**HPPQ: A Parallel Package Queries Processing Approach for**

**Large-Scale Data**

**Abstract:** A lot of scholars have focused on developing effective techniques for package queries, and a lot of excellent approaches have been proposed. Unfortunately, most of the existing methods focus on a small volume of data. The rapid increase in data volume means that traditional methods of package queries find it difficult to meetthe increasing requirements. To solve this problem, a novel optimization method of package queries (HPPQ) isproposed in this paper. First, the data is preprocessed into regions. Data preprocessing segments the dataset intomultiple subsets and the centroid of the subsets is used for package queries, this effectively reduces the volume of candidate results. Furthermore, an efficient heuristic algorithm is proposed (namely IPOL-HS) based on the preprocessing results. This improves the quality of the candidate results in the iterative stage and improves the convergence rate of the heuristic algorithm. Finally, a strategy called HPR is proposed, which relies on a greedy algorithm and parallel processing to accelerate the rate of query. The experimental results show that our method can significantly reduce time consumption compared with existing methods.

**Key words:** package queries; heuristic algorithms; parallel processing; opposition-based learning.

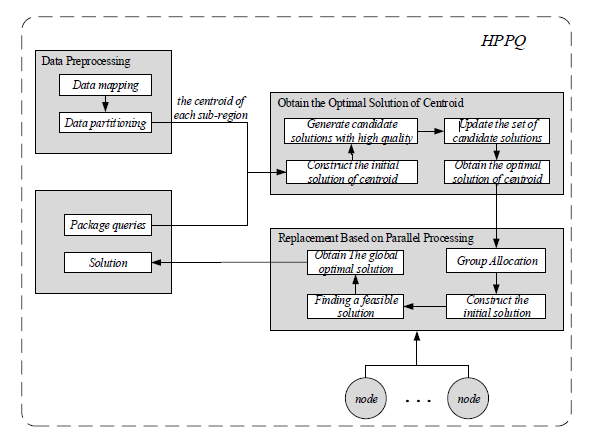
**Existing System:** Heuristic algorithm do not rely on gradient information and provide excellent performance in the problem of package queries. Compared with the exact methods, heuristic algorithms may get a suboptimal solution, but they greatly reduce the run time. However, with the rapid growth in data volume, heuristic algorithms used to solve package queries also find it difficult to meet the increasing demands of efficiency. Some scholars have used divide-and conquer algorithms to divide the problem into multiple sub-problems and improve query efficiency. But these use exact algorithms to solve the sub problems, which affect the run time to a certain extent.

**Disadvantages:** These existing algorithms are divided into several categories: exact algorithms, heuristic algorithms, and divide-and-conquer algorithms.All Operations not in single place.

**Proposed System:** It is necessary to design an efficient method of package queries for large volumes of data. In this, we present a method called HPPQ (Heuristic Parallel Package Queries), which is based on heuristic and divide-and-conquer strategies. It optimizes the method of package queries mainly through two aspects: improving the quality of the candidate solutions and accelerating the speed of query. To address these two aspects, we propose the IPOL-HS algorithm and the HPR strategy respectively.

Advantages: By Compbining heuristics,divide and conquer strategies we put all the operations at one place.

**Architectuire:**

****

**Modules:**

**Data Preprocessing**:

**Obtain Optimal Solution of Centroid**

**Replacement based Parallel Processing**

**Data Preprocessing**: Data preprocessing divides similar objects into the same sub-regions, and when dealing with package queries, the data objects in each sub-region are treated as a whole. Data preprocessing mainly has two parts: data mapping and data partitioning.

**Obtain Optimal Solution of Centroid:** Here, we propose a heuristic algorithm based on Improved Partial Opposition-based Learning (namely IPOL-HS) to obtain the optimal solution for the centroid.

**Replacement based Parallel Processing:** Each replacement program needs to be calculated and if replacing the centroids with real data objects this needs to happen sequentially. This takes a lot of run time and produces more combinations. So, in this section a strategy called HPR, based on a greedy algorithm and parallel processing, is proposed to accelerate the query speed. The HPR strategy divides into two parts: group allocation and parallel processing. First, we designed an allocation algorithm to balance the node load and improve parallelism. Then, using the greedy strategy and parallel processing, we effectively replaced the centroids with data objects of corresponding subregions to get the final solution.

# H/W System Configuration:-

# Processor - Pentium –IV

Speed - 3.5 GHz

RAM - 4GB (min)

Hard Disk - 500GB

# S/W System Configuration:-

* Operating System :Windows 8 or above
* Application Server : IIS 7.0
* Front End : ASP.Net (.Net 2015)
* Database : SQL SERVER 2014
* Database Connectivity : ADO.Net.

**Future Scope:** In the future, to further improve the convergence rate of the algorithm, the effect of different periods on the improved partial opposition operation probability should be considered. In addition, even though the method proposed in this paper reduces the run time required for package queries in the large-scale data environment to some extent, it still has the potential for further optimization. The parallel strategy mainly adopts parallel computing to reduce the run time, but there is no parallel strategy for obtaining the optimal solution of the centroid and no parallel replacement scheme for multiple sub-problems

**Conclusion:**

As a response to the requirements of large-scale data, we proposed a novel method for package queries combining heuristic methods and the divideand-conquer strategy, namely HPPQ. The method was found to reduce the number of candidate results by data preprocessing.