法索引做了压缩,但是其失去了 GT 方法的遍历优势,所以非常耗时,甚至超过了 BF 方法.由于 TG 和 TcG 方法具有对候选集的良好过滤能力,所以二者的查询时间基本维持在 3ms 左右.此实验证明了 5 种方法在不同数目的关键词下执行查询的可扩展性.

最后验证 BF,GT,GcT,TG 和 TcG 这 5 种方法在不同集群节点(1,2,4,8 个节点)上执行同一种约束查询的可扩展性.如图 7(c)所示,随着节点数目的增加,BF 方法的查询时间由 690ms 减少到 244ms,性能提高了将近 3 倍;GT 方法的查询时间基本维持在 65ms~70ms 范围内;GcT 方法的查询时间由 1 347ms 减少到 216ms,性能提高了 6 倍以上;TG 方法的查询时间由 21ms 减少到 5ms,性能提高了 4 倍以上;TcG 方法的查询时间由 21ms 减少到 3ms,性能提高了 7 倍.此实验说明,BF,GT,GcT,TG 和 TcG 这 5 种方法随着节点的增加,都具备有良好的可扩展性.

5 相关工作

本节主要回顾和分析与约束型查询有关的文献,主要分 3 类:(1) 基于定量和定性测度的子空间聚类;(2) 基于查询的双聚类;(3) 约束型双聚类.对每一类工作只做简单介绍,更详细的描述可查阅 Sim 等人的综述[44−48].

1) 基于定量和定性测度的子空间聚类

双聚类的概念最初由 Hartigan 等人[2]提出,作为对矩阵中的行与列同时聚类的一种方法,并将其命名为 Direct 聚类.Cheng 等人[3]提出了基因表达数据的双聚类,并引入了元素残差以及子矩阵的均方残差 MSR[3]的概念.他们提出了一种贪婪方法.首先,将整个数据矩阵作为初始化数据;接着,删除元素残差或者均方残差最大元素或者行列,依次递归下去,直到剩余矩阵的 MSR 低于某个阈值;然后,增加部分元素或者行列,保证所得矩阵的 MSR 也低于该阈值.该方法效率较低,因为一次只能挖掘 1 个双聚类.Ben-Dor[1]介绍了一种特殊的双聚类模型 OPSM,并证明了其是 NP 难问题.随后,研究者们提出了基于定量测度[4−14]和定性测度[15−33]的 OPSM 挖掘方法.定量测度包括均方残差 MSR[1]、平均相关值 ACV[10]、平均斯皮尔曼秩相关系数 ASR[10]、平均一致性相关指数 ACSI[10]等.定性测度包括上升、下降和无变化[24,28].

(1) 基于定量测度[4−14]的 OPSM 挖掘方法

基于 Cheng 等人[2]提出的δ-bicluster 模型,Yang 等人[4]为了减少数据缺失值的影响,给出一种δ-cluster 模型.Cho 等人[5]介绍了两种与 MSR 相似的平方残差测度,同时提出了两种有效的基于 k-means 的双聚类算法.Divina 等人[6]给出了一种基于进化计算的双向聚类方法,用来发现尺寸较大、重叠较少且 MSR 小于某阈值的双聚类.Deodhar 等人[7]提出了一种鲁棒的有重叠的双聚类方法 ROCC,能够有效地从大量的含有噪声的数据中挖掘出稠密的、任意位置的有覆盖的聚类.Cho 等人[8]给出了数据转换的方法,以解决现有的平方残差和测度方法只能有效地挖掘出在数值上具有偏移的双聚类,却不能很好地解决在数值上有缩放的双聚类问题.Odibat 等人[9]发现,现有方法并不能有效地挖掘矩阵数据中任意位置有重叠的双聚类,提出了确定性双聚类算法.该算法可以有效地发现正负相关的任意位置上有重叠的双聚类.Ayadi 等人[10]利用平均一致性相关指数 ACSI 来评估相干双聚类,并利用有向无环图组建这些双聚类.Truong 等人[11]提出了一种算法,用来生成若干个覆盖度小于阈值的双聚类,在一定程度上能够发现无冗余的双聚类.Ayadi 等人[12]给出了模因双聚类算法 MBA,以发现生物学意义上的重要的负相关双聚类.Chen 等人[13]利用最小均方错误 MMSE 测度来鉴别所有类型的线性模式(偏移、缩放、偏移与缩放联合模式).Denitto 等人[14]利用 Max-Sum 测度来提升双聚类的质量.

