URN etd-0407109-183712-941
Author Guo-Cheng Lan
Author's Email Address No Public.
Statistics This thesis had been viewed 47 times. Download 0 times.
Department
Year 2005
Semester 2
Degree Master
Type of Document Master's Thesis
Language English
Title Efficient Approaches for the Filtration Mechanisms of Data Mining
Date of Defense 0000-00-00
Page Count 278

Keyword
Data mining
association rules
sequential patterns
traversal patterns
the filtration mechanisms
the correlation about traversal and purchasing
combination patterns

Abstract
In recent years, the technology of data mining is generally applied to the various commercial domains. Many of association rules mining algorithms were proposed to improve the efficiency of data mining or save the utility rate of memory. In this thesis, we aim at the four research subjects which are association rule, sequential pattern, traversal pattern, and the correlation about traversal path and purchasing merchandises to propose several efficient approaches.
First, we propose new mining approaches of association rule such as EFI and GRA. The one of characters EFI (An Efficient Approach for Filtering Infrequent Itemsets) is the two phase filtration mechanisms. EFI only generates these itemsets which are the most possible to be frequent via the two filtration mechanisms. In the mining frequent itemsets, EFI does not generate candidate sets and scans the database four times, and then finish the mining task quickly. In addition, GRA (Gradation Reduction Approaches) is a level-wise technique. The one of the characters of GRA algorithm is the gradation filtration mechanisms, and the algorithm uses the simple mask method to generate itemsets. In the mining association rules, GRA can avoid generating a huge number of unnecessary itemsets via the gradation filtration mechanisms, and the algorithm does not need to generate the candidate sets, and then finish the mining task quickly. Besides, when the algorithms deal with the very large databases, EFI does not modify any mining process, and then perform the mining task. In addition, we propose another algorithm GRA-M (Gradation Reduction Approaches – Modified Version) which is modified from the GRA algorithm. EFI and GRA-M will first divide the large database into several sub-databases which are loaded in the memory by the algorithms. The algorithms only perform each sub-database four times I/O processes, and then finish the mining task.
Next, we also propose the algorithms SFA (Mining Sequential Patterns Using Filtering Approaches) and GRS (Gradation Reduction Approaches for Mining Sequential Patterns) to discover the sequential patterns. Because SFA is extended from the EFI algorithm, the algorithm also only generates these subsequences which are the most possible to be frequent via the two phase filtration mechanisms, and the algorithm scans the database four times without generating candidate sets, and then finishes the mining task. In the same way, GRS is modified from the GRA algorithm. GRS can effectively reduce a huge number of unnecessary subsequences via the gradation filtration

mechanisms, and the algorithm does not generate candidate set, and then finishes the mining task. Besides, when the algorithms deal with the very large databases, SFA does not modify any mining process, and then perform the mining task. In addition, we propose another algorithm GRS-M (Gradation Reduction Approaches for Mining Sequential Patterns – Modified Version) which is modified from the GRS algorithm. SFA and GRS-M will first divide the large database into several sub-databases which are loaded in the memory by the algorithms. The algorithms only perform each sub-database four times I/O processes, and then finish the mining task.
Now, the e-commerce websites are growing fast at the surprising speed. If we can understand the behaviors of users' traversal path in the websites, we can make a better target marketing. Therefore, we propose a new algorithm TFA (Mining Traversal Patterns Using Filtering Approaches) to discover the traversal patterns. The one of characters of TFA algorithm is the adjacency filtration mechanisms. TFA can effectively reduce a huge number of unnecessary continuity subsequences via the adjacency filtration mechanisms. The process of generating continuity subsequences is very simple, and the algorithm does not generate candidate sets, and then finishes the mining task. However, if we only consider the traversal path factor, the degree of accuracy is not enough. Some researchers later propose the correlation about traversal path and purchasing merchandises, and then increase the accuracy of patterns. Therefore, we aim at the subject to propose a new algorithm CFA (Mining the Correlation Using Filtering Approaches). The processes of CFA is combined the TFA algorithm with the EFI algorithm. CFA can reduce a huge number of unnecessary continuity subsequences and itemsets via the filtration mechanisms, and then get all of frequent combination patterns quickly. Besides, when the algorithms deal with very large databases, we do not modify the mining process of TFA or CFA, and then perform the mining task. In addition, the algorithms also use the method of dividing database, and the algorithm only perform each sub-database three times I/O processes, and then finish the mining task.
In the experimental, we aim at each parameter to compare the performance of algorithms. From the analysis of various results, we can obviously discover that our proposed algorithms have a better the performance and the utility rate of memory.
Advisory Committee Jen-peng Huang - advisor

Files No Any Full Text File.
Date of Submission 2006-07-23