首先证明$|ni(\alpha)|\geq|mo(\alpha)|$,即证明:对于情节$\alpha$相邻的最早转移发生$h_i$和$h_{i+1}$,若$h_i$是最小发生,则$h_i$和$h_{i+1}$一定是非交错发生.我们使用反证法证明.因为$h_i$是最小发生,所以有$h_i(t_k)<h_{i+1}(t_k)$,其中,$1\leq k\leq|\alpha|$.假设$h_i$和$h_{i+1}$不是非交错发生,则必然存在一个小于$|\alpha|$的$k$,使得$h_{i+1}(t_k)<h_i(t_{k+1})$,则有$h_i(t_k)<h_{i+1}(t_k)<h_i(t_{k+1})<h_{i+1}(t_{k+1})$.因$h_i(t_k)<h_{i+1}(t_k)<h_i(t_{k+1})$,且$h_i$和$h_{i+1}$是相邻的最早转移发生,所以有$h_i(t_{k+1})=h_{i+1}(t_{k+1})$,这与$h_i(t_{k+1})<h_{i+1}(t_{k+1})$矛盾,说明若$h_i$是最小发生,则$h_i$和$h_{i+1}$一定是非交错发生,从而证明了$|ni(\alpha)|\geq|mo(\alpha)|$.

其次证明$|mn(\alpha)|=|no(\alpha)|$.对于$no(\alpha)$中的任何一个发生$h$,若$h$是最小发生,则$h\in mn(\alpha)$,否则,根据推论2,一定存在一个发生$h'$,满足$h'(t_{|\alpha|})=h(t_{|\alpha|})$.显然,$h'$与$no(\alpha)$中的其他发生都是非重叠的.这表明,最小且非重叠发生的次数至少等于非重叠发生的次数,即$|mn(\alpha)|\geq|no(\alpha)|$.假设$|mn(\alpha)|>|no(\alpha)|$,则说明$no(\alpha)$不是$\alpha$的非重叠发生的最大集,这与推论4矛盾.因此,$|mn(\alpha)|=|no(\alpha)|$.

最后证明$|mo(\alpha)|\geq|mn(\alpha)|$.因为$mn(\alpha)$中的发生都是最小发生,所以$mn(\alpha)\subseteq mo(\alpha)$,即$|mo(\alpha)|\geq|mn(\alpha)|$. □

4 实验评估

我们使用文献[20,21]的合成数据集与真实数据集来进行实验评估.对于合成数据集,首先使用IBM合成数据生成器Quest Market-Basket的修改版生成了每个交易为单个项的交易序列,通过设置D=0.001,C=300000,N=20,S=300000,其中,参数D表示交易序列的个数(单位为1 000),C表示每个交易序列中交易的平均个数,N表示所有交易项的类型种数(单位为1 000),S为最长交易序列中交易的平均个数,这样就得到了一个20 000种交易项类型上的由300 000个交易组成的交易序列.然后,为该交易序列中的每个交易依次赋上一个连续的正整数以作为每个交易发生的时间戳,这样,就构造了一个20K种事件类型上的由300K个事件组成的事件序列.

对于真实数据集,考虑到作为国内最具影响力的知识传播与数字化学习平台,中国知网为全社会提供了最丰富、最全面的文献资源,为了能够发现中国知网相关文献之间的引用关系,并为广大学者展开相关研究提供个性化的推荐服务,我们选用了中国知网的一个Web服务器上从2010年11月1日~2010年11月30日的日志数据,该日志数据包括了相关读者对132 885种不同文献的211 665个阅读序列.

