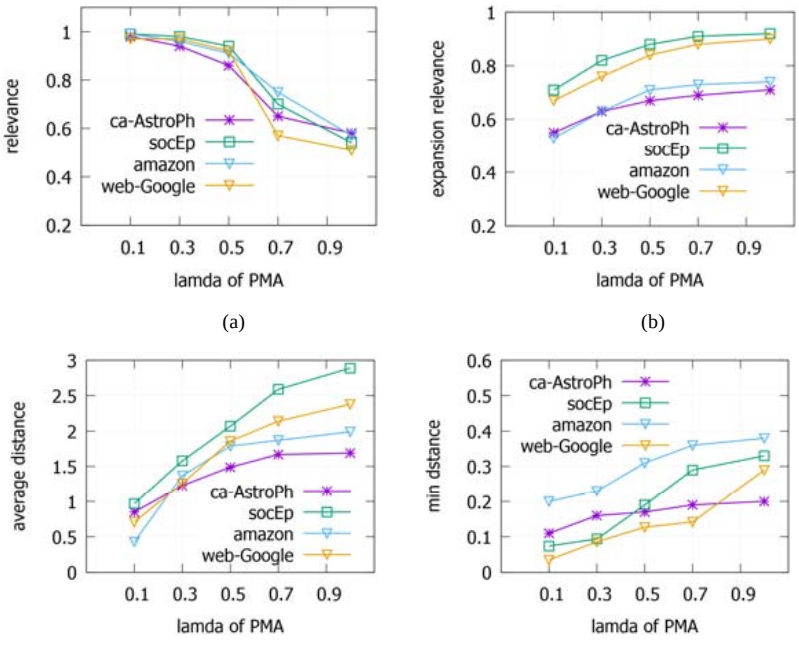


Fig.5 Effects of $\lambda$ to the performance of PMA

图 5 折中因子$\lambda$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.

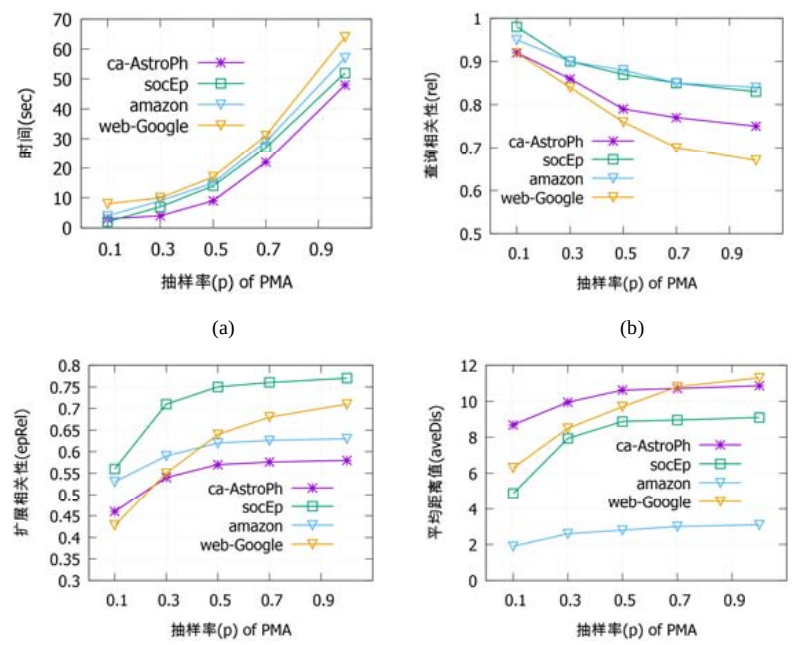


Fig.6 Effects of $p$ to the performance of PMA

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.

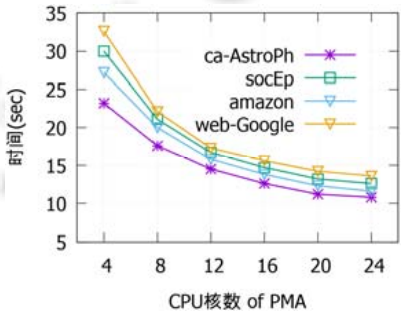


Fig.7 Effects of CPU Core number to time of PMA

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,$\lambda$=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.