(2) 基于定性测度[15−33]的 OPSM 挖掘方法

Wang 等人[15]提出并设计了基于 pScore 测度和 pCluster 模型的方法,以挖掘具有相似升降趋势的模式.Liu 等人[16]通过寻找在部分维度下表达值排序相同的基因等对象来挖掘 OPSM.Wang 等人[17]给出了一种基于最近邻的新的测度方法来挖掘相似模式.Kriegel 等人[18]提出了一种局部密度阈值的 OPSM 挖掘方法,试图改变现有的基于全局密度阈值的方法并不能适用于每个 OPSM 的现状.Jiang 等人[19]给出了一种质量驱动的 top-k 模式挖掘方法,以提升发现的有重叠的 OPSM 的质量.Gao 等人[20,29]提出了一种 KiWi 框架,利用 k 和 w 两个参数来约束计算资源和搜索空间.Zhang 等人[21]发现,现有的方法都假设基因表达数据是同质的,给出了称为 F-cluster 的模型来挖掘异质数据中的相干模式.Yan 等人[22]为挖掘非线性相关的模式,设计了适用于时序基因表达数据

的联合聚类方法 MI-TSB.Zhang 等人[23]提出了一种近似保序聚类模型 AOPC,以减少数据中噪声的影响.Chui 等人和 Yip 等人[24,30]利用多份数据模型 OPSM-RM 来消除数据噪声的影响.Zhao 等人[25]提出了一种最大化子空间聚类算法,来挖掘具有正相关和负相关的共调控基因聚类.Trapp 等人[26]为挖掘最优 OPSM,给出了一种基于线性规划的挖掘方法.Fang 等人[27]为挖掘放松的 OPSM,提出了包含以行或列为中心的 OPSM-Growth 方法.随后,Fang 等人[28,32]提出了基于桶和概率的方法,挖掘出放松的 OPSM.An 等人[31]利用互信息和核密度进行双聚类.Cho 等人[33]给出了一种基于坏字符规则的 KMP 算法,试图快速匹配保序模式.

2) 基于查询的双聚类[49]

该方法来自生物信息领域[50−53],应用对象是基因表达数据.首先,由用户根据经验来提供功能相关或共表达的种子基因;接着,利用该种子对搜索空间剪枝或双聚类进行指导.为了使现有挖掘方法能利用先验知识并回答指定的问题,Dhollander 等人[54]提出了基于贝叶斯的查询驱动的双聚类方法 QDB.同时,给出了一种基于实验条件列表的联合方法,以实现关键词的多样性并免除必须事先定义阈值等问题.随后,Zhao 等人[55]对 QDB 方法进行了改进,提出了 ProBic 方法.虽然二者在概念上相似,但是也有不同之处.QDB 方法利用概率关系模型扩展贝叶斯框架,并用基于期望最大化的直接指定方法来学习该概率模型.Alqadah 等人[34]提出了一种利用低方差和形式概念分析优势相组合的方法,以发现在部分实验条件下具有相同表达趋势的基因.为了便于 OPSM 的查询,Jiang 等人[38]提出了带有行列表头的前缀树索引 pIndex,同时给出 4 种 OPSM 查询方法.

3) 约束型双聚类

目前,该问题的相关研究相对较少,是一种对挖掘与分析基因表达数据的新方法.Pensa 等人[35,36]提出了一种从局部到整体的方法来建立间隔约束的二分分区.该方法是通过扩展从 0/1 数据集中提取出来的一些局部模式来实现的.其基本思想是:首先,将间隔约束转换成一个放松的局部模式;然后,利用 k 均值算法来获得一个局部模式的分区;最后,对上述分区做后续处理来确定数据之上的协同聚类结构.随后,Pensa 等人[37]对文献[35,36]进行了扩展,主要的不同点有:(i) 作者同时在行列之上应用目标函数来评价双聚类的好坏;(ii) 新工作[37]将文献[35,36]中的数据从 0/1 矩阵扩展到了实数数据;(iii) 提升了 must-link 与 cannot-link 两类约束在行列之上的处理性能.Tseng 等人[56]提出了基于相关约束完整链接的约束型双聚类方法.

