序列模式挖掘研究与发展

王 虎[1] 丁世飞[1,2]
(中国矿业大学计算机科学与技术学院 徐州 221008)[1]
(中国科学院计算技术研究所智能信息处理重点实验室 北京 100080)[2]

摘 要 序列模式挖掘是数据挖掘的一个重要研究课题,它在很多领域中都有着广泛的应用。首先讨论了序列模式挖掘的相关背景,然后对序列模式挖掘进行分类,并在此基础上对每一类序列模式挖掘算法的特点进行了介绍和比较;最后,对序列模式挖掘未来的研究重点进行展望,以便研究者对序列模式挖掘做进一步的研究。

关键词 数据挖掘,序列模式挖掘,闭合模式,增量式,多维模式

Research and Development of Sequential Pattern Mining (SPM)

WANG Hu[1] DING Shi-fei[1,2]
(School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221008, China)[1]
(Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China)[2]

Abstract Sequential pattern mining (SPM) is an important research subject of data mining, and used widely in many application fields. This paper firstly discussed background of sequential pattern minging, then classified it, and introduced as well as compared the features of sequential pattern minging algorithms based on the classification. Finally, discussed the future research on this field so that researchers can do further study.

Keywords Data mining, Sequential pattern mining (SPM), Closed pattern, Incremental pattern, Multi-dimentional pattern

1 引言

序列模式挖掘(Sequential pattern mining, SPM)是指从序列数据库中寻找频繁子序列作为模式的知识发现过程,它是数据挖掘的一个重要研究课题,在很多领域都有实际的应用价值,如客户购买行为模式的分析、Web访问模式的预测、疾病诊断、自然灾害预测、DNA序列分析等。通过对这些领域的数据运用序列模式挖掘技术,可以发现隐藏的知识,从而帮助决策者做出更好的决策,以获得巨大的社会价值和经济价值。深入理解高效的序列模式挖掘方法对在大型数据库中高效挖掘频繁子树、格、子图以及其他结构模式等有着重要意义。序列模式挖掘最早是由 Rakesh Agrawal 和 Ramakrishnan Srikant 针对购物篮数据分析提出来的,经过多年的发展,对序列模式挖掘的研究取得了比较大的进步,研究者们提出了很多性能良好的算法。本文将对每类算法的特点进行评述并对未来的研究重点进行展望。

2 序列模式挖掘算法的分类及研究现状

基本的序列模式挖掘的任务是找出序列数据库中满足用户定义的最小支持度阈值的所有序列的集合。序列模式挖掘和关联规则挖掘有着紧密的联系:一方面,关联规则挖掘以发现事物的内部联系为主,序列模式挖掘以发现事物之间的联系为主;另一方面,很多对于关联规则挖掘的研究都进一步促进了序列模式挖掘研究的发展。

2.1 基本的序列模式挖掘

(1)基于 Apriori 特性的算法

早期的序列模式挖掘算法都是基于 Apriori 特性发展起来的。Rakesh Agrawal 和 Ramakrishnan Srikant 在文献[1]中最早提出了序列模式挖掘的概念并且提出了 3 个基于 Apriori 特性的算法[1]: AprioriAll, AprioriSome, DynamicSome。基于这一思想,研究者又提出了 GSP[2] 算法,它对 AprioriAll 算法的效率进行了改进并且加入了时间限制、放宽交易的定义、加入了分类等条件,使序列模式挖掘更符合实际需要。GSP 算法是最典型的类 Apriori 算法,后来研究者又相继提出了 MFS[3] 算法和 PSP[4] 算法以改进 GSP 算法的执行效率。

基于 Apriori 特性的算法思想来源于经典的关联规则挖掘算法 Apriori,它满足一条重要的性质,即所有频繁模式的子模式也是频繁的。此类算法可以有效地发现频繁模式的完全集。但是类 Apriori 算法最大的缺点是需要多次扫描数据库并且会产生大量的候选集,当支持度阈值较小或频繁模式较长时这个问题更加突出。

到稿日期:2009-03-14 返修日期:2009-05-14 本文受国家自然科学基金资助项目(40574001),国家"863"计划基金资助项目(2006AA01Z128),中国科学院智能信息处理重点实验室开放基金资助项目(IIP2006-2)资助。

