

## 多样性图排序的研究 现状及展望\*

文 / 程学旗 孙冰杰 沈华伟 余智华  
中国科学院计算技术研究所 北京 100190

**【摘要】** 排序是信息检索、数据挖掘以及社会网络分析的基础工作之一。在线社交网络和社会媒体的快速发展积累了大量的图数据——由表示实体的节点和表示实体间关系的连边构成。图数据中节点之间连接关系复杂,通常缺少显式的全序结构,使得图排序在图数据分析中显得尤为重要。图排序算法主要包括2大类,面向节点中心度的图排序算法和面向节点集合多样性的图排序算法。与传统的图排序不同,多样性图排序考虑排序和聚类的融合,体现为节点集合对网络整体的覆盖程度。近年来,多样性图排序得到了广泛的关注,取得了一系列研究进展,研究成果成功应用到了搜索结果排序、文档自动摘要、信息推荐系统和影响最大化等诸多场景中。文章评述了多样性图排序的研究现状及主要进展,将现有的多样性图排序方法按照研究思路的不同分为边际效益最大化、竞争随机游走、聚类与排序互增强3类,分别评述了每类方法的优势和不足。最后指出,设计有效的评价指标和标准测试集、克服多样性图排序面临的精度和速度的矛盾等是多样性图排序未来的研究重点。

**【关键词】** 图数据,多样性图排序,社交网络

DOI 10.16418/j.issn.1000-3045.2015.02.012

### 1 引言

排序(Ranking)是信息检索、数据挖掘以及社会网络分析的基础工作之一<sup>[1-3]</sup>。搜索引擎中,较好的排序方法可以保证在有限的显示空间中呈现与用户查询相关性较高、信息冗余度较低的检索结果,从而最小化用户的查询放弃率<sup>[1,4,5]</sup>,对于改善用户的查询体验意义重大。在信息推荐系统

中,如何对推荐的结果进行排序从而向用户提供更相关更丰富的推荐,对于推荐系统具有重要意义。在线社交网络分析中,如何度量用户的社会影响力,如何通过较好的排序方法选择出影响大且信息冗余小的节点集合使其影响范围最大,是在线社交网络分析的研究热点之一。早期的排序技术主要关注排序对象的重要性,例如搜索引擎

\* 基金项目:国家重点基础研究发展计划(“973”)项目(2013CB329602),国家自然科学基金项目(61425016、61472400)  
修改稿收到日期:2015年1月30日

中检索结果和用户查询的相关性,推荐系统中推荐对象和用户兴趣的匹配程度,社交网络分析中节点的中心度等。近几年,人们越来越多地开始关注排序的多样性,即排序靠前的对象之间的差异性。同时,在线社交网络和社会媒体的发展积累了大量的图数据——由表示实体的节点和表示实体间关系的连边构成,而图数据中缺少显式的序使图排序显得尤为关键。

多样性图排序对于在线社交网络分析具有重要的理论和技术价值。首先,多样性图排序衔接着社交网络分析的两个主要技术手段——排序和聚类。排序侧重于对网络进行纵向分析,得到网络节点的一个全序。聚类则是对网络的横向分析,将网络节点划分成若干个内部连接紧密而类间连接稀疏的类<sup>[6]</sup>。多样性图排序则兼而有之,在节点排序过程中考虑节点间的连接情况,使排序靠前的节点尽可能多地覆盖整个网络,因此排序靠前的节点大多对应着网络聚类的类中心节点。其次,多样性图排序和社交网络分析中的影响最大化、病毒式营销等应用问题紧密相关,并为其提供了一系列高效的算法<sup>[7]</sup>。最后,多样性图排序对于揭示网络结构规则、预测网络演化和预测网络信息传播等具有重要意义。譬如,将网络抽象成多个偏序构成的层级结构,有助于提取网络结构骨架和主要结构规则<sup>[8]</sup>。