6 结 论

针对基因表达数据中保序子矩阵的约束查询问题,提出了适用于不同情形的索引和查询方法,即提出了基于数字签名和 Trie 的基因序列与实验条件序列索引方法,以及自顶向下与自底向上相结合的查询方法.索引方法极大地减小了索引的数据量并提升了索引速度;查询方法方便了不同类型的查询目的,且提升了查询性能.在真实数据集上的实验结果证明了所提出的方法具有很好的查询精确性和良好的可扩展性.

致谢 在此,向百忙之中评审我们论文的专家学者和对本文的工作给予支持与建议的同行表示感谢.

References:

[1] Ben-Dor A, Chor B, Karp R, Yakhini Z. Discovering local structure in gene expression data: The order-preserving submatrix problem. In: Proc. of the 6th Annual Int'l Conf. on Computational Biology (RECOMB). ACM Press, 2002. 49−57. [doi: 10.1145/565196.565203]

[2] Hartigan JA. Direct clustering of a data matrix. Journal of the American Statistical Association, 1972,67(337):123−129. [doi: 10.1080/01621459.1972.10481214]

[3] Cheng Y, Church GM. Biclustering of expression data. In: Proc. of the 8th Int'l Conf. on Intelligent Systems for Molecular Biology (ISMB). AAAI Press, 2000. 93−103.

[4] Yang J, Wang W, Wang H, Yu PS. δ-Clusters: Capturing subspace correlation in a large data sets. In: Proc. of the 18th Int'l Conf. on Data Engineering (ICDE). IEEE Press, 2002. 517−528. [doi: 10.1109/ICDE.2002.994771]

[5] Cho H, Dhillon IS, Guan Y, Sra S. Minimum sum-squared residue co-clustering of gene expression data. In: Proc. of the 12th SIAM Int'l Conf. on Data Mining (SDM). SIAM Press, 2004. 114−125. [doi: 10.1137/1.9781611972740.11]

[6] Divina F, Aguilar-Ruiz JS. Biclustering of expression data with evolutionary computation. IEEE Trans. on Knowledge and Data Engineering, 2006,18(5):590−602. [doi: 10.1109/TKDE.2006.74]

[7] Deodhar M, Gupta G, Ghosh J, Cho H, Dhillon IS. A scalable framework for discovering coherent co-clusters in noisy data. In: Proc. of the 26th Annual Int'l Conf. on Machine Learning (ICML). ACM Press, 2009. 241−248. [doi: 10.1145/1553374.155 3405]

[8] Cho H. Data transformation for sum squared residue. In: Proc. of the 14th Pacific-Asia Conf. on Advances in Knowledge Discovery and Data Mining (PAKDD). Berlin, Heidelberg: Springer-Verlag, 2010. 48−55. [doi: 10.1007/978-3-642-13657-3_8]

[9] Odibat O, Reddy CK. A generalized framework for mining arbitrarily positioned overlapping co-clusters. In: Proc. of the 19th SIAM Int'l Conf. on Data Mining (SDM). SIAM Press, 2011. 343−354. [doi: 10.1137/1.9781611972818.30]

[10] Ayadi W, Elloumi M, Hao JK. BicFinder: A biclustering algorithm for microarray data analysis. Knowledge and Information Systems, 2012,30(2):341−358. [doi: 10.1007/s10115-011-0383-7]

[11] Truong DT, Battiti R, Brunato M. Discovering non-redundant overlapping biclusters on gene expression data. In: Proc. of the 13th IEEE Int'l Conf. on Data Mining (ICDM). IEEE Press, 2013. 747−756. [doi: 10.1109/ICDM.2013.36]

[12] Ayadi W, Hao JK. A memetic algorithm for discovering negative correlation biclusters of DNA microarray data. Neurocomputing, 2014,145:14−22. [doi: 10.1016/j.neucom.2014.05.074]

[13] Chen S, Liu J, Zeng T. MMSE: A generalized coherence measure for identifying linear patterns. In: Proc. of the IEEE Int'l Conf. on Bioinformatics and Biomedicine (BIBM). IEEE Press, 2014. 489−492. [doi: 10.1109/BIBM.2014.6999206]

[14] Denitto M, Farinelli A, Bicego M. Biclustering gene expressions using factor graphs and the max-sum algorithm. In: Proc. of the 24th Int'l Conf. on Artificial Intelligence (IJCAI). AAAI Press, 2015. 925−931.

