SUMMARY AND FUTURE WORK

6.1 Summary of Research Contributions

6.2 Future Research Directions
SUMMARY AND FUTURE WORK

This chapter summarizes the work and discusses the future directions of this research.

6.1 SUMMARY OF RESEARCH CONTRIBUTIONS

The major contribution of this research work is a framework for frequent pattern generation from web logs. The important contributions to reinforce are listed below:

- An online incremental learning method using standard deviation (DS-OILSD) for web log is proposed to identify session boundaries dynamically by measuring the change of information in the sequence of requests.

- Enhanced Fuzzy Apriori Algorithm (EFAA) is proposed to exploit the frequent itemset count technique, in which an optimized minimum support value is calculated using fuzzy membership function to reduce number of the infrequent itemset.

- Frequency Based Ranking Method (FBRM) is developed for ranking the user preferred queries by calculating the weight based on the frequency of occurrence of the queries.

- Modified Density Based Clustering (MDBC) is proposed to cluster web users with common interest.

Summary of various chapters with the research findings are outlined below:

Preprocessing Using DS-OILSD: The initial stage of the work presents a novel dynamic log session identification method, which is based on an online incremental learning algorithm. It is different from existing methods since it basically assumes that all the databases query patterns to be dynamic. An information theoretic approach is used to identify session boundaries dynamically by measuring the change of information in the sequence of requests. The proposed Dynamic Session-Online Incremental Learning based on Standard Deviation (DS-OILSD) is a statistical probabilistic method which does not depend on the time based session identification, instead the dynamic web session identification is automatically done based on the standard deviation threshold value and automatic tuning of the web log session. Also, MAP estimation is
done for dynamic changing adaptation of query types based on the dynamic weight changes in the language model. The MAP estimation adjusts the parameters of DS-OILSD method capably, and keeps away from sharp changes. The method proves to be considerably better than the existing n-gram model designed for dynamic log session identification.

To determine the accuracy and the effectiveness of the DS-OILSD method and the existing statistical n-gram analysis model, the experimentation work is conducted on four web log benchmark datasets. DS-OILSD is found to be effective in solving the parameter range problem by proposing a standard deviation for automatic threshold. The entire training data samples are categorized into three types. Learning is introduced for training samples under labeled, semi labeled and unlabeled samples.

The results are analyzed based on the performance measures namely, F-measure and accuracy to evaluate the performance of dynamic log session identification. In case of LLD-SD which produces an accuracy of 84.00% for labeled samples, and there is an approximate increase of 4.0% and 14% accuracy when compared to LUD-E and timeout results for AOL Web Logs. It may be concluded that the proposed LSD-SD produces higher accuracy for AOL Web Logs when compared to other dataset samples.

A comparison of existing methods with the proposed LUD-SD, LLD-SD and LSD-SD methods reveals that there is significant improvement in performance measures such as precision, recall, F-measure and accuracy.

**Pattern Discovery Using EFAA:** In the second stage of the work, the Enhanced Fuzzy Apriori Algorithm (EFAA) is proposed for solving the pattern discovery problem in web log database. The goal is to overcome the problems of Apriori algorithm by exploiting the frequent itemset count technique. EFAA is a fuzzy based rule pruning algorithm in which itemset pruning is used to reduce the number of candidate itemsets in web log data. Itemsets are pruned using Fuzzy Intersection Pruning (FIP) to enhance pattern discovery results which reduces the number of candidate items in the frequent itemset mining. Fuzzy Automated Support Confidence Pruning (FASCP) technique arranges the rules in ascending order of support and confidence thresholds which is found from FAS for association rules generation.
In case of Apache web server log dataset, the proposed EFAA achieves 88.8% accuracy and there is an approximate increase of 5% and 13.8% when compared to the existing MAA and AA methods respectively. The experimental results obtained reveals that proposed EFAA outperforms existing techniques in terms of precision, recall, F-measure, accuracy and memory.

**Pattern Analysis Using FBRM And MDBC:** In the third stage of the work, Frequency Based Ranking Method (FBRM) is developed for ranking user preferred queries by calculating the weight based on the frequency of occurrence of the queries. The proposed personalized FBRM ranks the user query and provides the personalized or preferred results by reweighting the relevant results in accordance with user’s interest. For this purpose, here the user preferred query frequent pattern score is calculated and normalized by the summation of the edit distance of each frequent pattern in the dataset to the closest one in the top result. The larger the weight value of the score, the more representative is the result list.

Proposed FBRM algorithm achieves 90% accuracy which is 21.67% and 14.33% higher when compared to the existing PRA and RMAF methods respectively for Apache web server log dataset. A comparison of FBRM with the existing methods reveals that there is significant improvement in performance in terms of precision, recall, F-measure, and accuracy.

As a final contribution, Modified Density Based Clustering (MDBC) is proposed for carrying out clustering of user preferred pattern data. The ranked user preferred patterns are considered as datapoints for clustering. The common DBSCAN clustering techniques have the problem of border objects. Since the border points at times don’t take significant datapoints into consideration, here the MDBC is proposed to solve this problem that usually considers the datapoints from patterns dynamically during the clustering process.

The proposed MDBC algorithm achieves a completeness of 0.94, which is 0.25 and 0.09 higher when compared to K-means clustering and FCM methods respectively for Apache web server log dataset samples. A comparison of MDBC with the existing methods reveals that there is significant improvement in performance in terms of ARI, MI, homogeneity and completeness.
6.2 FUTURE RESEARCH DIRECTIONS

The present work can be extended and applied to DNA sequence analysis. Automatic threshold can be experimented using other methods. In association rule mining on web log data, there is a need to identify specific interestingness measures that address the web page issues like inter-connectedness through the website link structure. Future work includes investigating the effectiveness of the proposed approach on other data sets, and for other web usage mining problems, such as sequential pattern mining. Directions for future work include investigating the effect of coarse grained or noisy feedback on learning performance, learning preferences over sets of patterns instead of individual patterns, and shifting from the pool based active learning to query synthesis, i.e. directly mining patterns for queries. Though the research contributions have provided better accuracy, the computational complexity can be studied and improved. An in-depth research is still needed on several critical issues so as to bring about a deep impact in web usage mining applications.