近年来,多样性图排序得到了越来越多的关注,并取得了一系列研究进展,主要包括3大类研究思路。第一类基于边际效益最大化。1998年,杰米·卡博内尔(Jaime Carbonell)和杰德·戈尔茨坦(Jade Goldstein)<sup>[9]</sup>提出了最大边际相关度(MMR:Maximal Marginal Relevance)模型,提高搜索引擎中检索结果的排序多样性。该思路的关键问题是如何定义合适的目标函数。围绕

该问题,研究人员从相关度、差异性、子话题、网络覆盖等多个角度设计了不同的目标函数<sup>[4,10-17]</sup>,进而采用“贪心算法”实现多样性图排序,节点按照边际效益递减的顺序排列。第二类为网络随机游走方法。传统的随机游走中,节点排序主要取决于节点的中心度<sup>[18]</sup>,未考虑节点间连接的聚集性,导致排序靠前的节点集合中有很多冗余的信息,排序多样性差。为此,研究人员将竞争机制引入到随机游走中,通过节点之间的竞争实现多样性图排序<sup>[19-22]</sup>。第三类方法融合排序和聚类实现多样性图排序<sup>[23,24]</sup>。Sun等人<sup>[23]</sup>在RankClus的工作中首次将排序问题和聚类问题进行了结合,将排序和聚类视为两个相互促进的问题,按照聚类结果定义类别相关的排序,实现排序的多样性。除此之外,还有一些工作是通过寻找网络中影响力最大的节点来作为网络上多样性排序较好的节点集合<sup>[25,26]</sup>。

在线社交网络和社会媒体的快速发展,为多样性图排序的研究提供了很好的数据基础和应用场景。社会影响力度量、信息推荐、链路预测、社区发现<sup>[27-29]</sup>等诸多重要的社交网络分析问题,与多样性图排序之间有着紧密的关系。例如,社会化推荐中,选择合适的用户发起推荐,涉及用户间关系的结构多样性。因此,本文选择多样性图排序作为社交网络分析的切入点,对其研究现状进行评述,并展望其未来发展趋势。

## 2 多样性图排序问题

多样性图排序关注如何在节点的重要性和多样性之间做一个折中,使得选出的top- $k$ 节点集合更好地覆盖整个网络。多样性图排序的输入是图数据,例如社会关系网络,输出是节点的一种排序。一般而言,输出的排序综合考虑节点的中心度和节点间的多样性。排在前面的节点更能代表网络



中国科学院

的关键结构特征。经过10多年的发展,传统的多样性排序问题的基本理论已经基本形成,逐个迁移到图数据分析上,结合图数据分析的现有研究成果衍生出很多种多样性图排序方法。

图数据中,节点表示实体,边表示实体间的关系。通常,使用 $V$ 表示节点集合,矩阵 $W$ 表示节点间的连边,矩阵元素 $w_{ij}$ 表示连边权重。对于无权图,当 $i$ 和 $j$ 之间没有边相连时 $w_{ij}=0$ ,否则, $w_{ij}=1$ 。多样性图排序旨在将节点按照边际效益递减的方式进行排序。边际效益通常定义在节点集合上。对于节点集合 $S(S \subseteq V)$ ,边际效益通常是2个因素的折中:节点的中心度和节点间的多样性。节点中心度体现了节点在网络中的全局重要程度,例如度、PageRank值等;节点间的多样性通常采用节点相似度来刻画。

对于多样性图排序的评价,目前尚无一致认可的评价方式<sup>[30]</sup>。很多评价是将节点的中心度和多样性分别进行评价<sup>[19,20,31]</sup>。比如对中心度的评价有 Normalized Relevance<sup>[16]</sup>, nDCG等。对于多样性的评价主要有变化率(Difference Ratio)<sup>[31]</sup>、覆盖度等<sup>[19,20]</sup>。Tong等人<sup>[16]</sup>提出一个目标函数可以在考虑节点中心度和多样性的条件下综合评价集合的质量。但是他们提出的目标函数没有很好地反映出网络的拓扑结构特征,仍然没有解决多样性图排序的评价问题。

