 PageRank 算法识别了期刊中物理学家熟悉的一些杰出的具有影响力的论文。

3.3 大众标注标签分析及推荐

大众标注是自 Web2.0 以来倍受学术界关注的一个领域，中心度算法在大众标注中的应用研究主要体现在对标签的推荐和对标注用户的社会网络分析等方面。主要通过用户、资源、标签所构成的三元组关系构建不同的关系图或关系网络，并基于此进行中心度的测量和评估，进行对标签推荐或标签网络规律特点的分析。

Andreas H 根据大众分类法基于用户、资源、标签三元组关系的特点，提出 FolkRank 算法，将其主要用于特殊主题的标签、资源或用户的推荐。Robert 等^[17]对 FolkRank、基于改进

的PageRank及基于标签流行度的推荐方法进行了测试对比,发现前两种方法均比非个性化的推荐方法更有效,特别是FolkRank方法在探测超图结构、解决冷启动问题上都有优势。Iván^[18]基于共现图使用节点中心度算法进行社会标签推荐。通过关键词集抽取,检索相关书签,构建全局共现子图,结合TF-IDF算法计算相关标签的中心度,将具有最高中心度的标签作为推荐结果。

3.4 关键词抽取及文本摘要

关键词抽取和文本摘要也是中心度算法应用的重要领域,其主要被应用于测量文本中的词或句子的中心度,确定关键词或中心句,揭示文本的主题内容。特别是在关键词抽取研究中,中心度算法能够规避对低频但重要的词的忽略问题。

Kino^[19]使用 Wikify 系统从维基百科的文章中抽取关键词,利用基于随机冲浪的中心度算法进行主题识别。Zhang^[20]等提出利用 Hub-Authority 框架进行文本摘要的方法,结合线索短语、句子长度、首句等线索,将子主题的属性融入基于图的句子排序算法中来探测多文本主题。

3.5 语义结构探测及语义消歧

时至今日,对于语义的探测和挖掘成为研究者关注的热点,中心度在语义结构探测和语义消歧方面也有主要应用。Jason 等^[21]提出通过语义结构挖掘算法构建一致性本体和计算局部和全局语义中心度的思想,用于增强子团体的发现和资源的共享。

在语义消歧方面,Ravi^[22]应用词义相似度和中心度计算进行基于图的词义消歧,并使用测试数据集对入度中心度、接近中心度、中介中心度等算法分别结合基于 WordNet 的不同相似度计算进行了实验,取得了较好的效果。

3.6 科学前沿及创新探测

中心度被应用于引文网络、共被引网络测度具有重要影响力的论文,发现科学前沿和研究基础。在引文网络中,Chen^[23]提出在知识域中通过对标志性节点、中心节点和拐点的发现,查找重要文献的潜在特征。标志性节点是网络中具有特殊属性的节点,如高被引文献提供重要的标志,这种节点具有较大的半径。中心节点其节点度相对较高,在引文网络中体现为共被引文献,其具有重要智力贡献。拐点是连接不同网络的枢纽,是两个子网络共享的节点或子网络交互路径上的节点。Chen 在其开发的引文分析工具 Citespace 中利用中介中心度对引用网络中科学发展中的拐点进行识别。Shibata^[24]等通过被引次数和聚集中心度、中介中心度、接近中心度间的关系,探测作为基础创新萌芽的论文及预测论文未来被引的能力。

3.7 重要作者识别

PageRank、中介中心度等算法也被用于论文合著网络或作者引用网络中,根据网络属性特点,作者在网络中的贡献,如合作或被引的链接数量,合作者之间、引用者与被引用者之间的相互关系测量作者在网络中的影响力,对作者进行排序,识别网络中重要的、具有高度影响力及支持度的作者。Yan^[25]等将中心度算法作为一种影响力分析指标,通过来自 16 个图书情报期刊的 1988 年到 2007 年的数据生成合著者网络,利用接近中心度、中介中心度、度中心度和 PageRank 四种算法对作者排序,分析了各种算法在作者影响力评估上的局限性。

4、结语

本文较系统地对现有网络节点中心度计算方法研究及应用情况进行了总结和归纳,发现通过这些指标的计算能对不同的网络结构和特性进行揭示。就现有研究而言,由于中心度算法只是反映网络属性的一个方面,并且每种中心度算法都具有一定的适用范围和局限性,因此当前很多研究不仅考虑单一算法的运用更多的是注重多种算法的结合,并且特别关注于对网络的动态性变化和规模不断增长的情况的考虑,提出适于增量计算和具有高性能的改进算法。在应用研究中,除了在复杂网络分析中的应用,中心度算法在主题识别、语义探测等多个领域也都发挥了巨大潜力,并且有待于更深入的研究和拓展。由于不同网络具有不同特点,因此,基于特定网络特点的中心度算法研究将在未来得到更多的重视和关注。另外,在中心度算法的评估上研究者注意针对多种算法进行测评、比较及对算法不足的分析,这也是保证算法有效性的必要手段。

