于中间水平,使用这两个测试数据集的主要目的是测试算法的查全率和查准率.DBLP[①]的规模很大,节点数超过 1 亿,我们选择的测试数据仅仅是其中的一部分,但也保持了一定的规模,该数据集主要测试在数据规模较大情况下各算法的时间性能.本实验没有考虑数据集中 ID/IDREF 的关联情况.

Table 4 Data sets

表 4 测试数据集的基本信息

数据集	大小(KB)	节点总数	不同标签数	平均深度	最大深度
SIGMODRECORD	467	15 263	12	4.60	6
Mondial	1 743	69 846	50	3.59	5
DBLP	39 782	2 050 096	50	2.9	6

针对每个测试数据集,我们选取了具有不同查询意图的测试查询,并且每个查询测试的查询意图是明确的.为每个数据集设计了 7 个测试查询,S1~S7、M1~M7 和 D1~D7 分别表示针对 SIGMODRECORD、Mondial 和 DBLP 数据集的测试查询.具体见表 5.

Table 5 Keyword query case

表 5 关键字查询案例

数据集	查询编号	关键字	满足要求的 LCA 总数
SIGMODRECORD	S1	author,Anthony,article	9
	S2	David,Randy,article	3
	S3	Randy,Data	4
	S4	author,title,article	1 504
	S5	volume,27,article	4
	S6	System,initpage,endPage	285
	S7	initPage,47,article	9
Mondial	M1	Singapore,country	1
	M2	United States,border	1
	M3	AFG,border,country	6
	M4	United,Nations,organization	24
	M5	Moutain,Sweden	4
	M6	Russion,lake	8
	M7	desert,Syrian,country	1
DBLP	D1	Frank,Michael,article	15
	D2	2013,Springer,Net	50
	D3	ITC ,Machael,inproceedings	8
	D4	688,2011 Security	13
	D5	phdthesis,information 2011	119
	D6	California,university,2014	169
	D7	url Vassiliou author	7

5.2 查询质量

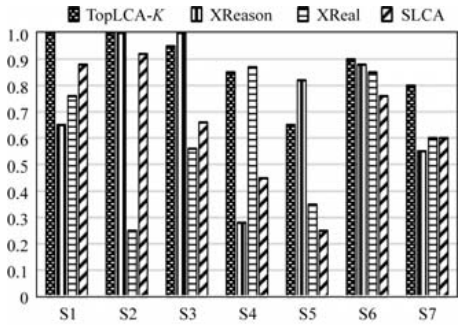
把本文提出的 TopLCA-*K* 算法与已有 XReason、Xreal 和 SLCA 等方法在查准率和查全率进行对比来比较它们的查询质量.

图 5 显示了本文提出的 TopLCA-*K* 与其他方法在查询准确率方面的比较情况.

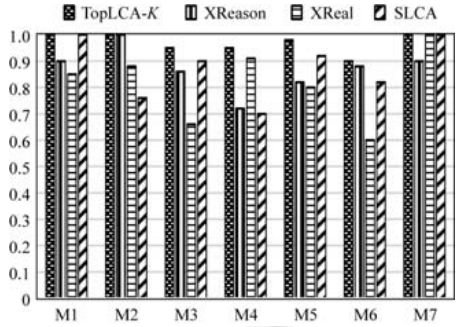
图 5 显示,在查准率方面,一般情况下,TopLCA-*K* 要高于其他 3 种方法,因为该算法考虑了 LCA 中横向关键字密度和纵向关键字密度两方面的情况,因此排在前面的更符合用户查询意图.但在某些查询情况下,准确率也不能达到 100%,主要是因为查询关键字存在一些歧义,例如关于 DBLP 数据集的 D4 查询,查询意图是查询 688 页,2011 年出版的标题中含有 Security 的论文、书籍或者会议出版的文章,但是由于 688 不仅仅出现在 page 中,而且在节点 ee 中也出现了该关键字.同理,2011 不仅仅出现在 year 节点中,而且在 incollection 的 url 中也出现了 2011.

图 6 显示了召回率的比较情况.