[15] Wang H, Wang W, Yang J, Yu PS. Clustering by pattern similarity in large data sets. In: Proc. of the 28th ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD). ACM Press, 2002. 394−405. [doi: 10.1145/564691.564737]

[16] Liu J, Wang W. OP-Clustering by tendency in high dimensional space. In: Proc. of the 3th IEEE Int'l Conf. on Data Mining (ICDM). IEEE Press, 2003. 187−194. [doi: 10.1109/ICDM.2003.1250919]

[17] Wang H, Pei J, Yu PS. Pattern-Based similarity search for microarray data. In: Proc. of the 11th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD). ACM Press, 2005. 814−819. [doi: 10.1145/1081870.1081978]

[18] Kriegel HP, Kroger P, Renz M, Wurst SHR. A generic framework for efficient subspace clustering of high-dimensional data. In: Proc. of the 5th IEEE Int'l Conf. on Data Mining (ICDM). IEEE Press, 2005. 250−257. [doi: 10.1109/ICDM.2005.5]

[19] Jiang D, Pei J, Zang A. A general approach to mining quality pattern-based clusters from expression data. In: Proc. of the 10th Int'l Conf. on Database Systems for Advanced Applications (DASFAA). Springer-Verlag, 2005. 188−200. [doi: 10.1007/11408079_18]

[20] Gao BJ, Griffith OL, Ester M, Jones SJM. Discovering significant OPSM subspace clusters in massive gene expression data. In: Proc. of the 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD). ACM Press, 2006. 922−928. [doi: 10.1145/1150402.1150529]

[21] Zhang X, Wang W. An efficient algorithm for mining coherent patterns from heterogeneous microarrays. In: Proc. of the 19th Int'l Conf. on Scientific and Statistical Database Management (SSDBM). IEEE Computer Society Press, 2007. 32. [doi: 10.1109/SSDBM.2007.30]

[22] Yan L, Sun Z, Wu Y, Zhang B. Biclustering nonlinearly correlated time series gene expression data. Journal of Computer Research and Development, 2008,45(11):1865−1873 (in Chinese with English abstract).

[23] Zhang M, Wang W, Liu J. Mining approximate order preserving clusters in the presence of noise. In: Proc. of the 24th Int'l Conf. on Data Engineering (ICDE). IEEE Press, 2008. 160−168. [doi: 10.1109/ICDE.2008.4497424]

[24] Chui CK, Kao B, Yip KY, Lee SD. Mining order-preserving submatrices from data with repeated measurements. In: Proc. of the 8th IEEE Int'l Conf. on Data Mining (ICDM). IEEE Press, 2008. 133−142. [doi: 10.1109/ICDM.2008.12]

[25] Zhao Y, Yu J, Wang G, Chen L, Wang B, Yu G. Maximal subspace coregulated gene clustering. IEEE Trans. on Knowledge and Data Engineering, 2008,20(1):83−98. [doi: 10.1109/TKDE.2007.190670]

[26] Trapp AC, Prokopyev OA. Solving the order-preserving submatrix problem via integer programming. INFORMS Journal on Computing, 2010,22(3):387−400. [doi: 10.1287/ijoc.1090.0358]

[27] Fang Q, Ng W, Feng J. Discovering significant relaxed order-preserving submatrices. In: Proc. of the 16th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2010. 433−442. [doi: 10.1145/1835804.1835861]

[28] Fang Q, Ng W, Feng J, Li Y. Mining bucket order-preserving submatrices in gene expression data. IEEE Trans. on Knowledge and Data Engineering, 2012,24(12):2218−2231. [doi: 10.1109/TKDE.2011.180]

[29] Gao BJ, Griffith OL, Ester M, Xiong H, Zhao Q, Jones SJM. On the deep order-preserving submatrix problem: A best effort approach. IEEE Trans. on Knowledge and Data Engineering, 2012,24(2):309−325. [doi: 10.1109/TKDE.2010.244]

[30] Yip KY, Kao B, Zhu X, Chui CK, Lee SD, Cheung DW. Mining order-preserving submatrices from data with repeated measurements. IEEE Trans. on Knowledge and Data Engineering, 2013,25(7):1587−1600. [doi: 10.1109/TKDE.2011.167]

[31] An P. Research on biclustering methods for gene expression data analysis [MS. Thesis]. Suzhou: Soochow University, 2013 (in Chinese with English abstract).