### 3 多样性图排序的研究现状

多样性排序的相关工作在1998年的最大边际相关度(MMR)工作中第一次被提出<sup>[9]</sup>。基于多样性排序的理论,本文总结出了多样性图排序研究工作的主要框架及其主要相关工作(图1)。

多样性图排序工作主要包含3种思路:边际效益最大化、竞争随机游走、排序和聚类互增强。本文将通过这3个思路下的代表工作对多样性图排序方法进行介绍。

#### 3.1 基于边际效益最大化的多样性图排序

边际效益最大化的一般框架为:定义一个衡

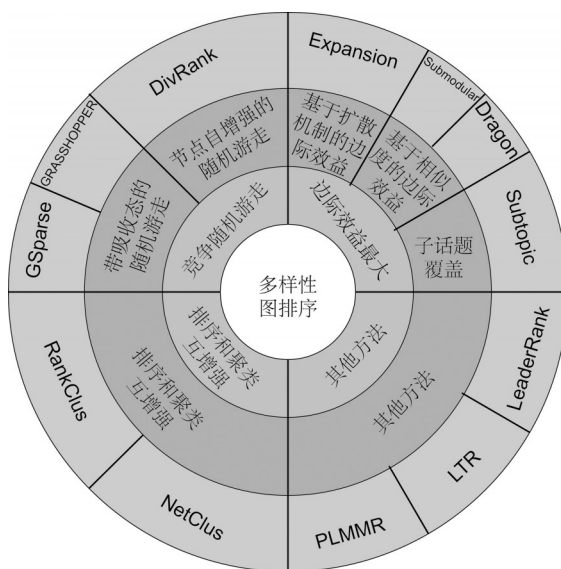


图1 多样性图排序研究现状

量节点集合的效益函数,然后逐个选择使效益函数增加最大的节点,从而形成一个按照边际效益递减的排序。效益函数定义大体上采用2种奖惩机制:(1)奖励扩散性强的节点,即奖励能使节点集合所覆盖的网络规模最大化的候选节点<sup>[10,11,17]</sup>;(2)惩罚冗余度大的节点,即惩罚与集合内部已选节点相似度高的候选节点<sup>[4,13-16]</sup>。基于上述效益函数的设计思路,研究人员在具体的应用背景下采用了不同的效益函数。代表性方法包括:

(1)子话题覆盖方法。该方法的应用场景是带查询的多样性图排序。考虑对与查询相关的子话题的覆盖程度<sup>[10,11]</sup>,其目标是利用统计语言模型对与查询相关的文档进行排序,希望排序靠前的文档可以覆盖到尽可能多的与用户查询话题相关的子话题。Zhai等人<sup>[10,11]</sup>首次提出子话题覆盖改善排序多样性的方法,致力于尽可能多地覆盖子话题。此后,Zhang等人<sup>[12]</sup>提出了邻近图模型(Affinity Graph Model),利用多样性和信息丰富程度(Information Richness)两个因素对话题进行重排序,前者反映话题的多样性,后者反映对话题的覆盖程度。邻近图模型的贡献在于将子话题覆盖问题抽象成了多样性图排序问题,利用贪心策略惩罚多样性差的节点,根据信息的覆盖程度和节点



多样性的综合得分对图中的节点进行重排序选择 top- $k$  的节点集合,实现了排序的多样性。

(2) 基于相似度的边际效益。MMR 方法最早提出了基于相似度的边际效益函数,综合候选节点和已选节点的相似度、候选节点和查询的相似度两个指标,实现搜索引擎检索结果的排序多样性。本质上,MMR 是一种启发式方法,直接定义了边际效益,但未明确给出所优化的效益函数。

Lin 等人<sup>[4,13-15]</sup>随后对 MMR 进行扩展,提出了效益函数应该具有的一般性质。具体地讲,效益函数  $f(\cdot)$  是定义在集合  $S$  上的一个函数,满足 3 个性质:

子模性 (Submodularity): 如果集合  $S \subseteq R \subseteq V$ , 则将一个元素  $k \in V$  加入到集合  $S$  所带来的边际效益不小于将该元素加入到集合  $R$  所带来的边际效益,即  $f(S \cup \{k\}) \geq f(R \cup \{k\})$ ;