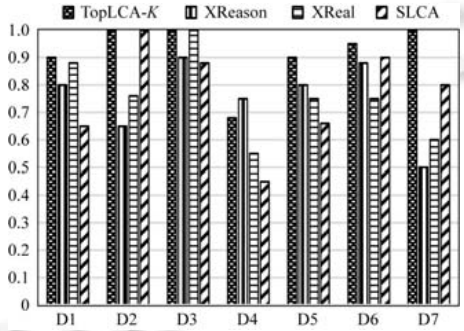
很明显,TopLCA-*K* 的召回率为 100%,不存在漏检问题,其他方法都不能达到 100%,XReal 采用扩展网页排名方法进行排名,在关键字歧义情况下,很容易漏检;SLCA 由于去掉了父 LCA,显然会丢失一些符合要求的结果;而 TopLCA-*K* 返回了所有的 LCA,包括有歧义的 LCA,然后从深度和广度计算 LCA 的值,把排名小的优先返回,当 K 值合理时,将返回所有可能满足查询意图的 LCA.



(a) SIGMODRECORD 数据集上的查准率



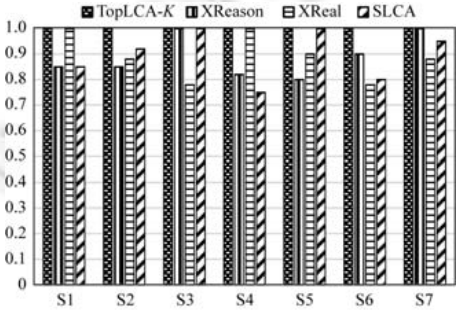
(b) Mondial 数据集上的查准率



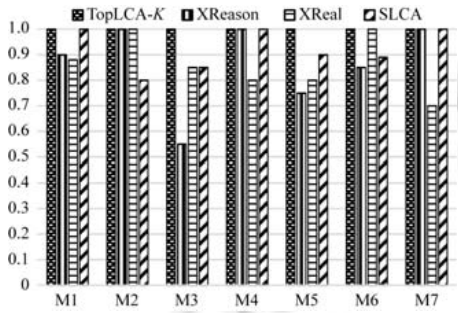
(c) DBLP 数据集上的查准率

Fig.5 Precision rate

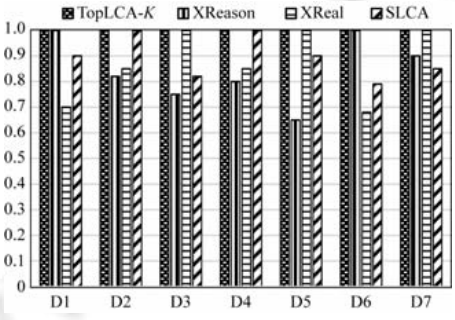
图 5 比较查准率



(a) SIGMODRECORD 数据集上的召回率



(b) Mondial 数据集上的召回率



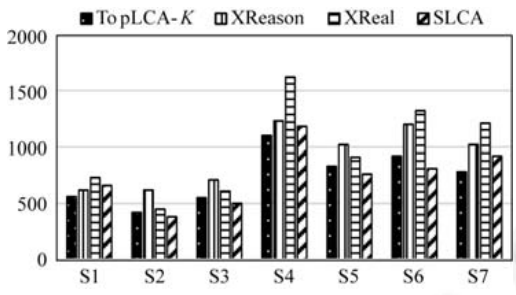
(c) DBLP 数据集上的召回率

Fig.6 Recall rate

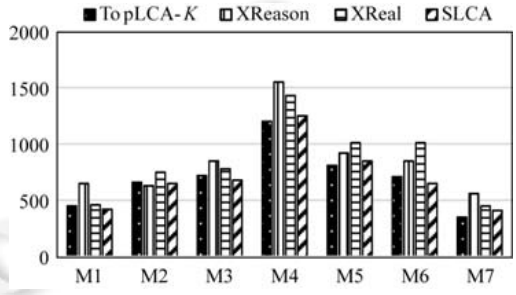
图 6 比较召回率

5.3 查询性能

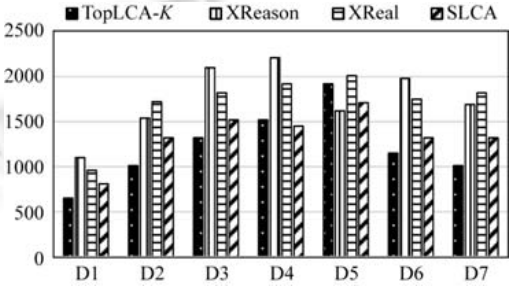
图 7 显示了 3 个数据集上算法运行时间情况,为了准确记录查询时间,每个关键字查询执行 5 次,取它们的平均值作为查询时间,根据规则 3 设置了 TopLCA-*K* 中 *K* 的值.从图 7 可以发现,TopLCA-*K* 的算法时间性能明显优于 XReason 和 XReal,与 SLCA 比较接近.因为 TopLCA-*K* 算法利用 CI 能够对查询空间树进行有效剪枝,例如在 D6 中,共有 8 849 个节点包含关键字 2014,其中 year 节点值有 4 051 个,其他节点,如 ee 包含的 2014 有 520 个.这种情况对 XReason 产生有效结构模式形成了很大干扰,这种情况也对采用 TF*IDF 排名的 XReal 方法形成了很多干扰信息.



(a) SIGMODRECORD 数据集上的算法时间性能



(b) Mondial 数据集上的算法时间性能



(c) DBLP 数据集上的算法时间性能

Fig.7 Time performance

图 7 算法时间性能对比

6 结束语

本文首先介绍了 LCA 过滤语义和结果排名方法,指出了在 XML 关键字查询中 LCA 过滤语义存在漏报问题,提出了用户查询意图与查询关键字在纵向和横向方面的两个规则,建立了利用边密度和路径密度对 LCA 节点进行评分的公式,采取中位节点索引 CI 来提高 TopLCA-*K* 算法效率.实验结果表明,本文提出的对 LCA 进行评分排名的方法在查准率和召回率方面效果较好,并且查询时间性能也较好,但需要进一步优化提高.下一步的研究重点考虑当在 LCA 之间存在包含、重复和交叉关系情况时,如何对 LCA 进行排序以及结果展示的问题,同时进一步优化算法.在未来的工作中,将研究如何减少编码长度以及基于新编码方案的 XML 关键字查询处理.

References:

[1] Guo L, Shao F, Botev C, Shanmugasundaram J. XRANK: Ranked keyword search over XML documents. In: Proc. of the SIGMOD Conf. 2003. 16–27.

[2] Schmidt A, Kersten M, Windhouwer M. Querying XML documents made easy: Nearest concept queries. In: Proc. of the Int'l Conf. on Data Engineering. IEEE, 2001. 321–329.

