软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2018,29(3):599−613 [doi: 10.13328/j.cnki.jos.005455]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于距离度量的多样性图排序方法*

李 劲[1,2], 岳 昆[3], 蔡 娇[1], 张志坚[3], 刘惟一[3]

[1](云南大学 软件学院,云南 昆明 650091)

[2](云南省软件工程重点实验室,云南 昆明 650091)

[3](云南大学 信息学院,云南 昆明 650091)

通讯作者: 岳昆, E-mail: kyue@ynu.edu.cn

摘 要: 有效结合查询相关性和多样性的扩展相关性,是多样性图排序问题的一种优化目标.基于扩展相关性的多样性图排序可建模为一个子模函数优化问题,贪心子模优化算法可近似求解该问题.然而,扩展相关性不能直接度量节点间的不相似性.子模优化算法是串行算法,不能充分利用诸如 Spark 等集群计算平台有效提高算法效率.针对这些问题,提出一种描述节点间不相似性的距离度量.基于该距离度量,将多样性图排序问题建模为一个在查询相关节点集上构造的带权完全图的最大和 *k*-dispersion 优化问题.提出了求解该问题的多项式时间 2-近似算法.鉴于不同节点对的距离度量计算是相互独立的,进一步提出了基于 MapReduce 编程模型的并行化多样性图排序算法.最后,在真实图数据集上验证了所提出算法的高效性和有效性.

关键词: 图数据;个性化 PageRank;多样性图排序;最大和 *k*-dispersion;MapReduce

中图法分类号: TP311

中文引用格式: 李劲,岳昆,蔡娇,张志坚,刘惟一.基于距离度量的多样性图排序方法.软件学报,2018,29(3):599−613. http://www.jos.org.cn/1000-9825/5455.htm

英文引用格式: Li J, Yue K, Cai J, Zhang ZJ, Liu WY. Distance metric based diversified ranking on large graphs. Ruan Jian Xue Bao/Journal of Software, 2018,29(3):599−613 (in Chinese). http://www.jos.org.cn/1000-9825/5455.htm

Distance Metric Based Diversified Ranking on Large Graphs

LI Jin[1,2], YUE Kun[3], CAI Jiao[1], ZHANG Zhi-Jian[3], LIU Wei-Yi[3]

[1](School of Software, Yunnan University, Kunming 650091, China)

[2](Key Laboratory of Software Engineering of Yunnan Province, Kunming 650091, China)

[3](School of Information Science and Engineering, Yunnan University, Kunming 650091, China)

Abstract: Expansion relevance which combines both relevance and diversity into a single function is resorted to a submodular optimization objective that can be solved by applying the classic cardinality constrained monotone submodular maximization. However, expansion relevance do not directly capture the dis-similarity over a pair of nodes. Existing submodular algorithms are sequential and not easy to take full advantage of the power of distributed cluster computing platform, such as Spark, to significantly improve the efficiency

* 基金项目: 国家自然科学基金(61562091, 61472345); 第二批"云岭学者"培养项目(C6153001); 云南省应用基础研究计划(2014FA023, 2016FB110); 云南大学中青年骨干教师培养计划项目; 云南大学青年英才培育计划(WX173602); 云南大学数据驱动的软件工程科技创新团队项目(2017HC012)

Foundation item: National Natural Science Foundation of China (61562091, 61472345); Program for the Second Batch of Yunling Scholar of Yunnan Province (C6153001); Natural Science Foundation of Yunnan Province (2014FA023, 2016FB110); Foundation of Backbone Teacher Development of Yunnan University (WX173602); Program for Excellent Young Talents of Yunnan University (XT412003); Project of Data Driven Software Engineeringinnovation Team of Yunnan University, Yunnan Province (2017HC012)

本文由基于图结构的大数据分析与管理技术专刊特约编辑林学民教授、杜小勇教授、李翠平教授推荐.

收稿时间: 2017-08-02; 修改时间: 2017-09-05; 采用时间: 2017-11-07; jos 在线出版时间: 2017-12-05

CNKI 网络优先出版: 2017-12-06 16:36:55, http://kns.cnki.net/kcms/detail/11.2560.TP.20171206.1636.031.html

of algorithm. To tackle this issue, in this paper, a distance metric, which is defined by a sum function of personalized PageRank scores over the symmetry difference of neighbors of a pair of nodes, is first introduced to capture the pairwise dis-similarity over pairs of nodes. Then, the problem of diversified ranking on graphs is formulated as a max-sum *k*-dispersion problem with metrical edge weight. A polynomial time 2-approximate algorithm is proposed to solve the problem. Considering the computational independence of different pairs of nodes, a MapReduce algorithm is further developed to boost the efficiency of the process. Finally, extensive experiments are conducted on real network datasets to verify the effectiveness and efficiency of the proposed algorithm.

Key words: graph data; personalized PageRank; diversified graph ranking; max-sum *k*-dispersion; MapReduce

当前,各种在线社交网络应用的快速发展累积了大量的图数据.图数据由节点和边组成,且节点之间往往存在复杂的连接关系,通常缺少显式的全序结构,使得图排序(graph ranking)成为图数据挖掘、分析的重要手段[1].

PageRank[2]是图中节点重要性度量的常用方法,其基本思想是:节点重要性由其邻接节点重要性决定,被重要节点连接的节点,重要性越高.一般来说,PageRank 值度量了图中节点的全局重要性或权威度,但 PageRank 值却不能很好地度量图上节点间的相似性.为此,研究者进一步提出了个性化 PageRank(personalized PageRank,简称 PPR)[3].但如文献[4,5]指出:在使用 PPR 方法进行相似性查询时,其返回的 top-*k* 结果只考虑查询相关性(relevance)而忽略了多样性(diversity).然而在实际图排序应用环境中,用户的查询意图很难准确描述,往往具有不确定性、模糊性以及多义性.因此,PPR 方法仍不能很好地满足图排序应用的要求.

当用户真实的查询意图难以准确获取时,对查询结果进行多样性处理,是信息检索系统解决此难题的有效方法[6].为提供高质量的查询结果,提高用户查询的满意度,提出能够有效折中相关性和多样性的图排序算法,是图数据挖掘、分析领域面临的研究挑战.为此,近年来许多学者提出了一系列的多样性图排序方法[4,5,7−10].

目前,已有的多样性图排序方法主要分为以下两类.

(1) 基于优化的方法[7−9].

这一类方法通过建立目标函数来刻画相关性和多样性,并将问题建模为该目标函数的优化问题,进而提出相应的优化算法.代表性工作有:Tong 等人[7]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,但是该目标函数没有考虑图数据固有的拓扑结构特征,不能很好地解决多样性图排序的评价问题.Li 等人[8]将多样性图排序建模为一个双目标优化问题,其中,用排序结果集的 PPR 值之和作为相关性度量,并提出扩展率(expansion ratio)来度量排序结果的多样性.Kucuktunc 等人[9]改进了文献[8]的工作,将相关性和扩展率进行融合,进一步提出了扩展相关性(expansion relevance)来度量排序结果的多样性.上述方法多样性度量指标不尽相同,但最终的目标函数均证明为非负、单调的子模函数(submodular function).因此,虽然多样性图排序问题是 NP-hard 的,但采用贪心算法可在多项式时间近似求解该问题;

(2) 基于随机游动的方法[4,5,10].

随机游动是度量图上节点重要性的有效方法,然而常规的随机游动方法未考虑节点之间的连接关系,因此排序结果多样性差.为此,相关学者提出了改进的随机游动方法来增强排序结果的多样性.代表性的研究工作有 DivRank[4]、GSparse[5]、GrassHopper[10].这类方法在随机游动过程中充分考虑了节点间的连接关系,通过引入节点间的竞争机制,让彼此相连的节点相互竞争,实现排序结果的多样性.然而,这类方法要么计算效率低,难以适用于大规模网络,要么缺少明确的优化目标,不具备可解释性[1].

