Eduardo Nogueira da Ponte

"Mineração Eficiente de Padrões Sequênciais da Indexação de Seqüências Candidatas"

Sequential Patterns represent an important type of information extracted in processes of data mining and has a great number of applications. The mining sequential patterns objective to discovery all the frequent sequences of events, extracted from a transactional database.

From the analysis of the GSP algorithm behavior, used in the extraction of sequential patterns, was identified a high computational cost in the phase of support counting of the candidates sequences. For this reason, in this dissertation, a new data structure was proposed, called ABS-SP (Array-Based Structure for Sequential Patterns), in substitution to the hash-tree used in GSP algorithm. With the ABS-SP structure, the candidates sequences are accessed directly, through the indexation of these candidates in a vector and a matrix, reducing the access time and, consequently, resulting in a substantial reduction of the execution time of the algorithm.

In experimental results, the algorithm with the ABS-SP structure presented, for minimum support and databases different, a substantial reduction in the execution time in relation to the GSP, that uses the hash-tree. Initially, ABS-SP structure was applied in the identification of the frequent sequences of size two. Later, it was verified that, with the reduction of the minimum support, the computational cost in the identification of the frequent sequences of size three becomes high. Then, in function of this observation, the structure proposed was extended for the support counting of the candidates in the third iteration of the algorithm. Moreover, it was developed a proposal of ABS-SP structure extension for the subsequent iterations.