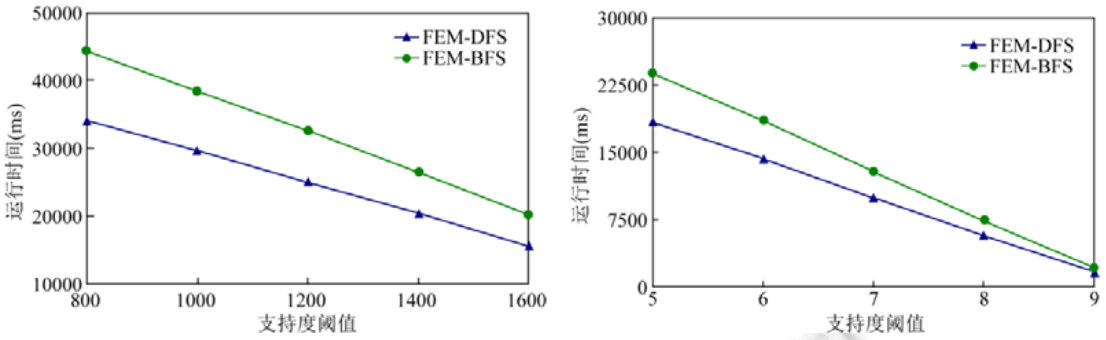
通过7组实验对比算法FEM-DFS和文献[19]提出的算法(为描述方便,简称为FEM-BFS)的时空性能,并分析FEM-DFS在不同支持度定义下的挖掘结果.实验的硬件环境为3.6GHz Intel(R) Core(TM) i7-4790 CPU,内存为8GB,操作系统为Windows 8,程序采用Java实现.

实验1. 运行时间vs.支持度阈值.合成数据集和真实数据集的窗口宽度分别设定为200和5天,通过改变支持度阈值,得到如图2所示的两种算法各自在7个支持度定义下的平均运行时间.

可以看出,随着支持度阈值的增加,两种算法的运行时间都在线性减少,且FEM-DFS要优于FEM-BFS,主要原因是:支持度阈值越大,频繁情节越少;FEM-BFS采用广度优先搜索策略,需要多遍扫描事件序列,而FEM-DFS采用深度优先搜索策略,只需单遍扫描事件序列.

尽管算法FEM-DFS和FEM-BFS在运行时间上存在差异,但它们的挖掘结果相同.例如,〈基于隐马尔可夫模型的多步攻击预测研究,面向分布数据安全的误用检测算法和入侵检测系统的研究,隐马尔可夫模型在入侵检测中的应用,基于入侵响应的入侵警报关联性的攻击预测算法,基于因果网络的攻击计划认别与预测,面向网络安全的基于入侵事件的早期预警方法〉是两种算法在真实数据集上都能挖掘得到的一个频繁6-情节,该情节刻画了一种行为模式:读者们旨在研究一个基于隐马尔可夫模型的攻击预测方法(见第1篇阅读文档),他们首先研究了两种常用的入侵检测模型,即误用检测模型(见第2篇阅读文档)和异常检测模型(见第3篇阅读文档),然后分析了现有攻击方法的不足(见后3篇阅读文档).算法FEM-DFS和FEM-BFS的挖掘结果有助于CNKI向读者提供个性化的文献阅读推荐服务.

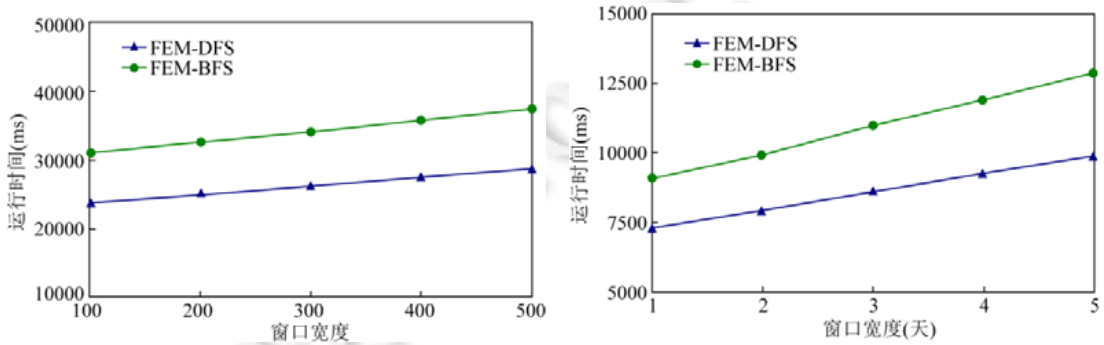
实验2. 运行时间vs.窗口宽度.合成数据集和真实数据集的支持度阈值分别设定为1 200和7,通过改变窗口宽度,得到如图3所示的两种算法各自在窗口发生、头发生和总发生支持度定义下的平均运行时间.