王 虎(1986—),男,硕士生,主要研究方向为数据挖掘与知识发现;丁世飞(1963—),男,博士后,教授,博士生导师,主要研究方向为机器学习与数据挖掘、人工智能与模式识别等。

(2)基于垂直格式的算法

最典型的是 SPADE[5]算法。它的基本思想是:通过把序列数据库转换成垂直数据库格式,然后利用格理论和简单的连接方法来挖掘频繁序列模式。SPADE 算法最大的优点是扫描数据库的次数大大减少,整个挖掘过程仅需扫描 3 次数据库,比 GSP 算法更优越。然而,SPADE 算法需要额外的计算时间和存储空间用以把水平格式的数据库转换成垂直格式,并且它的基本遍历方法仍然是广度优先遍历,需要付出巨大候选码的代价。

另一个典型的算法是 SPAM 算法。它实施了有效支持度计数与数据库垂直数位映象的表示方法相结合的搜索策略,挖掘长序列模式时效率特别高。

(3)基于投影数据库的算法

类 Apriori 算法由于会产生大量的候选集并且需要多次扫描数据库,因此在挖掘长序列模式方面效率很低。为了克服这些缺点,一些研究者开始另辟蹊径,提出了基于投影数据库的算法。此类算法采取了分而治之的思想,利用投影数据库减小了搜索空间,从而提高了算法的性能。比较典型的算法有 FreeSpan[6]和 PrefixSpan[7]。

FreeSpan 算法的基本思想是:利用当前挖掘的频繁序列集将数据库递归地投影到一组更小的投影数据库上,分别在每个投影数据库上增长子序列。FreeSpan 算法的优点是它能够有效地发现完整的序列模式,同时大大减少产生候选序列所需的开销,比典型的类 Apriori 算法 GSP 性能更优越。然而利用 FreeSpan 可能会产生很多投影数据库,如果一个模式在数据库中的每个序列中都出现,该模式的投影数据库将不会缩减;另外,由于长度为 k 的子序列可能在任何位置增长,搜索长度为(k+1)的候选序列需要检查每一个可能的组合,这是相当费时的。

针对 FreeSpan 的缺点,又提出了 PrefixSpan 算法。它的基本思想是:在对数据库进行投影时,不考虑所有可能的频繁子序列,而只是基于频繁前缀来构造投影数据库,因为频繁子序列总可以通过增长频繁前缀而被发现。PrefixSpan 算法使得投影数据库逐步缩减,比 FreeSpan 效率更高。并且它还采用了双层投影和伪投影两种优化技术以减少投影数据库的数量。PrefixSpan 算法的主要代价是构造投影数据库。在最坏的情况下,PrefixSpan 需要为每个序列模式构造投影数据库,如果序列模式数量巨大,那么代价也是不可忽视的。

除此之外,文献[8]中提出了一种无重复投影数据库扫描的算法 SPMDS。它通过对投影数据库的伪投影作单项杂凑函数,检测是否存在重复的投影,从而避免大量重复扫描数据库,很好地解决了密集数据集和长模式的挖掘问题。

(4)基于内存索引的算法

典型的算法是 MEMISP[9]。MEMISP 算法整个过程只需要扫描数据库一次,并且不产生候选序列也不产生投影数据库,大大地提高了 CPU 和内存的利用率。实验表明,MEMISP 比 GSP 和 PrefixSpan 更高效,而且对于数据库的大小和数据序列的数量也有较好的线性可伸缩性。

对于那些较大的不能一次装入内存的数据库,MEMISP 把它划分为能存储在内存中的部分数据库,然后对每个部分数据库应用 MEMISP 得到频繁序列,然后通过再一次扫描数据库得到最终的频繁序列。因此对于大型的数据库,MEMISP 也仅仅只需要扫描两次数据库。