参考文献

- [1] Noah E Friedkin.Theoretical Foundations for Centrality Measures.[2009-10-12].<http://www.soc.ucsb.edu/faculty/friedkin/Reprints/AJSCentrality.pdf>
- [2] Björn H. Junker.Analysis of Biological Networks[M].Wiley-Interscience,2008:17-20.
- [3] AiDong Z.Protein Interaction Networks:Computational Analysis[M].English: Cambridge University Press,2009:35.
- [4] Hildrun K, Theo K. Application of a New Centrality Measure for Social Network Analysis to Bibliometric and Webometric Data[C].ICDIM,2006:199-204.
- [5] Elbirt, B. S. The Nature of Networks: A Structural Census of Degree Centrality across Multiple Network Sizes and Edge Densities[C].SUNBELT,2008:1-82.
- [6] Yannick R.Closeness Centrality Extended To Unconnected Graphs:The Harmonic Centrality Index[C].ASNA,2009:1-14.
- [7] Ulrik Brandes.On Variants of Shortest-Path Betweenness Centrality and their Generic Computation[J].Social Networks,2008,30(2):136-145.
- [8] Franco S, Savitha S.Local Flow Betweenness Centrality for Clustering Community Graphs[C].WINE,2005:531-544.
- [9] Lee S, Yook S.H et al.Centrality Measure of Complex Networks Using Biased Random Walks[J].EPJB,2009,68(2):277-281.
- [10] David A. B, Shiva K, et al.Approximating Betweenness Centrality[C].WAW,2007:125.
- [11] Kazuya O, Wei C et al. Ranking of Closeness Centrality for Large-Scale Social Networks[C].FAW,2008:186-195.
- [12] Christian K,Paul Alexandru C et al., Efficient Parallel Computation of PageRank[C].ECIR, 2006:241-252.
- [13] Guangming T, Dengbiao T et al.A Parallel Algorithm for Computing Betweenness Centrality[C].ICPP,2009:340-347.
- [14] Nicola P,Santo F.Spectral Centrality Measure in Complex Networks[J/OL].Physical Review E, 78:1-11.
- [15] Francesco C,Enrico Z et al.Using Centrality Measures to Rank the Importance of the Components of a Complex Network Infrastructure[M].Critical Information Infrastructure Security.Springer,2008:155-167.
- [16] Chena P,Xie H et al.Finding Scientific Gems with Google's PageRank Algorithm[J].Journal of Informetrics,2007 (1) :8-15.
- [17] Robert J, Leandro Balby M et al.Tag Recommendations in Folksonomies[C].PKDD, 2007:506-514.
- [18] Iván C,David V.Measuring Vertex Centrality in Co-occurrence Graphs for Online Social Tag Recommendation[C].ECML PKDD,2009:497:17-33.
- [19] Kino C,Rada M.Topic Identification Using Wikipedia Graph Centrality[2009-10-12].www.aclweb.org/anthology/N/N09/N09-2030.pdf.
- [20] Junlin Z,Le S et al.A Cue-based Hub-Authority Approach for Multi-Document Text Summarization[C].IEEE NLP-KE,2005:642-645.
- [21] Jason J, Jérôme E.Measuring Semantic Centrality based on Building Consensual Ontology on Social Networks[C].ESWC,2006.1-60.
- [22] Sinha, R. Mihalcea, R.Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity Semantic Computing[C],ICSC,2007:363-369.
- [23] Chaomei C.Searching for Intellectual Turning Points: Progressive Knowledge Domain Visualization[J].PNAS,2004,101:5303-5310.
- [24] Shibata N,Kajikawa Y et al. Early Detection of Innovations from Citation

Networks[C].IEEM2009.

[25] Erjia Y, Ying D. Applying Centrality Measures to Impact Analysis: A Coauthorship Network analysis[J]. JASIST, 2009, 60 (10): 2107–2118.

作者简介: 吴思竹, 女, 1981年生, 博士研究生, 发表论文2篇。

张智雄, 男, 1971年生, 研究馆员, 发表论文70多篇。