“数据分析与计算专栏”主持人语

刘 滨

尊敬的读者朋友,您好! 在作者和编辑老师的共同努力下,本专栏又和您见面了! 文献大数据分析和重叠社区发现是大数据与社会计算领域的研究热点,也是本期关注的两个主题。文献大数据分析的研究目的之一,就是帮助广大研究人员快速进入目标领域,敏捷追踪其中的方向、方法和工具等,从而有效降低研究成本,缩短成果周期;重叠社区发现是复杂网络挖掘领域的基础性工作,成果应用于社交网络、通讯网络、蛋白质相互作用网络、代谢路径网络、交通网络等各种网络的数据分析中,可服务于人力资源管理、新药研制、交通规划、传染病防治、舆情控制等领域。

本期“数据分析与计算专栏”共推出 2 篇文章。

第 1 篇是翟君伟等撰写的“基于 LDA 主题模型的文献特征项多重共现可视化方法”。共现网络分析是文献计量研究中分析文献特征项数据关系的一种可视化方法,为解决多重共现网络中节点过多、连线密度过大、不利于发现数据的价值以及可视化效果较低等问题,引入 LDA 主题模型,采用空间划分的方法,将特征项全域可视化的问题转化为子空间可视化问题。首先,使用 SATI 文献题录信息分析软件抽取文献关键词,进行 TF-IDF 计算,以计算结果作为实验数据;其次,使用 Python 构建主题模型,对目标文献集合进行主题聚类分析;最后,使用 Ucinet 软件对不同主题子空间文献进行多重共现分析,并将子空间分析结果叠加和重构,完成多重共现可视化系统的结构化表达。结果表明:与原多重共现可视化方法相比,基于 LDA 主题模型的多重共现可视化改进方法降低了共现网络中的节点数量和节点间连线密度,使得多重共现可视化系统的结构更为清晰,增加了数据的可读性,突出了数据价值,有效提升了多重共现可视化效果。

第 2 篇是张妍等撰写的“基于社区森林模型的分布式重叠社区发现算法”。针对传统单机重叠社区发现算法已经不适于大规模复杂网络挖掘,以及不能支持网络表示学习任务相关要求的问题,提出了一种基于社区森林模型的分布式重叠社区发现算法(distributed community forest model,简称 DCFM 算法)。首先,将网络数据集存储到分布式文件系统,将数据分块,使用分布式计算框架在每个数据分块上执行 CFM 算法;然后,执行社区合并;最后,汇总社区划分结果,使用真实的 DBLP 数据集将算法运行于 Spark 集群上,采用 F 均值和运行时间对算法进行评估。结果表明:DCFM 算法的 F 均值稍逊于 CFM 算法,但其运算时间随着节点的增加接近线性下降,在牺牲小部分 F 均值的同时,DCFM 算法具备处理大规模网络数据的能力;分割份数对计算时间的影响也很大,在 com-dblp.ungraph.txt 数据集上,CFM 算法处理数据需要 192 min,而 DCFM 算法在将数据分成 6 份时,需要约 91 min,分成 100 份后仅需要约 13 min。因此,在大数据平台上采用分布式计算骨干度,进行社区划分、合并的 DCFM 算法是一种可行的大规模复杂网络挖掘方法,通过对网络进行分割,可以大幅加快社区划分速度,提高社区发现效率。

限于学术水平,以上文章定有可进一步完善之处,欢迎各位朋友不吝指正!

主持人简介

刘滨,男,1975 年出生,教授,硕士生导师,河北科技大学大数据与社会计算研究中心主任,主要研究领域包括网络新媒体、大数据、社会计算等。已发表学术论文 100 余篇,主持国家级科研课题 2 项、省级科研课题 9 项。