(5)其他

除此之外,文献[10]提出了基于改进的 FP 树的算法 STMFP。它通过改进 FP 树的结构,使得树的每个节点可以存储一个项集。在扫描一次数据库后,STMFP 树可以存储所有的序列信息。另外,该算法提出了一种新的挖掘方法,它可以找到 STMFP 树中每条路径上从叶节点到根节点所有的组合从而更有效地挖掘出序列模式。STMFP 算法的最大优点是在整个挖掘过程中只需要扫描数据库一次,提高了挖掘效率。然而,当序列数据库较大时,构建 STMFP 树的代价也会增大。

文献[11]提出一种基于 2-序列矩阵的算法 ESPE。它把一个序列分隔成两部分 X 和 Y。这里 X 是一个候选 2-序列,Y 是序列中去掉 X 后余下的序列。相比之前的算法,ESPE 有很多优点:首先,它只需要扫描数据库一次;其次,它不需要产生所有的候选序列,算法执行效率更高且减少了内存空间的浪费;最后,由于 ESPE 可以保存所有可能有趣的序列的支持度计数,因此,它不需要提前确定支持度阈值。另外此算法也可以应用于增量式序列模式的挖掘。

2.2 闭合序列模式挖掘

传统的序列模式挖掘的任务,是挖掘序列数据库中满足最小支持度阈值的频繁子序列的完全集。然而,当频繁模式较长或支持度阈值较低时,传统的算法性能会明显降低。这时,挖掘闭合序列模式是一个更好的选择。序列集中的一个序列 s 是闭合的,当且仅当此序列集中不存在和 s 的支持度相同的超序列。挖掘闭合序列模式比挖掘序列模式的完全集更加精简有效,而且具有相同的效力。

CloSpan[12]算法是第一个挖掘闭合序列模式的算法。它把挖掘过程分为两个阶段:在第一阶段产生候选集,通常这个候选集要大于最终的闭序列集;在第二个阶段是从候选集中删除那些非闭的序列。CloSpan 基于 PrefixSpan,采用了两种剪枝技术,用一个基于哈希的算法优化搜索空间,有效地找出了更紧凑的闭合序列集,而且运行效率很高。

由于 CloSpan 采用了候选维护-测试的方法,因此它需要维护已经挖掘出的闭序列的候选集。当模式较长或者支持度阈值较低时,它会付出巨大的维护代价。因此随后研究者又提出了 BIDE[13]算法。它采用了一种称为双向扩展的序列闭合检查方法,可以更有效地检查模式的闭合性。并且通过使用后向扫描剪枝方法和扫描跳跃优化技术,BIDE 可以更深地剪枝搜索空间。同时 BIDE 不需要对候选闭序列集进行维护。实验表明 BIDE 性能优于 CloSpan,并且具有良好的线性可伸缩性。

2.3 增量式序列模式挖掘

前面介绍的算法,如 AprioriAll, GSP, PrefixSpan, SPADE 等都是基于事务数据库是静止的这个假设而提出来的。然而现实世界中事务数据库总是随时间变化而变化的。例如顾客购物的数据库在新的商品被加入到现有的顾客中或者有新的顾客加入时,都会导致其发生变化。如果此时仍然采用之前的算法,每当事务数据库有变化时就重新运行算法从头开始挖掘的话,不仅会导致效率低下而且对资源也是一种极大的浪费。因此有必要对事务数据库的增量更新进行深入的研究。

GSP+和 MFS+[15]算法分别是基于 GSP 算法和 MFS 算法的增量式序列模式挖掘算法。GSP+和 GSP 的结构相同。不同之处在于 GSP+采用了和 GSP 不同的剪枝策略，它仅仅遍历更新的数据库来计算候选序列的支持数。类似地，MFS+采用了和 GSP+相同的剪枝策略。

在 SPADE 算法的基础上，研究者提出了 ISM[16]算法。ISM 构建了一个增量序列格，这个序列格包含旧数据库中所有的频繁序列和反向边界序列。当新的数据加入时，算法扫描一次增量部分并且把结果合并到格里。ISM 使用垂直数据存储方式，在建立数据结构方面的花销要比其他多数序列模式挖掘算法在速度上有所增加。但是如果交易数据库非常大，那么反向边界也会非常大，这将浪费很多内存空间。另外 ISM 算法只考虑了增加新的序列的情况。