综上,面对大规模图数据的多样性图排序工作,在优化目标、计算效率方面仍存在研究挑战.具体地:

首先,多样性度量指标是多样性图排序建模的核心问题.文献[8]提出了扩展率度量多样性,文献[9]将相关性和多样性进行融合,进一步提出扩展相关性.实际上,扩展相关性就是一种以节点 PPR 值为权重的带权扩展率.采用扩展率或者扩展相关性对多样性进行建模的基本思想是:任意两个节点的共同邻居越少,两个节点越不相似.节点集中,节点间不相似程度越高,则节点集的扩展率或者扩展相似性越大.然而,扩展率和扩展相关性的定义并非直接基于结点间的不相似性来描述节点集的多样性.换言之,具有高扩展率或高扩展相关性的节点集,其内部节点间的不相似性未必高.

下面举例说明此问题.先给出扩展率的形式化描述[8].令 $G=\langle V,E\rangle$是一个图,其中,V 是节点集,E 边集.令 $S\subseteq V$

是图上的任意节点集.$N(S)$是 S 的扩展集,即 $N(S)=S\cup\{v\in(V-S)|\exists u\in S,(u,v)\in E\}$,那么 S 的扩展率定义为 $er(S)=|N(S)|/|V|$.如图 1 所示:图 1(a)、图 1(b)子图中,节点集 $S=\{1,2\}$的扩展集均为 $N(S)=\{1,2\}\cup\{3,4,5,6\}$,由扩展率定义可知,两个子图中 S 的扩展率均为 $er(S)=|N(S)=\{1,2,3,4,5,6\}|/|V=\{1,2,3,4,5,6\}|=1$.然而,图 1(a)中节点 1 与节点 2 具有相同邻居,节点 1 和节点 2 应具有高度相似性.图 1(b)中的节点 1 与节点 2 无共同邻居,应不相似.因此,扩展率或者扩展相关性无法有效度量图 1 所示情形中节点之间的不相似性,进而无法有效度量结点集的多样性.

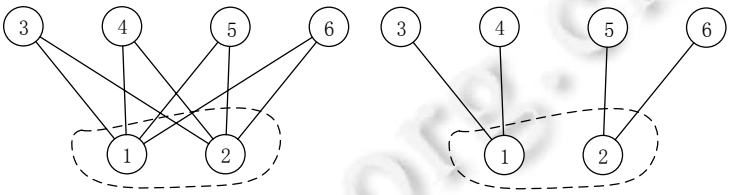


Fig.1 Illustration of deficiency of expansion ratio on measuring dis-similarity between two nodes

图 1 基于扩展率度量节点间的不相似性的不足之处

其次,已有的基于优化的方法均将多样性图排序建模为一个基数约束下的非负、单调子模函数最大化问题.该问题是 NP-hard 的,可采用贪心算法近似求解[11].算法从空的解集开始,每次迭代时,选取具有最大边际贡献的节点加入到当前解集中.对于含 m 个节点的查询相关集来说,为获取 top-k 排序结果,需要进行 $O(mk)$次边际贡献评估.当对大规模图数据进行多样性图排序时,计算代价高,算法的可扩展性成为很大的挑战.此外,由于贪心算法的执行过程是一个串行迭代过程,已有方法很难充分利用诸如 Spark GraphX[12]等图数据计算平台进行高效并行计算.

于是,对于多样性图排序研究,本文提出两个问题:能否提出基于节点间不相似性的节点集多样性度量指标?基于此度量指标提出的多样性图排序问题能否得到高效求解?

针对这两个问题,首先,提出了一种基于节点邻接信息的节点间不相似性的度量指标,并证明了该度量是节点间的一种距离度量(distance metric),即该度量满足非负、对称以及三角不等式等性质.基于该距离度量,将多样性图排序建模为一种带权完全图上的组合优化问题.具体地,以 PPR 查询节点构成节点集,以节点间的距离度量作为边权重构造带权完全图.求解 top-k 多样性图排序结果归结为求解该带权完全图的含 k 个结点的最大权子图.尽管该问题被证明是 NP-hard 的,得益于边权重满足距离度量性质,提出了多项式时间的 2-近似算法求解该问题.此外,由于带权图上不同节点对的距离度量计算是相互独立的,进而提出了基于 GraphX 的并行化算法.最后,在真实图数据集上,将本文方法与已有方法在算法执行时间、相关性以及多样性上进行了对比测试,实验结果验证了本文提出方法的有效性.

本文第 1 节给出节点间不相似性的距离度量定义,进而给出基于此距离度量定义的多样性图排序问题模型.第 2 节给出多样性图排序的(并行化)求解算法.第 3 节给出实验结果.第 4 节总结全文并展望将来的工作.

1 问题建模

本节中,先建立节点间不相似性的距离度量.基于此距离度量,将多样性图排序问题建模为一种带权完全图上的组合优化问题.

1.1 节点间的距离度量

节点间不相似性的度量标准是多样性图排序的基本问题.本文所讨论的多样性图排序问题不涉及图或边的属性信息,因此,节点的邻居节点集成为定义节点间不相似性的重要依据.直观地,两个节点的非共同邻居越多,则节点间的不相似程度越高.如果考虑到节点的 PPR 值,那么节点间非共同邻居的 PPR 值之和越大,则节点

之间越不相似.基于上述观察,节点的非共同邻居集可作为节点间不相似性的度量基础.于是,令 f 是定义在节点集上的函数,基于 f 可建立图 G 上任意两节点 v,u 之间的不相似度 $d_f(v,u)$.

定义 1. 令 $G=\langle V,E\rangle$是一个图,$N(v)$是节点 v 的邻居集.$\forall v,u\in V$,基于函数 f 的不相似度记为 $d_f(v,u)$,并定义:

$$d_f(v,u)=f(N(v)\oplus N(u)) \tag{1}$$

其中,⊕是集合间对称差运算,具体含义为:给定任意 2 个集合 $A,B,A\oplus B=(A-B)\cup(B-A)$.

在进一步讨论 $d_f(v,u)$前,先给出 v,u 之间距离度量的定义.

定义 2. 令 $d_f(v,u)$是 v,u 间的不相似度若 $d_f(v,u)$满足下面 3 条性质,则称 $d_f(v,u)$为 v,u 间的一种距离度量[13].

(a) $d_f(v,u)$满足非负性,即 $d_f(v,u)\geqslant 0$;

(b) $d_f(v,u)$满足对称性,即 $d_f(v,u)=d_f(u,v)$;

(c) $d_f(v,u)$满足三角不等式,即$\forall v,a,u\subseteq V,d_f(v,u)\leqslant d_f(v,a)+d_f(a,u)$.

由上述定义可知,$d_f(v,u)$的含义和性质决定于函数 f.下面证明:如果 f 是节点集上的权重和函数,则由 f 决定的 $d_f(v,u)$是 v,u 间的一种距离度量.

定理 1. 令 $A\subseteq V$ 是任意节点集,$w(v)\in\mathcal{R}^+$是节点的非负权重.令 f 是节点权重和函数,即 $f(A)=\sum_{v\in A}w(v)$,并记 $W=\sum_{v\in V}w(v)$.那么由 f 决定的不相似度是 v,u 间的一种距离度量.

$$d_f(v,u)=f(N(v)\oplus N(u))=\sum\nolimits_{v\in N(v)\oplus N(u)}(w(v)/W) \tag{2}$$

证明:设 v,a,u 是 V 集中任意 3 个节点,对于任意节点 v,其邻居节点集记为 $N(v)$.为简化记号,设 $N(v)=X,N(a)=Y$ 以及 $N(u)=Z$.按照定义 2,从以下 3 个方面分别进行证明.

(a) $d_f(v,u)$满足非负性,即 $d_f(v,u)\geqslant 0$.由定义 1 可知,$d_f(v,u)=f(X\oplus Z)$.由于 f 是节点集上的非负权重和函数,故 $f(X\oplus Z)\geqslant 0$,于是 $d_f(v,u)\geqslant 0$;

(b) $d_f(v,u)$满足对称性,即 $d_f(v,u)=d_f(u,v)$.$d_f(v,u)=f(X\oplus Y)$,且 $d_f(u,v)=f(Y\oplus X)$.⊕满足运算交换性,因此,$f(X\oplus Y)=f(Y\oplus X)$.于是,$d_f(v,u)=d_f(u,v)$;

(c) $d_f(v,u)$满足三角不等式.