[3] Zhang CJ, Wang XL, Zhou AY. XML filtering based-on probabilistic SLCA. Chinese Journal of Computers, 2014,(9):1959–1971 (in Chinese with English abstract).

[4] Bao Z, Ling TW, Chen B, *et al*. Effective XML keyword search with relevance oriented ranking. In: Proc. of the IEEE Int'l Conf. on Data Engineering. IEEE, 2009. 517–528.

[5] Bao Z, Lu J, Ling TW, *et al*. Towards an effective XML keyword search. IEEE Trans. on Knowledge & Data Engineering, 2010, 22(8):1077–1092.

[6] Chen LJ, Papakonstantinou Y. Supporting top-*K* keyword search in XML databases. In: Proc. of the IEEE Int'l Conf. on Data Engineering. IEEE, 2010. 689–700.

[7] Liu Z, Chen Y. Processing keyword search on XML: A survey. World Wide Web-Internet & Web Information Systems, 2011, 14(5-6):671–707.

[8] Liu X, Wan C, Chen L. Returning clustered results for keyword search on XML documents. IEEE Trans. on Knowledge & Data Engineering, 2011,23(12):1811–1825.

[9] Cohen S, Mamou J, Kanza Y, Sagiv Y. XSEarch: A semantic search engine for XML. VLDB Journal, 2003, 45–56.

[10] Li Y, Yu C, Jagadish HV. Schema-free XQuery. VLDB Journal, 2004, 72–83.

[11] Xu Y, Papakonstantinou Y. Efficient keyword search for smallest LCAs in XML databases. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. Baltimore: DBLP, 2005. 527–538.

[12] Li G, Feng J, Wang J, *et al*. Effective keyword search for valuable lcas over XML documents. In: Proc. of the 16th ACM Conf. on Information and Knowledge Management. ACM, 2007. 31–40.

[13] Chen ZY, Wang X, Tang X. Efficiently computing RKN for keyword queries on XML data. Journal on Communications, 2014, 35(7):46–55 (in Chinese with English abstract).

[14] Termehchy A, Winslett M. Using structural information in XML keyword search effectively. ACM Trans. on Database Systems, 2011,36(1):1–39.