ISE[17]算法不仅考虑增加新序列的情况，而且考虑了在序列中增加新后缀的情况。它通过把增量数据库中的序列和原始数据库中的序列进行连接产生整个数据库的候选序列。因此当原始数据库中的数据被更新时，它避免了保存大量的反向边界序列并且不需要重新计算这些序列。但是它需要更多地搜索数据库，从而而且 ISE 只是扩展了原始数据库中频繁序列的后缀，并没有扩展前缀。

算法 IUS[18]则对前缀和后缀都进行了扩展。IUS 使用了在 ISM 中用到的反向边界，并且定义了反向边界序列的最小支持数，只有支持度超过这个最小支持数的序列才能被反向边界包含，因此它需要更小的内存空间。

Jiawei Han 等人提出的 IncSpan[19]算法在挖掘增量式序列模式方面给出了一些新的思路，包括维护一个“几乎频繁”的序列集作为更新的数据库中的候选，并且提出了两个优化技术：反向模式匹配和共享投影。实验表明，IncSpan 的性能优于 ISM。另外，文献[14]中提出了一种基于投影数据库的增量式更新算法 ISPBP。在更新序列模式库时，ISPBP 采用了间接拼接的方法，只需计算增量数据库，从而避免了对更新后数据库的重新计算。对于因增量数据库新产生的频繁模式，它利用了在增量数据库中出现的频繁项集来减小投影数据库，这样可以得到更小的投影数据库，从而提高了算法的效率。

2.4 多维序列模式挖掘

多维序列模式挖掘在序列模式挖掘的基础上，考虑了其他的维度信息，从而挖掘出多维信息中感兴趣的模式，比如在顾客购买行为分析中考虑顾客的年龄、性别等信息得到的多维序列模式。这种模式包含了更多的信息，应用价值更高。文献[20]中首次提出了多维序列模式挖掘的概念，并提出了 3 种算法：UniSeq，Seq-Dim 和 Dim-Seq。

UniSeq 算法通过把多维信息嵌入到每个序列中形成扩展的序列数据库，然后利用 PrefixSpan 算法对扩展的序列数据库进行挖掘从而得到多维序列模式。UniSeq 可以有效地挖掘多维序列模式并且较容易实现。然而由于所有的维度值被当作序列的项来处理，因此它不能使用诸如 BUC 和 H-cubing 等一些有效的挖掘多维非序列的算法，因而当维度很多时其挖掘性能会下降。

Seq-Dim 算法和 Dim-Seq 算法将序列模式和多维模式分开进行挖掘。在挖掘序列模式时可以使用 PrefixSpan 算法，而挖掘多维模式时可以使用类 BUC 算法。Seq-Dim 算法先挖掘出所有的序列模式，然后对每个序列模式构造相应的投影数据库，并从中挖掘出多维模式。Dim-Seq 则采用和 Seq-Dim 相反的顺序，先挖掘多维模式，然后挖掘序列模式。实验表明，在上面 3 种算法中，Seq-Dim 性能最好。UniSeq 也是一种有效的和可伸缩的算法，特别是当维数较低时，它的性能比其他两种算法更好。Dim-Seq 算法相比其他两种算法效率最低。

文献[21]针对 Seq-Dim 算法需要多次扫描投影数据库而造成开销较大的问题，提出了一种改进的算法 Seq-mdp。该算法首先在序列信息中挖掘序列模式，然后对每个序列模式，在包含此模式的所有元组中的多维信息中挖掘频繁 1-项集，由得到的频繁 1-项集开始，循环地由频繁 $k-1$ 项集连接生成频繁 k 项集，从而得到所有的多维模式。该算法通过扫描不断缩小的频繁 $k-1$ 项集来生成频繁 k 项集，减少了扫描投影数据库的次数，因而减少了时间开销。

2.5 其他序列模式挖掘的研究

近几年，研究者对序列模式挖掘进行了扩展，结合其他领域，提出了一些新的挖掘序列模式的概念，包括文献[28-31]中提出的基于约束的序列模式挖掘、文献[32,33]中提出的并行序列模式挖掘、文献[22-25]中提出的周期序列模式挖掘、文献[26]中提出的分布式序列模式挖掘、文献[27]中提出的挖掘序列模式图等，这些研究都极大地丰富和发展了序列模式挖掘的研究内容。