要证明 $d_f(v,u)\leqslant d_f(v,a)+d_f(a,u)$,等价于证明 $f(X\oplus Z)\leqslant f(X\oplus Y)+f(Y\oplus Z)$.因为 $X\oplus Z\subseteq(X\oplus Y)\cup(Y\oplus Z)$,且 f 是单调非递减函数,我们有:

$$f(X\oplus Z)\leqslant f((X\oplus Y)\cup(Y\oplus Z)) \tag{3}$$

又因为 f 是非负权重和函数,因此对于任意节点集 A,B,C,均有 $f(A\cup B)\leqslant f(A)+f(B)$.于是:

$$f((X\oplus Y)\cup(Y\oplus Z))\leqslant f(X\oplus Y)+f(Y\oplus Z) \tag{4}$$

综合式(3)、式(4)可得:

$$f(X\oplus Z)\leqslant f(X\oplus Y)+f(Y\oplus Z) \tag{5}$$

于是,$d_f(v,u)\leqslant d_f(v,a)+d_f(a,u)$.

综上情形(a)~情形(c),$d_f(v,u)$是节点间的一种距离度量. □

基于定义 1、定理 1 的结论,定义图上节点之间的距离度量如下.

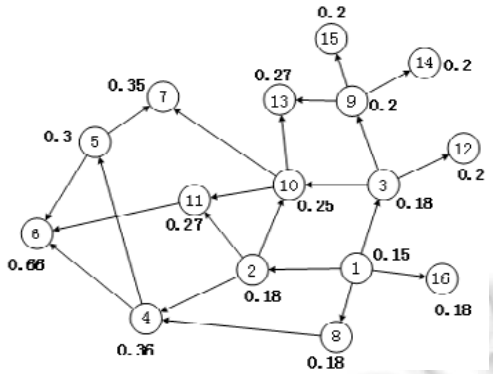
定义 3. 令 $\boldsymbol{r}$ 是个性化 PageRank 的排序值向量,$\boldsymbol{r}(v)\in\mathcal{R}^+$是 v 的 PPR 值, $f(A)=\sum_{v\in A}\boldsymbol{r}(v)$ 为节点集 A 的 PPR 值之和,并记 $R=f(V)$, 则可定义节点 v,u 之间的距离度量 $d_f(v,u)$如下:

$$d_f(v,u)=f(N(v)\oplus N(u))=\sum_{v\in N(v)\oplus N(u)}(\boldsymbol{r}(v)/R) \tag{6}$$

下面给出 $d_f(v,u)$的示例.图 2 给出一个节点带权有向图,设节点权值为 PPR 值,由公式(6)可得:

$$d_f(2,5)=\sum_{v\in N(2)\oplus N(5)}(\boldsymbol{r}(v)/R)=\sum_{\{2,4,5,6,7,10,11\}}(\boldsymbol{r}(v)/R)=0.574.$$

图 2 还给出了有向图上距离值最大的前 5 对节点及其距离值.



v	u	$d_f(v,u)$
4	10	0.599
2	5	0.574
4	9	0.537
1	4	0.533
1	5	0.53

Fig.2 Demostration of distance metric between nodes

图 2 节点之间距离度量计算示例

1.2 问题模型

多样性图排序问题需要定义相关性、多样性度量.先讨论相关性度量,然后基于第 1.1 节间节点的距离度量定义多样性度量,最后将多样性图排序建模为带权完全图上的优化问题.

首先讨论相关性度量.个性化 PageRank 是图数据上实现相关性查询的有效方法[3].具体地,令 G 是一个包含 n 个节点的图.节点的个性化 PageRank 排序向量可由如下迭代公式计算:

$$\boldsymbol{r}=\alpha\boldsymbol{A}'\boldsymbol{r}+(1-\alpha)\boldsymbol{q} \tag{7}$$

其中,$\alpha(0<\alpha<1)$为阻尼因子;$\boldsymbol{A}$ 是一个行规范化的图邻接矩阵,即 $\sum_{j=1}^{n}\boldsymbol{A}(i,j)=\mathbf{1},i=1,2,...,n$.给定查询向量 $\boldsymbol{q}(n\times1$ 的列向量 $\boldsymbol{q}(i)\geqslant0,\sum_{i=1}^{n}\boldsymbol{q}(i)=1)$,公式(7)会收敛到节点集 V 上的一个平稳分布,此分布即为排序值向量 $\boldsymbol{r}$.对于任意结点 $v\in V,\boldsymbol{r}(v)$度量了 v 与 $\boldsymbol{q}$ 的相似性.于是,给定一个排序结果 S,可用 S 的排序值之和,即 $rel(S)=\sum_{v\in S}\boldsymbol{r}(v)$ 度量相关性.

其次,如第 1.1 节讨论,$d_f(v,u)$给出了节点 v,u 间的一种距离度量.

于是,可用 S 中节点间的距离度量和 $div(S)=\sum_{v,u\in S}d_f(v,u)$ 度量其多样性.

给定查询向量 $\boldsymbol{q}$,令 $Q\subseteq V$ 是查询相关节点集,即$\forall v\in Q$,有 $\boldsymbol{r}(v)>0$.定义优化目标函数 $F(S)$如下:

$$F(S)=rel(S)+\lambda div(S)=\sum_{v\in S}\boldsymbol{r}(v)+\lambda\sum_{v,u\in S}d_f(v,u) \tag{8}$$

其中,$\lambda(0<\lambda<1)$是折中因子,用来折中相关性和多样性.如果$|S|=k$,注意到:公式(8)中,$rel(S)$为 k 项 $\boldsymbol{r}(v)$之和.同时,$div(S)$为 $k(k-1)/2$ 之和.为平衡 $rel(S)$与 $div(S)$,分别与不同的因子相乘,于是改写 $F(S)$为下面的公式(9):

$$F(S)=(k-1)\sum_{v\in S}\boldsymbol{r}(v)+2\lambda\sum_{v,u\in S}d_f(v,u) \tag{9}$$

于是,Topk 多样性图排序问题(TopkDRG)可定义为

$$S^*\leftarrow\arg\max_{S\subseteq Q\subseteq V,|S|=k}F(S) \tag{10}$$

下面证明由公式(10)定义的 TopkDRG 是 NP-hard 的.

定理 2. 令 $G=\langle V,E\rangle$是一个图,则图 G 上的 TopkDRG 问题是 NP-hard 的.

证明:在图 G 上给定查询向量 $\boldsymbol{q}$,Q 是 $\boldsymbol{q}$ 的查询相关节点集.引入如下融合节点相关性和不相似性的距离度量,记为 $d'_f(v,u)$.

$$d'_f(v,u)=\boldsymbol{r}(v)+\boldsymbol{r}(u)+2\lambda d_f(v,u) \tag{11}$$

不难验证, $d'_f(v,u)$ 仍然满足定义 2 中所述距离度量的非负、对称、三角不等式 3 条性质.

令 $D'_f(S)=\sum_{v,u\in S}d'_f(v,u)$ 为节点集 S 中所有节点对的度量距离和,由公式(11)可知:

$$D'_f(S)=\sum_{v,u\in S} d'_f(v,u)=(k-1)\sum_{v\in S}\boldsymbol{r}(v)+2\lambda\sum_{v,u\in S} d_f(v,u)=F(S).$$

也就是说,TopkDRG 的目标函数 $F(S)$等于节点集 S 中节点对的 $d'_f(v,u)$ 度量距离之和,即 TopkDRG 可描述为

$$S^*\leftarrow \arg\max_{S\subseteq Q\subseteq V,|S|=k} F(S)=\arg\max_{S\subseteq Q\subseteq V,|S|=k} D'_f(S) \tag{12}$$

至此,以 Q 集为节点集、以 $d'_f(v,u)$ 为 Q 中节点 v,u 之间的边权重构造了带权完全图 $C(Q)$.由公式(12)可知,TopkDRG 等价于在带权完全图 $C(Q)$上求解包含 k 个节点的最大边权和子图问题.该问题就是组合优化中的 max-sum k-dispersion 问题(简写为 MSkD)[13].MSkD 被证明是 NP-hard 的[14].因此,TopkDRG 是 NP-hard 的. □

由定理 2 结论可知,TopkDRG 是 NP-hard 的.因此,设计多项式时间近似算法成为本文后续要讨论的内容.