(a) 合成数据集 (b) 真实数据集

Fig.2 Average runtime vs. support threshold

图 2 运行时间 vs.支持度阈值



(a) 合成数据集 (b) 真实数据集

Fig.3 Average runtime vs. window width

图 3 运行时间 vs.窗口宽度

可以看出,随着窗口宽度的增加,两种算法的运行时间也在线性增加,这是因为窗口宽度越大,频繁情节越多.另外,FEM-DFS 要优于 FEM-BFS,这是因为两者采用了不同的搜索策略.

实验 3. 运行时间 vs.序列长度.设定合成数据集的支持度阈值和窗口宽度分别为 800 和 200,真实数据集的支持度阈值和窗口宽度分别为 7 和 5 天,并选择前 100K、前 150K、前 200K、前 250K 个和所有 300K 个合成数据作为 5 个合成子序列,选择前 6 天、前 12 天、前 18 天、前 24 天和所有 30 天真实数据作为 5 个真实子序列,通过改变序列长度,得到如图 4 所示的两种算法各自在 7 个支持度定义下的平均运行时间.

可以看出,两种算法的运行时间都随着序列长度的增加而线性增加,这是因为序列长度越大,频繁情节越多.另外,FEM-DFS 优于 FEM-BFS,这也源于两者不同的搜索策略.

实验 4. 内存开销 vs.支持度阈值.与实验 1 的设置相同,通过改变支持度阈值,得到如图 5 所示的两种算法各自在 7 个支持度定义下的平均内存开销.

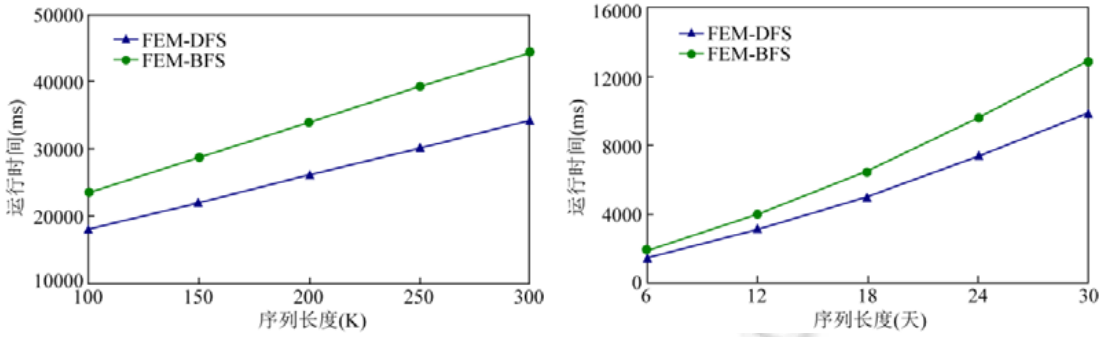
可以看出,随着支持度阈值的增加,两种算法的内存开销都在线性减少,且 FEM-DFS 要优于 FEM-BFS,原因与实验 1 相同.

实验 5. 内存开销 vs.窗口宽度.与实验 2 的设置相同,通过改变窗口宽度,得到如图 6 所示的两种算法各自在窗口发生、头发生和总发生支持度定义下的平均内存开销.

可以看出,随着窗口宽度的增加,两种算法的内存开销都在线性增加,但 FEM-DFS 要优于 FEM-BFS,原因与实验 2 相同.

实验 6. 内存开销 vs.序列长度.与实验 3 的设置相同,通过改变序列长度,得到如图 7 所示的两种算法各自在 7 个支持度定义下的平均内存开销.