3 序列模式挖掘的发展趋势

自从 Rakesh Agrawal 和 Ramakrishnan Srikant 1995 年提出序列模式挖掘的概念以来，经过十余年的发展，序列模式挖掘研究取得了比较大的进展，但是仍然存在着一些问题，比如支持度阈值的设定还没有好的评判方法，与相关领域知识的结合不够，针对海量数据的挖掘效率还不高等。这些都是下一步需要解决的问题。

另外，序列模式挖掘在未来的研究重点应该集中在：1）如何进一步提高算法效率，缩小搜索空间；2）与相关领域知识的结合；3）序列模式挖掘的拓展，如多维序列模式挖掘、增量式序列模式挖掘、并行挖掘、设计面向非关系数据库的序列模式挖掘等。

序列模式挖掘是数据挖掘领域的重要内容，已经应用于许多行业。可以预见，序列模式挖掘在未来仍将快速迅猛地发展。

参 考 文 献

[1] Agrawal R, Srikant R. Mining sequential pattern[C]// Proc. of the 11 th International Conference on Data Engineering. Taipei, 1995

[2] Srikant R, Agrawal R. Mining sequential patterns : Generalizations and performance improvements[C]// Proc. of the 5th International Conference on Extending Database Technology. Avignon, 1996

[3] Zhang M, Kao B, Yip C, et al. A GSP-based eficient algorithm for mining fequent sequences[C]// Proc. of International Conference on Artificial Intelligence. Nevada, 2001

[4] Masseglia F, Cathala F, Poncelet P. The PSP approach for mining sequential patterns[C]// Proc. of the 2nd European. Sympo-

sium on Principles of Data Mining and Knowledge Discovery. Berlin: Springer-Verlag, 1510: 176-184

[5] Zaki M J. SPADE: An eficient algorithm for mining frequent sequences[J]. Machine Learning, 2001, 41(1): 31-60

[6] Han J, Pei J, Mortazvi-asl B, et al. FreeSpan: frequent pattern-projected sequential pattern mining[C]// Proc. of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2000: 355-359

[7] Pei J, Han J. PrefixSpan: mining sequential patterns eficiently by prefix-projected pattern growth[C]// Proc. of the 7th International Conference on Data Engineering. Washington DC: IEEE Computer Society, 2001: 215-224

[8] 张坤，朱杨勇. 无重复投影数据库扫描的序列模式挖掘算法[J]. 计算机研究与发展，2007，44(1)：126-132

[9] Lin Ming-yen, Lee S Y. Fast discovery of sequential patterns by memory indexing[C]// Proc. of the 4th International Conference on Data Warehousing and Knowledge Discovery. London, UK: Springer-Verlag, 2002: 150-160

[10] Sui Yi, Shao Fengjing, Sun Rencheng, et al. A sequential pattern mining algorithm based on improved FP-tree[C]// Proc. of 9th ACIS Int. and SNPD 2008. 2008: 440-444

[11] Hsieh Chia-Ying, Yang Don-Lin, Wu Jungpin. An efficient sequential pattern mining algorithm based on the 2-sequence matrix[C]// Proc. of IEEE International Conference on Data Mining Workshops, ICDM Workshops 2008. Pisa, Italy: IEEE Computer Society, 2008: 583-591

[12] Yan X, Han J, Afshar R. CloSpan: mining closed sequential pattens in large datasets[J]. Data Mi-ning, 2003, 16(5): 40-45

[13] Wang J, Han J. BIDE: Efficient mining of frequent closed sequences[C]// Proc. of 20th International Conference on Data Engineering - ICDE 2004. Boston, USA: IEEE Computer Society, 2004: 79-90

[14] 陆介平，刘月波，等. 基于投影数据库的序列模式挖掘增量式更新算法[J]. 东南大学学报，2007，36(3)：457-462

[15] Zhang Ming-hua, Kao B, Cheung D W, et al. Eficient algorithms for incremental update of frequent sequences[C]// Proc. of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. London, UK: Springer-Verlag, 2002: 186-197