2 多样性图排序算法

首先给出 TopkDRG 的多项式时间求解算法,并分析该算法的近似性能.鉴于不同节点对的度量距离计算是相互独立的,进一步提出并行化的多样性图排序算法,并介绍算法在 Apache Spark 的并行图计算平台 GraphX 上的实现技术.

2.1 基于匹配的多样性图排序算法

由第 1.2 节可知,TopkDRG 归结为求解带权完全图 $C(Q)$上包含 k 个节点的最大权完全子图问题.虽然 TopkDRG 是 NP-hard 的,得益于 $C(Q)$中边权重满足距离度量性质,本节基于求解 $C(Q)$上最大权 $k/2$-匹配,提出一种 TopkDRG 的多项式时间近似算法.

算法的基本思想是:给定查询向量 $\boldsymbol{q}$,在图 G 上执行个性化 PageRank 获得查询相关节点集 Q.以 Q 为节点集,以 $d'_f(v,u)$ 为边 $e(v,u)$的边权重构造带权完全图 $C(Q)$.调用函数 *MWM* 求解 $C(Q)$的最大权 $k/2$-匹配,记为 M^* ($M^*=k/2$),M^*中所有边连接的 k 个节点即为多样性图排序的 top-k 排序结果.其中,函数 *MWM* 执行$\lfloor k/2\rfloor$次迭代,每次迭代时,选择 $C(Q)$中当前最大权重边加入到解集中,并删除该边的端点以及这些端点关联的边.下面,算法 1 以及函数 *MWM* 给出了上述过程的完整描述.

算法 1. MA(基于匹配的多样性图排序算法).

输入:图 $G=\langle V,E\rangle$,k,查询向量 $\boldsymbol{q}$,阻尼因子α,PPR 精度ε,折中因子λ,抽样概率 p;

输出:排序结果集 S.

(1) 给定查询向量 $\boldsymbol{q}$,执行个性化 PageRank 算法,获得查询相关节点集 Q:

$$Q\leftarrow personalizedPageRank(G,\boldsymbol{q},\alpha,\varepsilon);$$

(2) 以节点 PPR 值为权重,以 $p(0<p\leqslant 1)$作为抽样率,对 Q 进行随机抽样,得到随机子集:

$$Q_p\leftarrow ppr\_biased\_sampling(Q,p);$$

(3) 以 Q_p 为节点集,构造带权完全图 $k/2$,其中,任意边 $e(v,u)$的权重置为 $d'_f(v,u)$;

(4) $S\leftarrow\varnothing$;

(5) 调用 *MWM* 函数求解 $C(Q_p)$的最大权 $k/2$-匹配,即:$M^*\leftarrow MWM(C(Q_p),k)$;

(6) $S\leftarrow$包含在 M^*中所有边的端点;

(7) **RETURN** S

先分析算法 1 的时间复杂度.具体地,设 $G=\langle V,E\rangle$.算法 1 中第(1)行执行个性化 PageRank,其复杂度为 $O(|E|)$[8].第(3)行是算法 1 中计算代价最高的部分,代价主要由两部分构成:首先,收集节点的邻居节点信息,最坏情况下需要访问 G 中所有节点,代价为 $2|E|$;其次,构造 $C(Q_p)$,最坏情况下的时间复杂度为 $O(|V|^2)$.*MWM* 函数需要对所有节点对按距离度量值进行 1 次排序,其最坏情况下的复杂度为 $O(|V|^2\log(|V|^2))$.于是,算法 1 的时间复杂度为 $O(|E|+|V|^2\log(|V|^2))$.

需要说明的是:算法 1 中的第 2 行,得到查询相关节点集 Q 后,以 Q 中节点 PPR 值为权重,$p(0<p\leqslant 1)$作为抽样率,对 Q 进行随机抽样得到随机子集 Q_p,进而构造 $C(Q_p)$,并在其上进行求解.如果 $p=1.0$,即 $Q_p=Q$,如果 $p<1$,则

在 Q 的一个随机子团(randomsub-clique)上进行求解.下面先分析 p=1.0 的情况下,算法 1 求解 TopkDRG 的理论近似性能.对于 p<1 的情况,本文第 3 节将通过实验验证在稍微降低求解质量的情况下,算法的效率将得到很大提高.

函数 1. *MWM*(求解 $C(Q_p)$的最大权 k/2-匹配).

Input:$C(Q_p)$,k;

Output:最大权 k/2-匹配 M^*($|M^*|=k/2$).

(1) $M\leftarrow\varnothing$;

(2) **FOR** $i\leftarrow$1 to $\lfloor k/2\rfloor$ **DO**

(3) $(v,u)\leftarrow \arg\max_{x,y\in Q_p} d'_f(x,y)$;

(4) $M\leftarrow M\cup e(v,u)$;

(5) 删除 $C(Q_p)$中与 v,u 节点关联的边,删除 v,u;

(6) **END FOR**

(7) **RETURN** M

为分析算法 1 的近似性能,先证明一个基于匹配方法求解带权图的最大权重和 k 子图的一般性结论,即定理 3.首先引入相关数学记号,令 Q 是任意节点集,$C(Q)=\langle Q,E_Q\rangle$是 Q 上任意边权重满足距离度量性质的带权完全图,其任意一条边 $e(v,u)\in E_Q$ 的权重为某一距离度量 $dis(v,u)$.给定任意节点集 $A\subseteq Q$, $D(A)=\sum_{v,u\in A\subseteq Q} dis(v,u)$ 为 A 集中节点对的边权重和.类似地, $D(E)=\sum_{e(v,u)\in E\subseteq E_Q} dis(v,u)$ 为边集 E 中所有边的权重和.令 A^*($|A^*|=k$)是 $C(Q)$的含 k 个节点的最大权重和子图的节点集.M^*是 A^*决定的子图的最大权和 k/2 匹配边集.M^{OPT} 是 $C(Q)$的最大权重和 k/2 匹配边集.M^{OPT} 是 M^{OPT} 中的边决定的节点集.记 $\tilde{M}$ 是 M^{OPT} 的α-近似解. $\tilde{A}$ 是 $\tilde{M}$ 决定的节点集.

定理 3 描述了这样一个事实:通过求解 $C(Q)$的最大权重和 k/2 匹配的α-近似解 $\tilde{M}$, $\tilde{A}$ 是 $\tilde{M}$ 决定的含 k 个节点的节点集,那么对于边权重和这一度量标准而言, $\tilde{A}$ 是 A^*的 2α-近似解.换句话说,定理 3 给出了基于匹配方法求解带权图上含 k 个节点的最大权重和子图的一般性近似结论.

定理 3. 对于任意一个边权重满足距离度量性质的带权完全图 $C(Q)$,总有 $D(A^*)\leqslant 2\alpha D(\tilde{A})$.

证明:

首先,令 A($|A|=k>2$)是任意节点集,M 表示 A 决定的带权完全子图 $C(A)$上的最大权和 k/2 匹配边集, $\bar{D}(A)$ 表示 $C(A)$中边的平均权重值, $\bar{D}(M)$ 表示匹配边集 M 的平均权重值.根据文献[13]可知:

$$\bar{D}(A)\leqslant\bar{D}(M)\Leftrightarrow D(A)\leqslant(k-1)D(M) \tag{13}$$

其次,由于 $\bar{D}(A^*)\leqslant\bar{D}(M^*)$,进而有:

$$\frac{D(A^*)}{k(k-1)/2}\leqslant\frac{D(M^*)}{k/2}\Leftrightarrow D(A^*)\leqslant(k-1)D(M^*) \tag{14}$$

此外,由于 M^{OPT} 是 $C(Q)$上的最大权重和 k/2 匹配边集,因此 $D(M^*)\leqslant D(M^{OPT})$,进而基于公式(14)的结论,有:

$$D(A^*)\leqslant(k-1)D(M^*)\leqslant(k-1)D(M^{OPT}) \tag{15}$$

于是,根据文献[13]可知:

$$\frac{\bar{D}(M)}{2}<\bar{D}(A)\Leftrightarrow\frac{D(M)}{k/2}<\frac{D(A)}{k(k-1)/2}\Leftrightarrow D(M)<(2/k-1)D(A) \tag{16}$$

因为 $D(\tilde{M})<(2/k-1)D(\tilde{A})$,同时, $\tilde{M}$ 是 M^{OPT} 的α-近似解,即 $D(M^{OPT})\leqslant\alpha D(\tilde{M})$,综合公式(13)~公式(16)的结果可知:

$$D(A^*)\leqslant(k-1)D(M^*)\leqslant(k-1)D(M^{OPT})\leqslant(k-1)\alpha D(\tilde{M})\leqslant 2\alpha D(\tilde{A}) \tag{17}$$

证毕. □

基于定理 3 给出的结论,可分析算法 1 求解 TopkDRG 问题的近似性能.在算法 1 中,对于 p=1.0 的情况,我们采用函数 *MWM* 求解 $C(Q)$的最大权 k/2 匹配边集.由文献[15]的结论可知,基于 *MWM* 可得到α=1-近似的 $C(Q)$的最大权 k/2 匹配边集.于是,作为定理 3 结论的一个推论,推论 1 给出了算法 1 求解 TopkDRG 问题的近似性能:

推论 1. 当 p=1.0 时,如果算法 1 采用 *MWM* 求解 $C(Q)$的最大权 $k/2$ 匹配边集,那么基于定理 3 的结论,算法 1 可得到 TopkDRG 问题的 $2\alpha(\alpha=1)$-近似解.