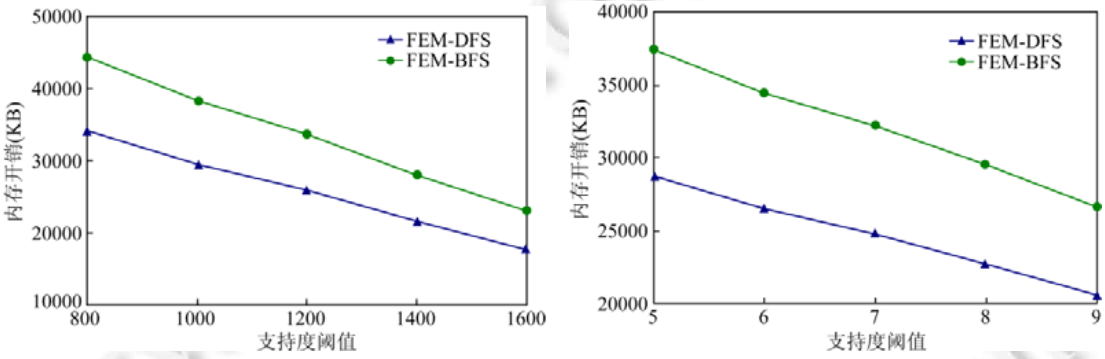
可以看出,两种算法的内存开销都随着序列长度的增加而呈线性增加,但 FEM-DFS 要优于 FEM-BFS,原因与实验 3 相同.