[32] Fang Q, Ng W, Feng J, Li Y. Mining order-preserving submatrices from probabilistic matrices. ACM Trans. on Database Systems, 2014,39(1):No.6. [doi: 10.1145/2533712]

[33] Cho S, Na JC, Park K, Sim JS. A fast algorithm for order-preserving pattern matching. Information Processing Letters, 2015,115(2): 397−402. [doi: 10.1016/j.ipl.2014.10.018]

[34] Alqadah F, Bader JS, Anand R, Reddy CK. Query-Based biclustering using formal concept analysis. In: Proc. of the 12th SIAM Int'l Conf. on Data Mining (SDM). SIAM Press, 2012. 648−659. [doi: 10.1137/1.9781611972825.56]

[35] Pensa RG, Robardet C, Boulicaut JF. Towards constrained co-clustering in ordered 0/1 data sets. In: Proc. of the 16th Int'l Symp. on Methodologies for Intelligent Systems (ISMIS). Springer-Verlag, 2006. 425−434. [doi: 10.1007/11875604_49]

[36] Pensa RG, Robardet C, Boulicaut JF. Constraint-Driven co-clustering of 0/1 data. In: Proc. of the Constrained Clustering: Advances in Algorithms, Data Mining and Knowledge Discovery Series. 2008. 123−148. [doi: 10.1201/9781584889977.ch6]

[37] Pensa RG, Boulicaut JF. Constrained co-clustering of gene expression data. In: Proc. of the 8th SIAM Int'l Conf. on Data Mining (SDM). SIAM Press, 2008. 25−36. [doi: 10.1137/1.9781611972788.3]

[38] Jiang T, Li Z, Chen Q, Li K, Wang Z, Pan W. Towards order-preserving submatrix search and indexing. In: Proc. of the 20th Int'l Conf. on Database Systems for Advanced Applications (DASFAA). Part II. Berlin, Heidelberg: Springer-Verlag, 2015. 309−326. [doi: 10.1007/978-3-319-18123-3_19]

[39] Helmer S, Moerkotte G. Evaluation of main memory join algorithm for joins with subset join predicates. In: Proc. of the 23rd Int'l Conf. on Very Large Database (VLDB). ACM Press, 1997. 386−395.

[40] KiWi Software 1.0. 2012. http://www.bcgsc.ca/platform/bioinfo/ge/kiwi/

[41] Broad Institute. Datasets.rar and 5q_gct_file.gct. 1999. http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi

[42] Jiang T, Li Z, Chen Q, Wang Z, Li K, Wang Z. Parallel partitioning and mining gene expression data with butterfly network. In: Proc. of the 24th Int'l Conf. on Database and Expert Systems Applications (DEXA). Part I. Berlin, Heidelberg: Springer-Verlag, 2013. 129−144. [doi: 10.1007/978-3-642-40285-2_13]

[43] Jiang T, Li Z, Chen Q, Wang Z, Li K, Pan W. OMEGA: An order-preserving submatrix mining, indexing and search. In: Proc. of the European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD). Part III. Berlin, Heidelberg: Springer-Verlag, 2015. 303−307. [doi: 10.1007/978-3-319-23461-8_35]

[44] Sim K, Gopalkrishnan V, Zimek A, Cong G. A survey on enhanced subspace clustering. Data Mining and Knowledge Discovery, 2013,26:332−397. [doi: 10.1007/s10618-012-0258-x]

[45] Kriegel HP, Kroger P, Zimek A. Clustering of high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Trans. on Knowledge Discovery from Data, 2009,3(1):1−58. [doi: 10.1145/1497577.1497578]

[46] Yue F, Sun L, Wang K, Wang Y, Zuo W. State-of-the-Art of cluster analysis of gene expression data. Acta Automatica Sinica, 2008,34(2):113−120 (in Chinese with English abstract). [doi: 10.3724/SP.J.1004.2008.00113]