2.2 并行化的多样性图排序算法

构造 $C(Q)$是算法 1 计算代价最高的部分,其关键在于计算 Q 中节点对的度量距离值.幸运的是,不同节点对的距离计算是相互独立的.基于此,提出并行化的多样性图排序算法 PMA.PMA 与 MA 最大的不同在于:PMA 采用 MapReduce 编程模型完成 $C(Q)$上 $O(|Q|^2)$的节点对度量距离计算,从而提高了多样性图排序算法的效率.由于 PMA 与 MA 其余部分相同,不再给出 PMA 的算法描述.

从以下 3 个方面介绍基于 MapReduce 编程模型[16]构造 $C(Q)$具体方法.

(1) 键值对(key-value pairs)生成

设由个性化 PageRank 获得查询相关节点集 Q,其中,$v\in Q$,$v.ppr>0$ 为节点 v 的 PPR 值.基于 Spark GraphX 中的 *aggregateMessages* 函数,并行化地收集 Q 所有节点的邻居信息.对任意节点 v,其邻居集信息为 $v.Nbrs$.由此生成节点 RDD,内容格式:$(v,v.ppr,v.Nbrs)$.对此,RDD 做 cartesian 运算(笛卡尔积),生成键值对集,其中,任意键值对为

$$\textbf{\textit{key}}:(v,u),\textbf{\textit{value}}:(v.ppr,u.ppr,v.Nbrs,u.Nbrs) \tag{18}$$

(2) Map 函数

实现 v,u 节点邻居集的对称差运算,即,将公式(18)所示键值对映射为如下键值对:

$$\textbf{\textit{key}}:(v,u),\textbf{\textit{value}}:(v.ppr,u.ppr,v.Nbrs\oplus u.Nbrs) \tag{19}$$

(3) Reduce 函数

对每个键值对,根据公式(6)完成 v,u 之间的距离计算.

基于上述 MapReduce 编程模型,可并行计算 Q 上节点对的距离度量计算,从而 $C(Q)$构造.然后,调用 *MWM* 函数即可输出 top-k 排序结果.

本文基于 Apache Spark[12]的并行化图数据计算组件 GraphX 实现了 PMA 算法.弹性分布式数据集(resilient distributed dataset,简称 RDD)是 Spark 上分布式数据的基本抽象.在 Spark 平台上,图数据也以 RDD 的方式进行存储.基于 GraphX 的 PMA 算法也是通过对图 RDD 数据的一系列变换(transformation)以及行动(action)而得以实现的.

图 4 给出了图 3 所示基于 GraphX 的 PMA 算法实现所涉及到的 RDD 以及施加在这些 RDD 上的一系列的变换和行动,其中,PPR 查询的输入节点为节点 1.从步骤 1~步骤 5 完成节点对的度量距离计算.步骤 6 的循环过程求解 Topk 多样性排序节点集.

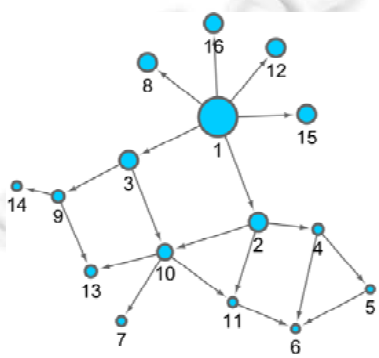


Fig.3 A graph used in the demonstration of PMA algorithm

图 3 PMA 算法示例用图

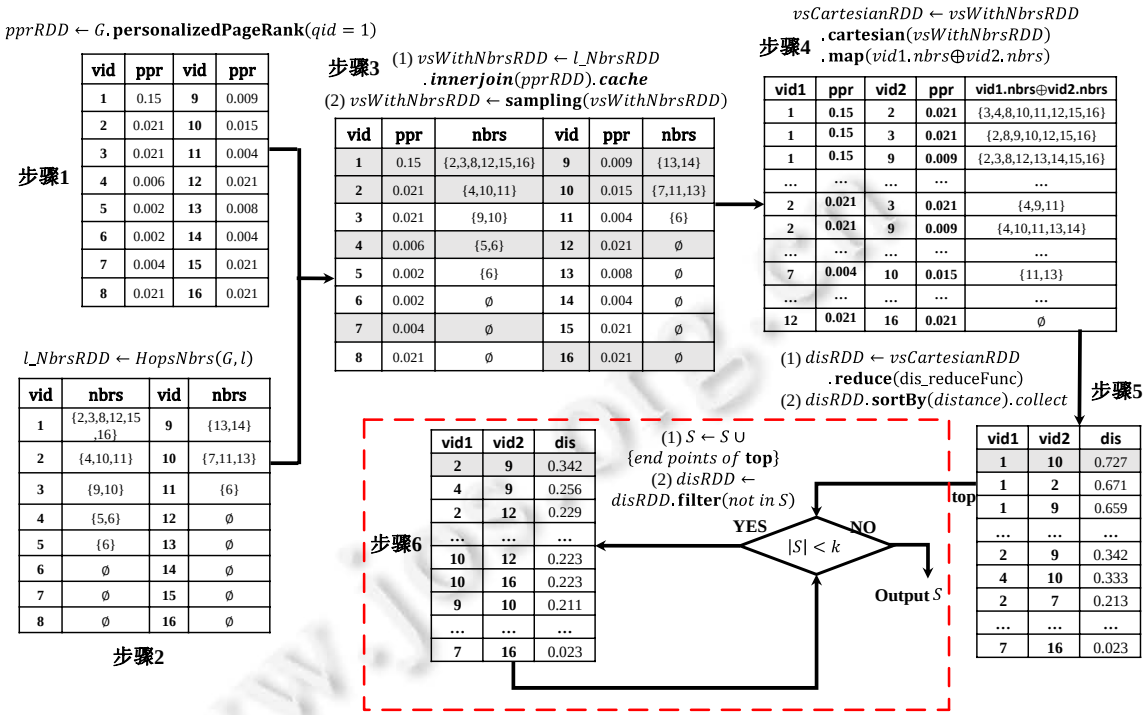


Fig.4 A demonstration of transformations and actions of RDDs in PMA algorithm

图 4 PMA 算法中 RDD 变换和行动示例

3 实 验

3.1 实验环境设置

(1) 实验图数据集

在表 1 所示 4 个真实的、不同类型的图数据集(均下载自 http://snap.stanford.edu/)上测试了本文算法.其中,ca-AstroPh 是论文作者合作关系网络,soc-Epinions1 是社交网站 Epinions.com 内用户之间的信任关系网络,amazon0601 是 Amazon 商品共购关系网,Web-Google 是来自 Google 的 Web 页面关系网.