(a) 合成数据集 (b) 真实数据集

Fig.4 Average runtime vs. sequence length

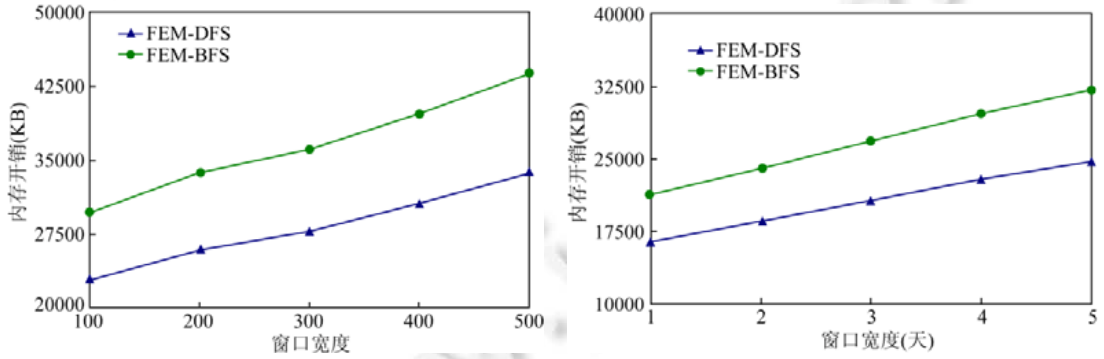
图 4 运行时间 vs.序列长度



(a) 合成数据集 (b) 真实数据集

Fig.5 Average memory vs. support threshold

图 5 内存开销 vs.支持度阈值

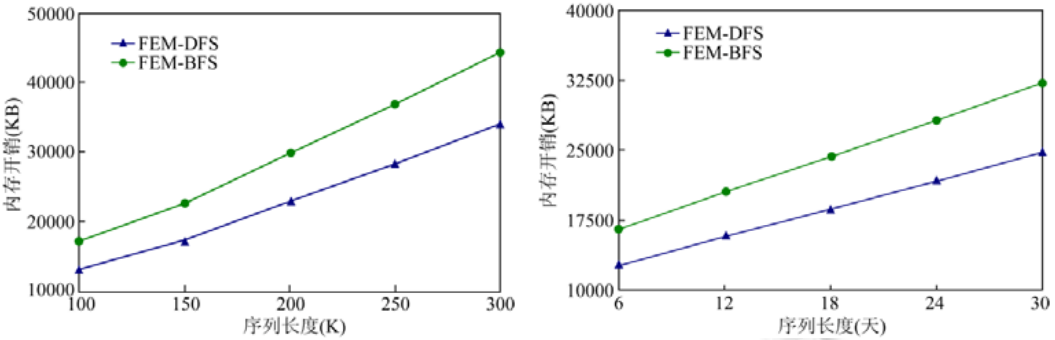


(a) 合成数据集 (b) 真实数据集

Fig.6 Average memory vs. window width

图 6 内存开销 vs.窗口宽度

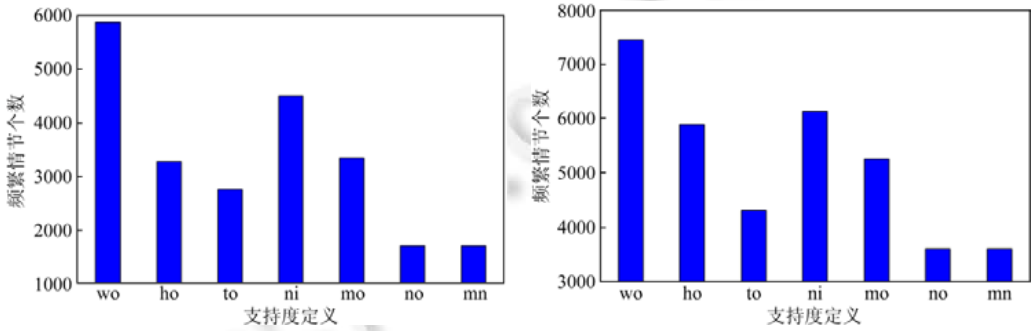
实验 7. 频繁情节个数 vs.支持度定义.选择 300K 合成数据集和 30 天真实数据集,两个数据集的支持度阈值分别设定为 800 和 7,窗口宽度分别设定为 200 和 5 天,通过改变支持度定义,得到如图 8 所示的算法 FEM-DFS 在不同支持度定义下的频繁情节个数.



(a) 合成数据集 (b) 真实数据集

Fig.7 Average memory vs. sequence length

图 7 内存开销 vs.序列长度



(a) 合成数据集 (b) 真实数据集

Fig.8 Episode number vs. support definition

图 8 频繁情节个数 vs.支持度定义

可以看出,基于窗口发生、头发生、总发生的 3 种支持度定义,发现的频繁情节个数依次递减;基于非交错发生、最小发生、非重叠发生的 3 种支持度定义,发现的频繁情节个数也在依次递减;基于非重叠发生、最小且非重叠发生的两种支持度定义,发现的频繁情节个数相同.主要原因:同一情节的发生次数在不同支持度定义下有着固定的比较关系,这与定理 5 一致.

5 总结与展望

针对现有频繁情节挖掘算法存在的不足,本文提出了一个采用深度优先搜索方式和共享前/后缀树存储结构的频繁情节挖掘算法 FEM-DFS,满足了实际情况下用户多变的支持度定义需求,实验评估证实了算法 FEM-DFS 能够有效地发现事件序列上的频繁情节.

未来将基于事件流环境,研究一种能够融合多种支持度定义的频繁情节挖掘算法.

References:

[1] Mannila H, Toivonen H, Verkamo AI. Discovery of frequent episodes in event sequences. Data Mining and Knowledge Discovery, 1997,1(3):259–289.

[2] Méger N, Rigotti C. Constraint-based mining of episode rules and optimal window sizes. In: Proc. of PKDD. 2004. 313–324.

[3] Lin YF, Huang CF, Tseng VS. A novel methodology for stock investment using high utility episode mining and genetic algorithm. Applied Soft Computing, 2017,59:303–315.

[4] Ma X, Pang HH, Tan KL. Finding constrained frequent episodes using minimal occurrences. In: Proc. of the ICDM. 2004. 471–474.

[5] Zhou WZ, Liu HY, Cheng H. Mining closed episodes from event sequences efficiently. In: Proc. of the PAKDD. 2010. 310–318.