[15] Li G, Li C, Feng J, *et al*. SAIL: Structure-aware indexing for effective and progressive top-*k*, keyword search over XML documents. Information Sciences, 2009,179(21):3745–3762.

[16] Li J, Liu C, Zhou R, *et al*. Top-*k* keyword search over probabilistic XML data. In: Proc. of the IEEE Int'l Conf. on Data Engineering. IEEE, 2011. 673–684.

[17] Amer-Yahia S, Lalmas M. XML search: Languages, INEX and scoring. ACM SIGMOD Record, 2006,35(4):16–23.

[18] Hristidis V, Papakonstantinou Y, Balmin A. Keyword proximity search on XML graphs. In: Proc. of the Int'l Conf. on Data Engineering. IEEE, 2003. 367–378.

[19] Zhou J, Bao Z, Ling TW, *et al*. MCN: A new semantics towards effective XML keyword search. Lecture Notes in Computer Science, 2009,5463:511–526.

[20] Zhou R, Liu C, Li J. Fast ELCA computation for keyword queries on XML data. In: Proc. of the Int'l Conf. on Extending Database Technology. ACM, 2010. 549–560.

[21] Zhou J, Wang W, Chen Z, *et al*. Top-down XML keyword query processing. IEEE Trans. on Knowledge & Data Engineering, 2016, 28(5):1340–1353.

[22] Dimitriou A, Theodoratos D, Sellis T. Top-*k*-size keyword search on tree structured data. Information Systems, 2014,(47):178–193.

[23] Hristidis V, Papakonstantinou Y. DISCOVER: Keyword search in relational databases. VLDB Journal, 2002,26(2):670–681.

[24] Li QS, Wang QY, Wang S. Query understanding for XML keyword search. Ruan Jian Xue Bao/Journal of Software, 2012,23(8): 2002–2017 (in Chinese with English abstract). http://www.org.cn/1000-9825/4122.htm [doi: 10.3724/SP.J.1001.2012.04122]

[25] He H, Wang H, Yang J, *et al*. BLINKS: Ranked keyword searches on graphs. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. Beijing: ACM, 2007. 305–316.

[26] Aksoy C, Dimitriou A, Theodoratos D, *et al*. XReason: A semantic approach that reasons with patterns to answer XML keyword queries. In: Database Systems for Advanced Applications. Berlin, Heidelberg: Springer-Verlag, 2013. 299–314.

[27] Aksoy C, Dimitriou A, Theodoratos D. Reasoning with patterns to effectively answer XML keyword queries. VLDB Journal, 2015, 24(3):441–465.

[28] Li J, Liu C, Zhou R, *et al*. Suggestion of promising result types for XML keyword search. In: Proc. of the Int'l Conf. on Extending Database Technology. ACM, 2010. 561–572.

[29] Zhou J, Bao Z, Wang W, *et al*. Efficient query processing for XML keyword queries based on the IDList index. VLDB Journal, 2014,23(1):25–50.

[30] Liu X, Wan C, Liu D. Keyword query with structure: towards semantic scoring of XML search results. Information Technology & Management, 2016,17(2):151–163.

[31] Nguyen K, Cao J. Top-*k*, answers for XML keyword queries. World Wide Web-Internet & Web Information Systems, 2012,15(5-6): 485–515.

[32] Liu Z, Chen Y. Identifying meaningful return information for XML keyword search. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. ACM, 2007. 329–340.

附中文参考文献:

[3] 张晨静,王晓玲,周傲英.基于概率 SLCA 的 XML 过滤.计算机学报,2014,(9):1959–1971.

[13] 陈子阳,王璿,汤显.面向 XML 关键字查询的高效 RKN 求解策略.通信学报,2014,35(7):46–55.

[24] 李求实,王秋月,王珊.XML 关键词检索的查询理解.软件学报,2012,23(8):2002–2017. http://www.org.cn/1000-9825/4122.htm [doi: 10.3724/SP.J.1001.2012.04122]

覃遵跃(1974－),男,湖南张家界人,博士,副教授,主要研究领域为数据库技术.

徐洪智(1974－),男,副教授,主要研究领域为嵌入式系统,并行计算.

汤庸(1964－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为协同工作,数据库.

黄云(1976－),男,博士,副教授,CCF 专业会员,主要研究领域为数据挖掘,智能信息计算.