ABSTRACT

Now-a-days, in many applications, data mining has become an essential part. It can be observed that many business deals depend on the statistical data which can be easily done by data mining algorithms. Data mining has a vital role in many of the applications like market-basket analysis, bio-technology field etc. In data mining, frequent itemsets also play an important role which is used to identify the correlations among the fields of database. This thesis proposes an approach in examining user interest, based on the web pages accessed. It delineates how clustering these clients into interest groups helps in customized servicing and how the service that needs to be provided is configured. The implementation aspects of web client clustering using prefetching and socket clone methods are presented. Then, the utilization of various clusters is measured based on the number of accesses on the clusters by various clients.

HBMFI-LP algorithm which uses hashing technology to store the database in vertical data format is described in this thesis. To avoid hash collisions, linear probing technique is utilized. The proposed algorithm generates the exact set of maximal frequent itemsets directly by removing all the non-maximal itemsets. The proposed algorithm is compared with the recently developed MAFIA algorithm. It proves to be better than MAFIA algorithm.
Also, a new algorithm is presented in this thesis, which associates a weight to each item in the transaction database, depending on the significance of the corresponding item. Weighted support is calculated using the weight and the frequency of occurrence of the item in the transactions. This weighted support is used to find the frequent itemsets. This algorithm partitions the database among ‘N’ processors and generates closed frequent itemsets in parallel without losing any interesting and significant itemsets. The parallel algorithm used minimizes communication by exchanging only weighted supports among the processors. The performance of the proposed algorithm is compared to other similar algorithms in terms of scale up, speedup and sizeup.

One of the most frequently used data mining algorithms is sequential pattern mining. Normally, sequential pattern mining will be carried out based on the minimum support value. If the support alone is considered, there is a chance of obtaining spurious results or the interesting results may be lost. So, in this thesis, a new sequential pattern mining method, based on closed sequential patterns, is proposed. The highlight of this method is that it considers the grade of each item along with the support by generating frequent itemsets without losing any interesting patterns and also dispenses with any uninteresting patterns.