非减性 (Non-decreasing): 对于任意的  $S \subseteq R \subseteq V$ , 满足  $f(S) \leq f(R)$ ;

标准性:  $f(\emptyset) = 0$ 。

对于满足上述 3 个性质的效益函数,采用贪心算法得到的集合  $S$  不会比最优的集合  $S^*$  的  $1 - 1/e$  差,即  $f(S) \geq (1 - 1/e)f(S^*)$ 。

效益函数和贪心算法,为多样性图排序提供了一个一般性的框架。定义不同的效益函数对应着不同的多样性图排序算法,贪心算法保证算法的精度。定义效益函数时,通常综合考虑单个节点的中心度和节点间的差异性。例如, Tong 等人提出的效益函数<sup>[16]</sup>,采用个性化 PageRank 算法度量单个节点的中心度,节点集合  $S$  的多样性采用连边权重来度量。但由于中心度和多样性的计算标准不一,导致算法执行前期侧重于选择中心度较高的节点,后期则侧重于节点多样性。为此,研究人员开始考虑设计新的效

益函数,度量对象从单个节点的角度转向节点集合  $S$ ,代表性方法是随后介绍的基于扩散机制的边际效益。

(3) 基于扩散机制的边际效益。基本思路是考虑如何选择节点使得节点的中心度较好且加入该节点能使节点集合  $S$  扩散到的网络规模增长较大。该思路与之前基于节点相似度的效益函数定义方式异曲同工,不同的是基于扩散机制的方法不再侧重于节点间的相似度或连边权重,转而考虑了节点集合对整个网络的覆盖程度,例如 L-阶共同邻居特征和扩展度<sup>[17]</sup>。

以扩展度为例来介绍<sup>[17]</sup>。普通的 PageRank 只考虑了节点的中心度,得到的是访问频率高的节点。由于没有考虑节点多样性, top- $k$  的集合中会存在冗余信息。为此,基于扩展度的效益函数,利用个性化 PageRank 计算节点的中心度之后,采用 L-阶共同邻居数量作为评价指标来衡量集合  $S$  中节点的多样性。扩展度越大,说明集合  $S$  内部节点的多样性越好。

### 3.2 基于竞争随机游走的多样性图排序

网络上的随机游走可以用于识别中心度高的节点,随机游走中频繁访问的节点具有较高的中心度。然而,随机游走未考虑节点之间的连接关系,排序靠前的节点多样性差。基于随机游走的多样性图排序,大多在随机游走过程中引入节点间的竞争机制,通过让彼此相连的节点相互竞争,实现节点多样性。按照竞争策略不同,有 2 类代表性方法:节点自增强 (Vertex-reinforced) 的随机游走策略和带吸收态的随机游走策略。

(1) 节点自增强的随机游走策略。着眼于改变随机游走中的转移概率矩阵,实现节点间的竞争。代表性工作是 Mei 等人提出的 DivRank<sup>[19]</sup>。DivRank 提供了一种转移概率随时间变化的随机游走策略,在随机游走



中国科学院

中引入“富者愈富”(Rich-get-richer)机制<sup>[32]</sup>。具体而言,DivRank是对随机游走采用了动态更新的机制,即节点之间的转移概率随时间而变化,节点倾向于到达排序值更高的节点,得分较大的节点会竞争得到相邻节点的得分。DivRank最终得到的节点既具有高的中心度,又具有高的多样性。然而,该方法的计算效率低,难以适用于大规模的网络。

(2)带吸收态的随机游走策略。代表性工作是Zhu等人提出的GRASSHOPPER算法<sup>[20]</sup>、Onur Küçüktunç等人提出的GSparse<sup>[21]</sup>以及Cheng等人提出的带汇点的流形排序算法MRSP<sup>[22]</sup>。此类方法利用带吸收态的随机游走(Absorbed Random Walk)依次选择节点。对于被选中的节点,将其置为吸收态,然后再进行随机游走,选择下一个节点。GRASSHOPPER和MRSP每次将随机游走访问频率最高的节点置为吸收态,处于吸收态的节点,转移出去的概率为0,类似一个随机游走的“黑洞”。GSparse将与候选节点相连的所有连边去掉,使得候选节点周围的节点被随机游走访问的概率降低,从而实现多样性图排序。此类方法缺少明确的优化目标,可解释性差。