[47] Jiang D, Tang C, Zhang AD. Cluster analysis for gene expression data: A survey. IEEE Trans. on Knowledge and Data Engineering, 2004,16(11):1370−1386. [doi: 10.1109/TKDE.2004.68]

[48] Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: A survey. IEEE Trans. on Computational Biology and Bioinformatics, 2004,1(1):24−45. [doi: 10.1109/TCBB.2004.2]

[49] Smet RD, Marchal K. An ensemble method for querying gene expression compendia with experimental lists. In: Proc. of the 2010 IEEE Int'l Conf. on Bioinformatics and Biomedicine (BIBM). IEEE Press, 2010. 314−318. [doi: 10.1109/BIBM.2010. 5706583]

[50] Zou Q, Guo M, Liu Y, Wang J. A classification method for class-imbalanced data and its application on bioinformatics. Journal of Computer Research and Development, 2010,47(8):1407−1414 (in Chinese with English abstract).

[51] Zou Q, Li X, Jiang W, Lin Z, Li G, Chen K. Survey of mapreduce frame operation in bioinformatics. Briefings in Bioinformatics, 2014,15(4):637−647. [doi: 10.1093/bib/bbs088]

[52] Chen W, Cheng Y, Zhang S, Pan Q. Heuristic clustering method based on neighbor-seeds for 454 sequencing data. Ruan Jian Xue Bao/Journal of Software, 2014,25(5):929−938 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4547.htm [doi: 10.13328/j.cnki.jos.004547]

[53] Zou Q, Hu Q, Guo M, Wang G. HAlign: Fast multiple similar DNA/RNA sequence alignment based on the centre star strategy. Bioinformatics, 2015,31(15):2475−2481. [doi: 10.1093/bioinformatics/btv177]

[54] Dhollander T, Sheng QZ, Lemmens K, Moor BD, Marchal K, Moreau Y. Query-Driven module discovery in microarray data. Bioinformatics, 2007,23(19):2573−2580. [doi: 10.1093/bioinformatics/btm387]

[55] Zhao H, Cloots L, Bulcke TV, Wu Y, Smet RD, Storms V, Meysman P, Engleen K, Marchal K. Query-Based biclustering of gene expression data using probabilistic relational models. BMC Bioinformatics, 2011,12(s1):S37. [doi: 10.1186/1471-2105-12-S1-S37]

[56] Tseng VS, Chen LC, Kao CP. Constrained clustering for gene expression data mining. In: Proc. of the 12th Pacific-Asia Conf. on Advances in Knowledge Discovery and Data Mining (PAKDD). Berlin, Heidelberg: Springer-Verlag, 2008. 759−766. [doi: 10.1007/978-3-540-68125-0_73]

附中文参考文献:

[22] 闫雷鸣,孙志挥,吴英杰,张柏礼.联合聚类非线性相关的时序基因表达数据.计算机研究与发展,2008,45(11):1865−1873.

[31] 安平.基因表达数据的双聚类分析方法研究[硕士学位论文].苏州:苏州大学,2013.

[46] 岳峰,孙亮,王宽全,王永吉,左旺孟.基因表达数据的聚类分析研究进展.自动化学报,2008,34(2):113−120. [doi: 10.3724/SP.J.1004.2008.00113]

[50] 邹权,郭茂祖,刘扬,王峻.类别不平衡的分类方法及在生物信息学中的应用.计算机研究与发展,2010,47(8):1407−1414.

[52] 陈伟,程咏梅,张绍武,潘泉.邻域种子的启发式 454 序列聚类方法.软件学报,2014,25(5):929−938. http://www.jos.org.cn/1000-9825/25/929.htm [doi: 10.13328/j.cnki.jos.004547]

姜涛(1983－),男,河南滑县人,博士,讲师,CCF 专业会员,主要研究领域为生物信息检索,数据管理,数据挖掘.

陈伯林(1985－),男,博士,副教授,CCF 专业会员,主要研究领域为生物信息学,数据挖掘,数据管理.

李战怀(1961－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据管理,数据挖掘.

李卫榜(1979－),男,博士生,主要研究领域为数据质量,云计算,数据管理.

尚学群(1973－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据挖掘,生物信息学,数据管理.

殷知磊(1986－),男,博士生,讲师,主要研究领域为生物信息学,数据挖掘,数据库管理.