[16] Parthasarathy S, Zaki M J, Ogihara M, et al. Incremental and interactive sequence mining[C]// Proc. of the 8th International Conference on Information and Knowledge Management. Kansas City, New York: ACM Press, 1999: 251-258

[17] Masseglia F, Poncelet P, Teisseire M. Incremental mining of sequential patterns in large databases[J]. Data and Knowledge Engineering, 2003, 46(1): 97-121

[18] Zheng Qing-guo, Xu Ke, Ma Shi-ling, et al. The algorithms of updating sequential patterns[C]// Proc. of the 5th International Workshop on High Performance Data Mining. Washington DC, 2002

[19] Cheng Hong, Yan X, Han J. IncSpan: incremental mining of sequential patterns in large database[C]// Proc. of the 10th International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2004: 527-532

[20] Pinto H, Han J, Pei J, et al. Multi-dimensional sequential pattern mining[C]// Proc. of the 10th International Conference on Information and Knowledge Management. Atlanta, New York: ACM Press, 2001: 81-88

[21] 肖仁财，薛安荣. 一种挖掘多维序列模式的有效方法[J]. 计算机工程与应用，2008，44(6)：187-190

[22] Han J, Dong G, Yin Y. Efficient mining of partial periodic patterns in time series database[C]// Proc. of the 15th International Conference on Data Engineering. Washington DC: IEEE Computer Society, 1999

[23] Yang J, Wang Wei, Yu P S. Mining asynchronous periodic patterns in time series data[C]// Proc. of the 6th International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2000: 275-279

[24] Elfeky M G. Incremental mining of partial periodic patterns in time series databases[EB/OL]. (2000). http://citeseer.ist.psu.edu/421296.html

[25] Bettini C, Wang X S, Jajodia S. Mining temporal relationships with multiple granularities in time sequences[J]. Data Engineering Bulletin, 1998, 21(1): 32-38

[26] 邹翔，张巍，刘洋，等. 分布式序列模式发现算法的研究[J]. 软件学报，2005，16(7)：1262-1269

[27] 吕静，王晓峰. 序列模式图及其构造算法[J]. 计算机学报，2004，27(6)：782-787

[28] Garofalakis M N, Rastogi R, Shim K. Spirit: sequential pattern mining with regular expression constraints[C]// Proc. of the 25th International Conference on Very Large Databases. San Francisco, CA: Morgan Kaufmann Publishers Inc, 1999: 223-234

[29] Pei J, Han J, Wang W. Mining sequential patterns with constraints in large databases[C]// Proc. of 11th International Conference on Information and Knowledge Management. McLean, USA: Association for Computing Machinery, 2002: 18-25

[30] Capelle M, Masson C, Boulicaut J. Mining frequent sequential patterns under a similarity constraint[C]// Proc. of Third International Conference Intelligent Data Engineering and Automated Learning-IDEAL 2002. Berlin, Germany: Springer-Verlag, 2002: 1-6

[31] Lin Ming-Yen, Hsueh Sue-Chen, Chang Chia-Wen. Mining closed sequential patterns with time constraints[J]. Journal of Information Science and Engineering, 2008, 24(1): 33-46

[32] Zhou Lijuan, Qin Bai, et al. Research on parallel algorithm for sequential pattern mining[C]// Proc. of SPIE-The International Society for Optical Engineering. Orlando, USA: SPIE, P. O. Box10, Bellingham WA, WA 98227-0010, United States, 2008

[33] Wang Jinlin Chen Xi, et al. Parallel research of sequential pattern data mining algorithm[C]// Proc. Int. Conf. Comput. Sci. Softw. Eng., CSSE. Wuhan, China: Inst. of Elec. And Elec. Eng. Computer Society, 445 Hoes Lane-P. O. Box 1331, Piscataway, NJ 08855-1331, United States, 2008: 348-353

[34] 陈卓，杨炳儒，等. 序列模式挖掘综述[J]. 计算机应用研究，2008，25(7)：1960-1963

[35] 马传香，张凌. 序列模式挖掘算法的分析与比较[J]. 湖北大学学报，2006，28(2)：138-143