### 3.3 基于排序和聚类互增强的多样性图排序

排序和聚类长期以来作为2个单独的问题来求解。最近,Sun等人首次将排序问题和聚类问题进行了结合,利用混合模型来改善排序的方法。RankClus<sup>[23]</sup>和NetClus<sup>[24]</sup>等通过排序<sup>[33]</sup>和聚类<sup>[34]</sup>的互增强来同时改善排序和聚类。RankClus针对双关系型异质网络(例如:“会议-作者”网络)进行排序和聚类时,使用了混合模型来建模节点之间边的生成过程,每个节点在 $k$ 个类别的分布符合多项分布,利用目标类型(Target Type)的聚类结果优化排序结果,然后再用优化之后的排序结果反过来对目标类型的聚类结果进行优化,直到优化前后结果变化较小为止。与RankClus不同的是,NetClus不是对某一类型的节点进行排序和聚类,而是针对多关系类型的星形网络进行综合排序和聚

类。

### 3.4 其他多样性图排序方法

除上述3类主要方法之外,概率化隐式最大边际相关度方法PLMMR<sup>[35]</sup>为MMR提供了一种隐空间的概率解释形式,不需要手动调节MMR中参数的取值。基于有监督学习来训练排序模型的代表工作有基于用户点击行为数据(Click-through Data)进行分析来优化排序的方法<sup>[4,36,37]</sup>;通过从训练数据<sup>[38]</sup>中学习得到边际效益函数来预测文档的排序,如SVMDIV<sup>[39]</sup>、INDSTRSVM<sup>[40]</sup>,以及Learning-to-rank(LTR)<sup>[41-43]</sup>等。也有单纯通过改善PageRank方法来识别重要的节点(Leader Node)从而找到一个节点排序的工作,比如,LeaderRank<sup>[25]</sup>通过添加一个公共节点使得全网可以满足强连通并且使得出度较大的节点具有更小的跳转概率,更好地识别出高重要度的节点。

## 4 多样性图排序问题的趋势展望

多样性排序最初是为改善信息检索而进行的重排序(Reranking)过程。在近几年的发展中,多样性排序在信息检索、文档分类和聚类、协同过滤、关键词抽取、问答系统、多文档摘要、观点挖掘、情感分析、机器翻译、句法分析等研究问题中体现出了应用价值<sup>[44-46]</sup>。近年来,社交网络和社会媒体的兴起提供了大量的图数据,使得多样性排序被更多地应用到了图数据上,多样性图排序的研究日益增多,并在一系列社交网络分析问题中取得了成功的应用,比如通过多样性排序进行影响最大化和病毒式营销等。

多样性图排序方法借助已有的多样性排序理论,发展到现在可以在合适规模的网络上找到一个给定数目的节点的排序集合。下一步面临的问题主要有以下几点:(1)如何评价找到的集合 $S$ 的质量,如何构造一个标准答案(Ground Truth)或者评测集(Benchmark)使得对发现的节点集合的评价更合理;(2)现有的方法需要遍历图中的所有节点,不适用于大规模的网络,需要寻找一种能够体

现网络拓扑结构特征并且快速高效的方法,比如,考虑抽样方法的应用;(3)现有方法都是一个节点一个节点利用“贪心方式”进行节点的选择,每次选择的节点会依赖于已选的节点,而已选节点不会再重新进行更新,因此所得到的集合并不是全局最优的<sup>[47]</sup>,我们需要考虑一种可以更新已选节点的方法以找到更接近全局优化的集合。

