CHAPTER 9

CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

The R*-tree is an improved model of the variants of R-tree. It forms the basis of all the multidimensional indexing methods based on data partitioning that are in existence today. The overall objective of this research is to study and ameliorate the R*-tree principles that would assist in improved multidimensional query processing. The conclusions of the various ameliorations carried out are given below.

9.1 Conclusions

The first principle explored was the estimation of the number of node accesses in an R*-tree for window queries. The existing models assumed uniform distribution of data and incorporated only the length parameters of the MBRs for estimation. A new model was developed that accommodates any distribution of data. The model also incorporates the location parameters of the MBRs along with length parameters of the MBRs for estimation. Experiments were conducted and the results show that the newly developed model outperforms the existing models in a significant way.

The second principle explored was the estimation of the number of node accesses in an R*-tree for kNNQs. The existing methods do not differentiate smaller value of $k$ from larger value of $k$. The experimentation showed that the methods were very inaccurate in estimating the node accesses for smaller value of $k$. The reason for the inaccuracy was explored as part of this research. It was found that the location of the query point influenced the accuracy of the estimation for smaller value of $k$. A new method for estimation was developed that eliminates the deficiencies of the existing models for smaller value of $k$ irrespective of the location of the query point. Experimental results uphold the superiority of the new method.

The third amelioration was for DaTQs. The existing methods in the literature handled the 2-dimensional data efficiently for answering DaTQs. They were seldom
scalable to higher dimensions. Hence, a new improved representation scheme and an algorithm to retrieve multidimensional objects based on DaTRs using R*-trees were developed. Subsequently, a model for the estimation of the number of node accesses in the R*-tree for the given DaTQs was also developed. The results of the experiments established the applicability, scalability and accuracy of the MDTQ methods.

The next exploration delved into the possibility of improving the R*-tree performance by efficiently organizing it in secondary storage devices. As a consequence, a new method was proposed that efficiently clusters the nodes of the R*-tree into input-output units of the hard disk within the constraint that the independence between the logical and physical organization of the R*-tree should be preserved. Moreover, to preserve the structural and functional properties of R*-tree at any point in the process of clustering, a concept called controlled duplication had been introduced. Experimental results show a significant improvement on various parameters.

Re-indexing the R*-tree formed the next investigation avenue for the improvement of the performance of the R*-tree. While research literature seldom emphasized on this, the commercial literature projected this as an invaluable exercise to maintain the peak performance of the database systems. A process called recursive re-grouping was developed to re-index the R*-tree. Comparisons were made with existing methods for re-indexing. Recursive re-grouping outperforms the existing methods.

The last principle explored to enhance the performance of the R*-tree was the usage of MBPs for representing multidimensional objects instead of MBRs and MBSs. Usage of polygons reduced the dead space and this in turn improved the performance of the R*-tree. Experimental results make it evident.

The above said ameliorations improve the multidimensional database performance both individually and collectively.

9.2 Future Research Directions

Apart from the above said principles of R*-tree, a lot of other avenues still unfold for research. Besides the window queries, NNQs and DaTQs a spectrum of other
query types are in existence and can be explored. One of the major query category is the join query which decides the cost of queries that involve multiple R*-trees.

Another interesting ramification that can be envisaged for the R*-tree indexing method, is its construction for fragmented and distributed tables. The challenge here is the handling of the R*-tree structure for replicas of data and the updating of replicas synchronously.

It is already well established through this research work that the characteristics of data influence the estimation models. Apart from the location and length parameters, other parameters may be explored for the inclusion into the models for better estimations.