Table 1 Information of experimental graph datasets

表 1 图数据信息

名称	节点数	边数	图直径	平均聚类系数
ca-AstroPh	18 772	198 110	14	0.630 6
soc-Epinions1	75 879	508 837	14	0.137 8
amazon0601	403 394	3 387 388	21	0.417 7
Web-Google	875 713	5 105 039	21	0.514 3

(2) 度量标准

采用文献[8]定义的查询相关性(relevance)作为查询相关性度量标准,记为 *rel* 并定义如下:

$$rel(S)=\frac{\sum_{v\in S}\boldsymbol{r}(v)}{\sum_{v\in A}\boldsymbol{r}(v)} \tag{20}$$

其中,$\boldsymbol{r}(v)$是节点 v 的 PPR 值,S 是 PMA 算法返回的 top-k 排序结果,A 是个性化 PageRank 算法返回的 top-k 排序结果.可见,$rel(S)$值越大,S 的查询相关性越高.

至今,排序结果多样性没有统一度量标准.为便于对比,本文采用文献[9]提出的扩展相关性(expansion

relevance)作为多样性度量标准,记为 *epRel* 并定义如下.

定义 4[9]. 令 S 是任意节点集,$N_l(S)$是 S 的 l-步邻居节点集,即:

$$N_l(S)=S\cup\{v\in(V-S)|\exists v\in S,d(v,u)\leqslant l\}.$$

则 S 的 l-扩展相关性定义为

$$epRel_l(S)=\sum_{v\in N_l(S)}\boldsymbol{r}(v) \tag{21}$$

epRel 融合了 S 的查询相关性和扩展率,*epRel* 越大,则在此度量标准下的排序结果多样性程度越高.

此外,本文通过定义节点间距离来度量节点间的不相似性,因此引入了 2 种新的多样性度量标准,定义如下.

定义 5. 令 S 是包含 k 个节点的节点集,$d_f(v,u)$是本文定义的节点 v,u 之间的距离度量,那么,S 的平均距离度量定义为

$$aveDis(S)=\sum_{v,u\in S}d_f(v,u)\Big/(k(k-1)/2) \tag{22}$$

不难看出:*aveDis* 越大,S 内部节点之间的差异度越大,S 多样性程度越高.

定义 6. 令 S 是包含 k 个节点的节点集,$d_f(v,u)$是本文定义的节点 v,u 之间的距离度量,那么,S 中节点间的最小距离度量记为

$$\min Dis(S)=\min_{v,u\in S}d_f(v,u) \tag{23}$$

易知:min*Dis*(*S*)越大,S 内部节点之间的差异度越大,S 多样性程度越高.

(3) 参与对比的其他图排序算法

第 1 种算法是 PPR,即个性化 PageRank 算法.本文个性化 PageRank 采用 Apache GraphX 提供的 API 实现.

第 2 种是文献[9]提出的基于子模函数优化 *epRel* 的算法,记为 SM.根据文献[9]的结论,*epRel* 是以节点 PPR 值为权重的带权扩展率,由于 *epRel* 融合了查询相关性和扩展率,对于排序结果的多样性度量更具优越性[9].因此,本文与文献[9]提出的算法进行实验对比.SM 是一种子模目标函数的贪心算法,我们采用文献[17]提出的 CELF 方法对贪心过程进行加速以提高算法执行效率.

(4) 实验环境

PMA,SM 算法均采用 scala 语言实现.所有实验在一台 32G 内存,2×12 核(2 个 Intel Xeon 2.66 GHz CPU)的 Linux 14.4 服务器上完成.Apache Spark 采用 2.0 版本.

3.2 参数对算法性能的影响

通过实验验证算法 1 中的参数对 PMA 算法性能的影响.具体包括以下方面.

(1) λ因子对 PMA 算法的影响

由公式(8),λ用来折中优化目标中相关性和多样性.实验验证不同的λ对于 PMA 算法性能的影响.通过调整个性化 PageRank 中的ε参数,使得$|Q|$制在 2000~3000 之间.每次实验时,随机给定查对询节点,分别针对 λ=0.1,0.3,0.5,0.7,1.0 这 5 种情况查询 top-k=30 排序结果.最终结果是 50 次实验的平均值.

图 5 给出了在 4 个测试数据集上,PMA 算法的 *rel*,*epRel*,*aveDis*,min*Dis* 这 4 个度量标准与λ参数的关系图.如图 5(a)所示:随着λ值的增大,进行距离计算时,多样性部分的权重增大,而相关性权重降低,因此,*rel* 逐渐降低,而其他多样性指标得到增强.图 5(b)~图 5(d)给出的实验结果验证了这一事实,*epRel*,*aveDis*,min*Dis* 这 3 个多样性指标随λ提高而增大.由于λ大于 0.5 后 *rel* 显著降低,因此在后续的算法对比实验中,PMA 算法的λ均设为 0.5.

(2) 抽样率 p 对 PMA 算法的影响

如第 2.1 节所述,PMA 算法可以对查询相关节点集 Q 以 PPR 值为权重进行随机抽样得到随机节点子集 Q_p,进而在 Q_p 形成的子团上执行 PMA 算法求解.实验验证不同的 p 对于 PMA 算法性能的影响.为有效检验参数 p 的影响,进一步降低了个性化 PageRank 算法中的ε参数,提高了查询相关集的节点数,使得$|Q|$在 4000~ 5000 之间.每次实验时,随机给定查询节点,分别对 p=0.1,0.3,0.5,0.7,1.0 这 5 种情况查询多样性图排序 top-k=30 排序结果.最终结果是 50 次实验的平均值.

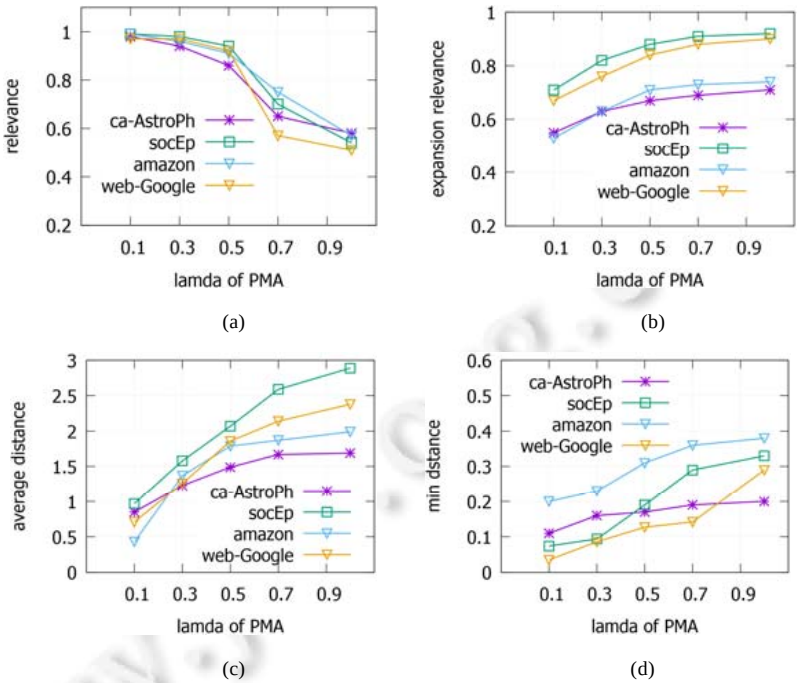


Fig.5 Effects of λ to the performance of PMA

图 5 折中因子λ对 PMA 算法的影响

图 6 给出了抽样率 p 对 PMA 算法影响的实验结果.