既然是多样性图排序,我们希望得到的节点集合在评价方式上更符合图的评价方式要求。多样性图排序目标是寻找一个节点集合能够较好地表示图的结构特征(如,社区的划分)。因此可以通过将集合  $S$  中的  $\text{top-}k$  个节点扩展为  $k$  个社区然后以社区划分的质量来评价多样性图排序方法的优劣,而社区划分的质量是可以通过一个统一的指标来衡量的,也有大家比较认可的评测集构造方法<sup>[48]</sup>。

## 5 结论

目前多样性图排序方法的理论基础已经基本完善,但是缺乏令人信服的评价方式和可以用于评价的标准答案以及评测集的构造方法,并且难以高效地应用于实际网络。后续研究需要借鉴网络结构分析等领域的一些工作进行。

### 参考文献

- 1 Clarke C L A, Kolla M, Cormack G V et al. Novelty and diversity in information retrieval evaluation In Proc. of SIGIR, 2008, 659-666.
- 2 Drosou M, Pitoura E, Search result diversification. SIGMOD Rec., 2010, 39: 41-47.
- 3 Gollapudi S, Sharma A. An axiomatic approach for result diversification, In Proc. of WWW, 2009, 381-390.
- 4 Agrawal R, Gollapudi S, Halverson A et al. Diversifying search results. In Proc. of WSDM, 2009, 5-14.
- 5 Sarma A D, Gollapudi S, Jeong S. Bypass rates: reducing query abandonment using negative inferences. In KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008, 177-185.

- 6 Fortunato S. Community detection in graphs. Physics Reports, 2010, 486(3-5): 75-174.
- 7 Cheng S, Shen H, Huang J et al. StaticGreedy: solving the scalability-accuracy dilemma in influence maximization, In Proc. of CIKM, 2013, 509-518.
- 8 Cheng X Q, Ren F X, Shen H W et al. Bridgeness: A local index on edge significance in maintaining global connectivity. Journal of Statistical Mechanics, 2010, 10011.
- 9 Carbonell J, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries, in Proc. of SIGIR, 1998, 335-336.
- 10 Zhai C X, Cohen W W, Lafferty J. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval In Proc. of SIGIR, 2003, 10-17.
- 11 Zhai C X, Lafferty J. A risk minimization framework for information retrieval. Information Processing & Management, 2006, 42(1): 31-55.
- 12 Zhang B, Li H, Liu Y et al. Improving web search results using affinity graph, In Proc. of SIGIR, 2005, 504-511.
- 13 Lin H, Bilmes J, Xie S. Graph-based submodular selection for extractive summarization. In automatic speech recognition and understanding workshop, 2009.
- 14 Lin H, Bilmes J. Multi-document summarization via budgeted maximization of submodular functions In Proc. of HLT, 2010, 912-920.
- 15 Lovász L. Submodular functions and convexity. Mathematical programming-The state of the art, (eds. A. Bachem, M. Grotschel and B. Korte) Springer, 1983, 235-257.
- 16 Tong H, He J, Wen Z et al. Diversified ranking on large graphs: an optimization viewpoint, In Proc. of KDD, 2011, 1028-1036.
- 17 Li R H, Yu J X. Scalable diversified ranking on large graphs. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(9): 2133-2146.