[6] Wu JJ, Wan L, Xu ZR. Algorithms to discover complete frequent episodes in sequences. In: Proc. of the PAKDD. 2011. 267–278.

[7] Wu CW, Lin YF, Yu PS, Tseng VS. Mining high utility episodes in complex event sequences. In: Proc. of the SIGKDD. 2013. 536–544.
[8] Lin SK, Qiao JZ. An episode mining method based on episode matrix and frequent episode tree. Control and Decision, 2013,28(3): 339–344 (in Chinese with English abstract).
[9] Ao X, Luo P, Li CK, Zhuang FZ, He Q, Shi ZZ. Discovering and learning sensational episodes of news events. In: Proc. of the WWW. 2014. 217–218.
[10] Huang KY, Chang CH. Efficient mining of frequent episodes from complex sequences. Information Systems, 2008,33(1):96–114.
[11] Iwanuma K, Ishihara R, Takano Y, Nabeshima H. Extracting frequent subsequences from a single long data sequence. In: Proc. of the ICDM. 2005. 186–193.
[12] Avinash A, Ibrahim A, Sastry PS. Pattern-growth based frequent serial episode discovery. Data & Knowledge Engineering, 2013,87:91–108.
[13] Laxman S. Discovering frequent episodes: Fast algorithms, connections with HMMs and generalizations [Ph.D. Thesis]. Bangalore, 2006.
[14] Laxman S, Sastry PS, Unnikrishnan K. Discovering frequent episodes and learning hidden Markov models: A formal connection. IEEE Trans. on Knowledge and Data Engineering, 2005,17(11):1505–1517.
[15] Laxman S, Sastry PS, Unnikrishnan KP. A fast algorithm for finding frequent episodes in event streams. In: Proc. of the SIGKDD. 2007. 410–419.
[16] Zhu HS, Wang P, He XM, Li YJ, Wang W, Shi BL. Effcient episode mining with minimal and non-overlapping occurrences. In: Proc. of the ICDM. 2010. 1211–1216.
[17] Zhu HS, Wang P, Wang W, Shi BL. Discovering frequent closed episodes from an event sequence. In: Proc. of the IJCNN. 2012. 2292–2299.
[18] Liao GQ, Yang XT, Xie S, Yu PS, Wan CX. Two phase mining for frequent closed episodes. In: Proc. of the WAIM. 2016. 55–66.
[19] Avinash A, Laxman S, Sastry PS. A unified view of the apriori-based algorithms for frequent episode discovery. Knowledge and Information Systems, 2012,31:223–250.
[20] Zhu HS, Wang W, Shi BL. Extracting non-redundant episode rules based on frequent closed episodes and their generators. Chinese Journal of Computers, 2012,35(1):53–64 (in Chinese with English abstract).
[21] Zhu HS. Research on stream prediction based on episode rule matching [Ph.D. Thesis]. Shanghai: Fudan University, 2011 (in Chinese with English abstract).

附中文参考文献:

[8] 林树宽,乔建忠.一种基于情节矩阵和频繁情节树的情节挖掘方法.控制与决策,2013,28(3):339–344.
[20] 朱辉生,汪卫,施伯乐.基于频繁闭情节及其生成子的无冗余情节规则抽取.计算机学报,2012,35(1):53–64.
[21] 朱辉生.基于情节规则匹配的数据流预测研究[博士学位论文].上海:复旦大学,2011.

朱辉生(1968－),男,博士,教授,CCF 高级会员,主要研究领域为数据库,数据挖掘,大数据.

汪卫(1970－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为数据库,数据挖掘,大数据.

陈琳(1981－),女,副教授,CCF 专业会员,主要研究领域为图像处理,模式识别.

施伯乐(1935－),男,教授,博士生导师,CCF 高级会员,主要研究领域为数据库,数据挖掘.

倪艺洋(1986－),女,博士,副教授,主要研究领域为移动通信,机器学习.