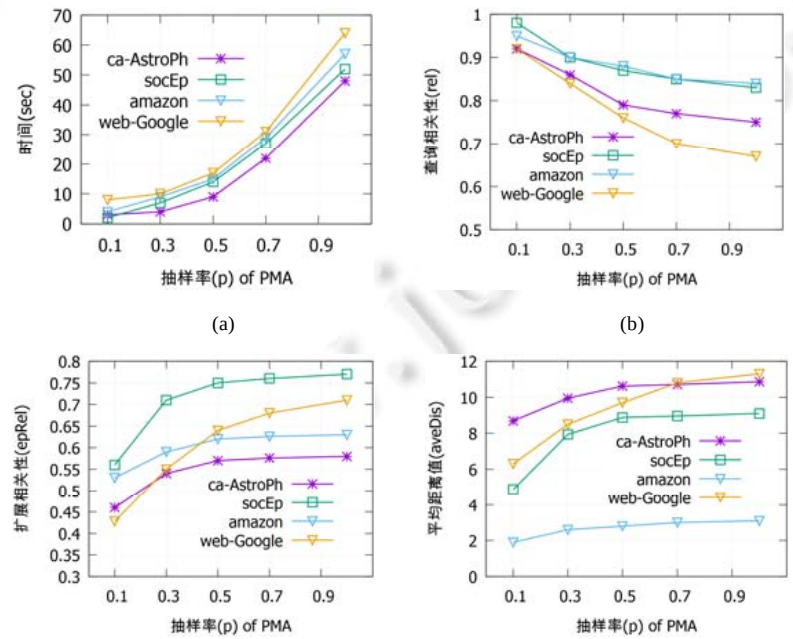


Fig.6 Effects of p to the performance of PMA

图 6 抽样率 p 对 PMA 算法的影响

如图 6(a)所示:随着 p 的增加,Q_p 节点数也随之增加,构造 $C(Q_p)$子团所需的计算距离对数目大幅增加(需 $O(|Q_p|^2)$距离计算),PMA 算法时间也随之上升.从图 6(b)可见:当 p<0.5 时,由于按节点的 PPR 值为权重进行抽样,高相关性节点得以高概率选中进入 Q_p 集,最终排序结果具有高相关性.随着 p 的增加,可在更大范围内选择节点,因此 *rel* 降低.但与此同时,距离度量值更大的节点对会进入最终的排序结果,因此如图 6(c)、图 6(d)所示,*epRel* 和 *aveDis* 这两个多样性指标也随 p 的增大而增加.同时注意到:当 p>0.5 后,除 web-Google 外,在其余几个数据集上, *epRel*,*aveDis* 指标增长趋势减缓.这意味着增加 Q 的节点数并不能显著提升多样性指标.

从实验结果可知:采取对查询相关节点集进行随机抽样时,在保证排序结果的查询相关性和多样性的前提下,可大幅降低算法的执行时间.这一特性使得 PMA 算法能够高效完成大规模图数据的多样性图排序任务.

(3) CPU 核数对 PMA 执行之间的影响

PMA 算法基于 ApacheSpark 的并行图计算平台 GraphX 实现,可通过设置 Spark 的并行核数来调整 PMA 算法的效率.Q_p 中节点对的距离计算是 PMA 算法中计算密集部分,恰好也是适于并行计算的部分.实验验证了核数分别在 4,8,12,16,20,24 的情况下,PMA 算法查询 top-k=30 的执行时间.实验结果如图 7 所示.

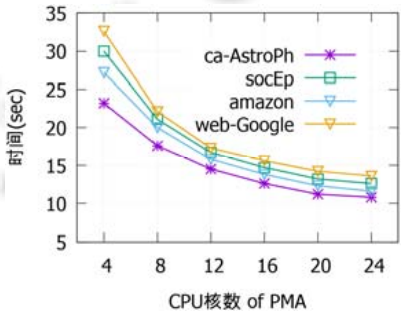


Fig.7 Effects of CPU Core number to time of PMA

图 7 CPU 核数对 PMA 执行时间的影响

由实验结果可知:随着核数的增加,PMA 算法的执行时间减少.这一结果也验证了距离计算的可并行性.后续的算法对比实验中,PMA 的 CPU 核数均设为 24.

3.3 算法对比实验结果

我们比较了 PPR,SM,PMA,rPMA(以 PPR 权重进行节点随机抽样的 PMA 算法)这 4 种算法的执行时间、相关性以及多样性指标.实验参数设置为:|Q|=2000~3000,λ=0.5,p=0.5.每次实验随机给定查询节点,分别得到 k=10,20,30,50,100 的 top-k 排序结果.最终结果是 50 次重复实验的平均值.

表 2 给出了 SM,PMA,rPMA 这 3 种算法的执行时间比较结果.

- 首先,PMA 和 rPMA 执行时间明显优于 SM.特别地,在完成同样的 top-k 排序时,rPMA 比 SM 快 5 倍~10 倍,且数据集越大,速度优势越明显;
- 其次,由于 SM 算法是迭代过程,随着 k 值增加,其执行时间也显著增加.PMA 与 rPMA 中无计算密集的迭代过程,其执行时间随 k 值增加趋势平缓.

Table 2 Time comparisons of different diversified ranking algorithms (s)

表 2 不同算法执行时间比较 (s)

k	ca-AstroPh			socEp			Amazon			Web-Google		
	SM	PMA	rPMA	SM	PMA	rPMA	SM	PMA	rPMA	SM	PMA	rPMA
10	4.6	5.8	**2.1**	11.2	14.1	**2.7**	5.8	5.4	**2.3**	7.6	8.8	**3.6**
20	6	5.6	**2.7**	14.2	14.6	**3.2**	12.2	6.4	**4.3**	17	10.6	**6.6**
30	12.8	10.2	**3.2**	18	15	**3.6**	16.4	10.2	**4.6**	23.4	11.2	**6**
50	19.8	13.2	**3.8**	27.2	17.6	**4.6**	25	12	**5.6**	43	12.4	**8.6**
100	50	16	**6.8**	81	24	**8.5**	58.3	17.3	**10**	104	24	**16**

表 3 是 SM、PMA 以及 rPMA 在 *rel* 指标上的比较结果.

Table 3 Relevance (*rel*) comparisons of algorithms

表 3 不同算法的查询相关性比较(*rel*)

k	ca-AstroPh			socEp			Amazon			Web-Google		
	SM	PMA	rPMA	SM	PMA	rPMA	SM	PMA	rPMA	SM	PMA	rPMA
10	0.962	0.935	0.95	0.99	0.94	0.95	0.84	0.85	0.87	0.94	0.98	0.97
20	0.849	0.92	0.94	0.95	0.93	0.94	0.84	0.86	0.87	0.92	0.94	0.96
30	0.825	0.89	0.92	0.91	0.92	0.94	0.83	0.9	0.9	0.89	0.92	0.94
50	0.81	0.874	0.9	0.88	0.9	0.92	0.81	0.92	0.93	0.88	0.91	0.92
100	0.8	0.84	0.88	0.82	0.88	0.9	0.79	0.93	0.94	0.87	0.89	0.9

由于 rPMA 以 PPR 权重进行节点随机抽样构造子团,高 PPR 值的节点被高概率选入 Q_p 集,其 *rel* 指标优于 SM,PMA.在所有测试数据集上,PMA 的 *rel* 指标稍优于 SM.

表 4 是 *epRel* 指标的比较结果.

Table 4 Expansion relevance (*epRel*) comparisons of algorithms

表 4 不同算法的扩展相关性比较(*epRel*)

k	ca-AstroPh				socEp				Amazon				Web-Google			
	PPR	SM	PMA	rPMA	PPR	SM	PMA	rPMA	PPR	SM	PMA	rPMA	PPR	SM	PMA	rPMA
10	0.3	0.33	**0.413**	0.38	0.48	0.51	**0.56**	0.51	0.41	0.47	0.45	0.42	0.5	0.59	**0.73**	0.7
20	0.44	0.52	**0.59**	0.54	0.68	0.71	**0.82**	0.77	0.48	0.58	0.54	0.51	0.58	0.71	**0.82**	0.78
30	0.52	0.62	**0.7**	0.68	0.73	0.79	**0.89**	0.83	0.57	0.7	0.69	0.65	0.67	0.76	**0.85**	0.82
50	0.74	0.84	**0.83**	0.79	0.85	0.93	**0.95**	0.89	0.63	0.75	0.74	0.69	0.73	0.85	**0.88**	0.85
100	0.81	0.9	**0.94**	0.87	0.89	0.97	**0.95**	0.9	0.75	0.85	0.82	0.78	0.8	0.9	**0.93**	0.88

SM,PMA 以及 rPMA 增强了排序结果的多样性,这 3 种算法的 *epRel* 指标明显优于 PPR.值得注意的是:虽然 PMA 算法并非直接优化 *epRel*,但在参与测试的 ca-AstroPh,socEp 以及 web-Google 这 3 个图数据上,其 *epRel* 指标仍优于以 *epRel* 为优化目标的 SM.此外,由于 PMA 和 rPMA 直接优化 *aveDis*,由表 5、表 6 的实验结果可见,PMA 和 rPMA 算法在 *aveDis*、min*Dis* 指标下明显优于 PPR 和 SM.