中国科学院

- 18 Brin S, Page L. Pagerank: Bringing order to the web , Stanford digital library project, Tech. Rep., 1997.
- 19 Mei Q, Guo J, Radev D. Divrank: the interplay of prestige and diversity in information networks. In Proc. of KDD, 2010, 1009-1018.
- 20 Zhu X, Goldberg A B, Van Gael J et al. Improving diversity in ranking using absorbing random walks. In Proc. of HLT-NAAACL, 2007, 97-104.
- 21 Küçüktunç O, Saule E, Kaya K et al. Diversifying citation recommendation. Technical Report arXiv:1209.5809, ArXiv, Sep 2012.
- 22 Cheng X Q, Du P, Guo J et al. Ranking on data manifold with sink points. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(1) : 177-191.
- 23 Sun Y, Han J, Zhao P et al. Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In Proc. of EDBT, 2009: 565-576.
- 24 Sun Y, Yu Y, Han J. Ranking-based clustering of heterogeneous information networks with star network schema. In Proc. of KDD, 2009, 797-806.
- 25 Lü L, Zhang Y C, Yeung C H et al. Leaders in social networks, the delicious case. PloS One, 2011, 6(6) : e21202.
- 26 Huang S, Cui H, Ding Y. Evaluation of node importance in complex networks. arXiv preprint: 1402.5743, 2014.
- 27 Newman M E J, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004, 69(2) : 026113.
- 28 Newman M E J, Leicht E A. Mixture models and exploratory analysis in networks. Proceedings of the National Academy of Sciences, 2007, 104(23) : 9564-9569.
- 29 Shen H W, Cheng X Q, Guo J F. Exploring the structural regularities in networks. Physical Review E, 2011, 84(5) : 056111.
- 30 Radlinski R, Bennett P N, Carterette B et al. Redundancy, diversity and interdependent document relevance. SIGIR Forum, 2009, 43(2) : 46-52.
- 31 Küçüktunç O, Saule E, Kaya K et al. Diversified recommendation on graphs: pitfalls, measures, and algorithms. In Proc. of WWW, 2013, 715-726.
- 32 Reiman J H, Leighton P. The rich get richer and the poor get prision: Ideology, class, and criminal justice. New York: Macmillan, 1990.
- 33 Altman A, Tennenholtz M. On the axiomatic foundations of ranking system. In Proc. of IJCAI, 2005, 917-922.
- 34 Kleinberg J. An Impossibility Theorem for Clustering. In Proc. of NIPS, 2003.
- 35 Guo S, Sanner S. Probabilistic latent maximal marginal relevance. In Proc. of SIGIR, 2010, 833-834.
- 36 Joachims T. Optimizing search engines using clickthrough data. In Proc. of KDD, 2002, 133-142.
- 37 Radlinski F, Kleinberg R, Joachims T. Learning diverse rankings with multi-armed bandits. In Proc. of ICML, 2008, 784-791.
- 38 Duh K, Kirchhoff K. Learning to rank with partially-labeled data. In Proc. of SIGIR, 2008, 251-258.
- 39 Yue Y, Joachims T. Predicting diverse subsets using structural SVMs. In Proc. of ICML, 2008, 271-278.
- 40 Li L, Zhou K, Xue G R et al. Enhancing diversity, coverage and balance for summarization through structure learning. In Proc. of WWW, 2009, 71-80.
- 41 Agichtein E, Brill E, Dumais S T, et al. Learning user interaction models for predicting web search result preferences. In Proc. of SIGIR, 2006, 3-10.
- 42 Amini M R, Truong T V, Goutte C. A boosting algorithm for learning bipartite ranking functions with partially labeled data. In Proc. of SIGIR, 2008, 99-106.
- 43 Huang J C, Frey B J. Structured ranking learning using cumulative distribution networks. In Proc. of NIPS, 2008, 697-704.
- 44 Ailon N, Mohri M. An efficient reduction from ranking to classification. In Proc. of COLT, 2008.
- 45 Balcan M F, Bansal N, Beygelzimer A et al. Robust reductions from ranking to classification. In Proc. of COLT, 2007.
- 46 Goldstein J, Mittal V, Carbonell J et al. Multi-document summarization by sentence extraction. In NAACL-ANLP 2000 Workshop on Automatic summarization, 2000, 40-48.
- 47 McDonald R. A study of global inference algorithms in multi-document summarization. Lecture Notes in Computer Science, 2007, 4425, 557.



48 Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms. Phys-

ical Review E, 2008, 78(4): 046110.

## Research Status and Trends of Diversified Graph Ranking

Cheng Xueqi Sun Bingjie Shen Huawei Yu Zhihua

(CAS Key Laboratory of Network Data Science and Technology, Institute of Computing Technology,  
Chinese Academy of Sciences, Beijing 100190, China)

**Abstract** Ranking is a fundamental task in information retrieval, data mining, and social network analysis. With the rapid proliferation of online social network and social media, a great deal of graph data has been accumulated. Graph data is made up of nodes, representing entities, and edges, characterizing relationships among entities. In graph data, nodes are connected through heterogeneous relationships, lacking explicit order among them. Therefore, graph ranking is particularly important for graph data analysis.