Table 5 Average distance (*aveDis*) comparisons of algorithms

表 5 不同算法的平均距离值比较(*aveDis*)

k	ca-AstroPh				socEp				Amazon				Web-Google			
	PPR	SM	PMA	sPMA	PPR	SM	PMA	sPMA	PPR	SM	PMA	sPMA	PPR	SM	PMA	sPMA
10	0.73	0.81	**1.2**	1.1	1.27	1.4	**1.87**	1.67	0.86	0.58	**1.7**	1.4	0.79	0.65	**1.45**	1.2
20	1.03	1.4	**1.53**	1.4	1.2	1.23	**2.01**	1.78	0.88	0.51	**1.63**	1.28	0.68	0.52	**2.1**	1.9
30	0.83	1.16	**1.51**	1.35	1.34	1.37	**2.05**	1.81	0.89	0.61	**2.1**	1.56	0.71	0.5	**1.89**	1.78
50	1.49	1.73	**2.27**	1.91	1.26	1.34	**1.99**	1.76	0.86	0.43	**1.41**	1.2	0.39	0.28	**1.7**	1.4
100	0.98	0.97	**1.54**	1.3	1.24	1.34	**1.89**	1.71	0.44	0.24	**0.91**	0.72	0.13	0.23	**0.92**	0.79

Table 6 Minimum distance (min*Dis*) comparisons of algorithms

表 6 不同算法的最小距离值比较(min*Dis*)

k	ca-AstroPh				socEp				Amazon				Web-Google			
	PPR	SM	PMA	rPMA	PPR	SM	PMA	rPMA	PPR	SM	PMA	rPMA	PPR	SM	PMA	rPMA
10	0.31	0.3	0.39	0.35	0.18	0.23	0.25	0.21	0.06	0.06	0.17	0.16	0.19	0.03	0.58	0.54
20	0.17	0.23	0.25	0.23	0.07	0.1	0.17	0.16	0.02	0.02	0.06	0.06	0.04	0.02	0.38	0.4
30	0.07	0.13	0.16	0.14	0.06	0.067	0.15	0.13	0.02	0.02	0.04	0.03	0.03	0.008	0.15	0.12
50	0.05	0.08	0.07	0.06	0.02	0.03	0.06	0.07	0.01	0.01	0.03	0.03	0.01	0.004	0.11	0.09
100	0.03	0.03	0.04	0.03	0.02	0.001	0.04	0.03	0.01	0.01	0.02	0.02	0.01	0.001	0.08	0.08

综上,在进行多样性图排序时,PMA 和 rPMA 在保证查询结果的相关性和多样性的前提下,通过并行计算和随机抽样,大幅提高了算法的执行效率.相较于 SM,在查询质量和查询效率上均有优势.

4 总 结

在用户真实的查询意图难以准确获取的情况下,为提供高质量的图排序结果并提高用户查询的满意度,能

够有效折中相关性和多样性的图排序算法是图数据检索面临的研究挑战.

针对已有的研究工作在排序结果多样性建模和算法效率这两方面存在的不足之处,本文提出了一种描述节点间不相似度的的距离度量,以此为基础,建立了新的多样性度量标准,并将多样性图排序建模为一种带权完全图上的组合优化问题.给出了求解此问题的 2-近似算法以及该算法在 MapReduce 编程模型上的并行化实现方法.在真实的图数据上测试了本文方法,实验结果表明,本文方法在算法执行时间、查询相关性和多样性指标上均优于已有方法.

本文方法并未涉及节点或边的信息,在很多实际的图数据检索应用中,节点和边往往带有丰富的属性信息[18],如何将本文方法拓展到面向属性图(attributed graph)的多样性图排序中,这是我们下一步可研究的工作.

References:

[1] Cheng XQ, Sun BJ, Shen HW, Yu ZH. Research status and trends of diversified graph ranking. Bulletin of Chinese Academy of Science, 2015,30(2):248−256 (in Chinese with English abstract). [doi: 10.16418/j.issn.1000-3045.2015.02.012]

[2] Page L, Brin S, Motwani R, *et al*. The PageRank citation ranking: Bringing order to the Web. Stanford InfoLab, 1999.

[3] Haveliwala TH. Topic-Sensitive pagerank. In: Proc. of the 11th Int'l Conf. on World Wide Web. ACM Press, 2002. 517−526.

[4] Mei Q, Guo J, Radev D. DivRank: The interplay of prestige and diversity in information networks. In: Proc. of the 16th Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2010. 1009−1018.

[5] Zhu X, Goldberg AB, Van Gael J, Andrzejewski D. Improving diversity in ranking using absorbing random walks. In: Proc. of the Human Language Technology Conf. of the North American Chapter of the Association of Computational Linguistics. the Association for Computational Linguistics, 2007. 97−104.

[6] Zheng K, Wang H, Qi Z, Li JZ, Gao H. A survey of query result diversification. Knowledge & Information Systems, 2017,51:1−36. [doi: 10.1007/s10115-016-0990-4]

[7] Tong H, He J, Wen Z, Konuru R, Lin CY. Diversified ranking on large graphs: An optimization viewpoint. In: Proc. of the 17th Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2011. 1028−1036.

[8] Li RH, Yu JX. Scalable diversified ranking on large graphs. IEEE Trans. on Knowledge and Data Engineering, 2013,25(9): 2133−2146. [doi: 10.1109/TKDE.2012.170]

[9] Küçüktunç O, Saule E, Kaya K, Çatalyürek ÜV. Diversified recommendation on graphs: Pitfalls, measures, and algorithms. In: Proc. of the 22nd Int'l Conf. on World Wide Web. ACM Press, 2013. 715−726.

[10] Küçüktunç O, Saule E, Kaya K, Çatalyürek ÜV. Diversifying citation recommendations. ACM Trans. on Intelligent Systems and Technology (TIST), 2015,5(4):55. [doi: 10.1145/2668106]

[11] Buchbinder N, Feldman M, Naor J, Schwartz R. Submodular maximization with cardinality constraints. In: Proc. of the 25th Annual Symp. on Discrete Algorithms. SIAM, 2014. 1433−1452. [doi: 10.1137/1.9781611973402.106]

[12] Apache Spark—Lightning-fast cluster computing. http://spark.apache.org/

[13] Ravi SS, Rosenkrantzt DJ, Tayi GK. Approximation algorithms for facility dispersion. Gonzalez TF, ed. Handbook of Approximation Algorithms and Metaheuristics. Chapman & Hall/CRC, 2007. 38.1−38.17. [doi: 10.1201/9781420010749]

[14] Hassin R, Rubinstein S, Tamir A. Approximation algorithms for maximum dispersion. Operations Research Letters, 1997,21(3): 133−137. [doi: 10.1016/S0167-6377(97)00034-5]

[15] Gollapudi S, Sharma A. An axiomatic approach for result diversification. In: Proc. of the 18th Int'l Conf. on World Wide Web. ACM Press, 2009. 381−390.

[16] Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. In: Proc. of the 6th Symp. on Operating System Design and Implementation. USENIX Association, 2004. 137−150.

[17] Leskovec J, Krause A, Guestrin C, Faloutsos C, Van Briesen J, Glance N. Cost-Effective outbreak detection in networks. In: Proc. of the 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2007. 420−429. [doi: 10.1145/1281192.1281239]

[18] Muller E, Sanchez PI, Mulle Y, Böhm K. Ranking outlier nodes in subspaces of attributed graphs. In: Proc. of the 29th Int'l Conf. on Data Engineering, IEEE. 2013. 216−222. [doi: 10.1109/ICDEW.2013.6547453]

附中文参考文献:

[1] 程学旗,孙冰杰,沈华伟,余智华.多样性图排序的研究现状及展望.中国科学院院刊,2015,30(2):248−256. [doi: 10.16418/j.issn.1000-3045.2015.02.012]

李劲(1975－),男,云南大理人,博士,副教授,CCF 专业会员主要研究领域为数据与知识工程.

岳昆(1979－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据与知识工程.

蔡娇(1992－),女,硕士生,主要研究领域为数据与知识工程.

张志坚(1980－),男,讲师,主要研究领域为数据与知识工程.

刘惟一(1950－),男,教授,博士生导师,CCF 高级会员,主要研究领域为数据与知识工程.