Existing graph ranking algorithms could be roughly classified into two categories, respectively graph ranking based on centrality of nodes and graph ranking on diversity of a set of nodes. For centrality-based graph ranking, spectral analysis is main-stream technique, where nodes are ranked according to their magnitude in primary eigenvectors of certain matrix, like adjacency matrix, Google matrix, and modularity matrix. These methods are particularly useful at offering a global ranking of nodes. However, a global ranking is not sufficient in many scenarios, such as personalized recommendation, influence maximization, and ranking of search results. Hence, researchers begin to study the diversified ranking on graphs. Different from traditional ranking on graph, diversified graph ranking focuses on the interplay of ranking and clustering, capturing the intrinsic structure regularity of graph. In recent years, diversified ranking on graph attracts great attention, and several methods are proposed and successfully applied in many scenarios, including search results ranking, information recommendation system, document summarization, and influence maximization. In this paper, we summarize main research progress about diversified ranking on graph.

We roughly classified existing methods into three categories, respectively based on marginal-benefit maximization, random walk with node competition, and reinforcement between clustering and ranking. For marginal-benefit maximization, a benefit function is generally defined for optimization. With such a benefit function, nodes are selected one by one according to their marginal benefit. Each time, the node with the maximum marginal benefit is selected. In this way, nodes are ranked in the decreasing order of marginal benefit. The performance of this kind of methods depends on the benefit function. The other kind of methods are based on random walk. Traditional random walk is essentially a kind of centrality-based method. To guarantee the diversity of ranking, researchers try to introduce competition among adjacent nodes into traditional random walk. Two typical examples are DivRank and Grasshopper. DivRank revises the transition probability among nodes to implement a rich-get-richer mechanism, forcing adjacent nodes to compete for scores and consequently form a diversified ranking. Grasshopper adopts a multi-step manner, selecting one node at each iteration and taking the selected nodes as sink point to penalize its adjacent nodes in following iterations. The third kind of diversified



中国科学院



graph ranking exploits the reinforcement between ranking and clustering. Cluster-dependent ranking and ranking-based clustering are iteratively updated to form a diversified ranking.

Finally, we introduce some potential research trends for diversified graph ranking. We point out the community lacks a commonly-available and widely-accepted benchmark for diversified ranking on graphs. A benchmark is particularly valuable for filtering out state-of-the-art methods. Meanwhile, we also need to design good evaluation metrics for diversified graph ranking. With the rapid increase of the scale of graph data, to offer an off-the-shelf method, scalability is another key issue for future research.

**Keywords** graph data, diversified graph ranking, social network

**程学旗** 中科院计算技术所副总工程师, 中科院网络数据科学与技术重点实验室主任、研究员、博士生导师。1994、1996 年分别获东北大学计算机科学与技术专业学士与硕士, 2006 年获中科院计算技术所计算机系统结构博士学位。在信息网络建模与社区结构分析、Web 搜索与挖掘等领域发表学术论文 150 余篇, 获授权发明专利 17 项, 软件著作权 27 项。中国计算机学会大数据专委会秘书长、中文信息学会信息检索与内容安全专委会常务副主任、国家信息安全专项计划管理专家。E-mail: cxq@ict.ac.cn

**Cheng Xueqi**, is a full Professor in the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS). He is the director of CAS Key Lab of Network Data Science and Technology. He received his bachelor and master degree from Northeastern University of China in 1994 and 1996, separately. He received his Ph.D. degree from ICT-CAS in 2006. His research interests include network science, social and information network analysis, Web search and data mining. He has published more than 150 papers in these areas. He also has 17 patents and 27 software copyrights. He is the Secretary General of Big Data Society, China Computer Federation. He is also the vice president of the Society of Information Retrieval and Content Security, Chinese Information Processing Society of China. E-mail: cxq@ict.ac.cn