

How to Cite:

Babu, V. N., & Babu, A. S. (2022). A hybrid pre-local skyline and post-global skyline filter Based optimized GPMRS framework using multiple mappers and multiple reducers framework. *International Journal of Health Sciences*, 6(S1), 9206–9224. <https://doi.org/10.53730/ijhs.v6nS1.7088>

A hybrid pre-local skyline and post-global skyline filter Based optimized GPMRS framework using multiple mappers and multiple reducers framework

V. Narendra Babu

Research Scholar, Department of Computer Science and Engineering, JNTUK, Kakinada-533001, Andhra Pradesh, India

Email: narendrababu.v06@gmail.com

A. Suresh Babu

Associate Professor, Department of Computer Science and Engineering, JNTUA, Anantapuramu-515001, Andhra Pradesh, India

Email: asureshjntu@gmail.com

Abstract---Due to the rapid and exponential development of spatial databases, traditional spatial relationships and its co-locating measures have become inaccurate and inefficient. Many different spatial co-locating algorithms have been tried to detect nearest neighbor patterns within limited datasets in order to evaluate their spatial patterns. Finding co-locating patterns using skyline computational models is one of the active research areas on large spatial databases. However, most of the conventional spatial skyline co-locating models require high computational memory and time on large spatial databases. Another limitation with traditional skyline models is finding the filtered candidate sets in the local and skyline computational models. In order to improve the accuracy of the traditional skyline models, a hybrid parallel filter based skyline model is developed on the large spatial databases to optimize data filtering along with ranking process. Experimental results describes the efficiency of proposed model on conventional skyline computational models on large databases.

Keywords---Map reduce Skyline processing, Spatial data minings Hadoop.

Introduction

Spatial data mining techniques provide the key to implicit knowledge of events, spatial relationship and Spatial data. It can influence the selection and judgment of those instances. Association rule mining technique stresses that frequent patterns in transactional data mining is very important. Transactions cannot be automatically determined by means of spatial data. There are exchanges and collaborations among different transactions at the same region. For spatial data security detection, it is also essential to detect explicit sets of spatial data out of spatial data. Spatial techniques like networking, reference features, and windows-oriented to identify transaction-centric data[1].

It means that a special technique can be used for deriving out features from the whole pattern set. The biggest drawbacks due to the exponential growth of pattern repositories are to require experts to retrieve useful information and records from a large amount of distributed pattern repositories. Distributed data mining features that provide an opportunity to translate large volumes of high dimensional data into smaller and manageable datasets with sufficient resources. Big Data entails big amounts of relevant data, and these large volumes of data are essential to remain relevant quickly. Big Data is the set of information relating to human behaviours, institutions, spatial classes, industries, economic, and social issues. In a Data Warehouse, data are assumed to be precise, complete, clean and consistent. However, when dealing with big data, the data itself is unproven, inaccurate, and imprecise in nature so data must be processed through virtual machines. It doesn't matter what specifications is your system. Virtual machines that run this type of applications are easily configurable and can be integrated with a variety of cloud storage systems such as Dropbox and Google Drive to improve the performance of distributed computing. There are plenty of available cloud computing systems such as Google Compute Engine, Windows Azure, Amazon EC2 and Rackspace[2].

This is basically a set of applications that are planning for construction of groups. This is called small clique, where spatial instances are closely related with each other, making the complete clique. Sometimes two separate group members from within a corporation or group have different characteristics. They are both located near each other. At a single level, a clique consists of a number of members with high cohesion among them. Still other scholars have discovered that co-location mining algorithm usually identifies features having frequent co-located instances[3].

Detecting spatial co-location is a tough task and very challenging. By analyzing spatial relationships, co-location emphasizes the significance of spatial information in the design of the environment. Places where many spatial objects gather in a locality can be defined as spatial co-location patterns. This information can be used to predict weather, make judgments on political actions, and so on. Thus, co-location patterns can be efficiently identified by means of spatial statistics and spatial data mining methods. These spatial techniques depend on data measure of spatial correlation. In this analysis, the information is broken up into several elements, and analyzed with one another. Spatial data

mining method bases on two other algorithms: association rule mining and pattern clustering[4].

Maximal cliques concept is basically a top-down graph-based frequent pattern mining framework[5]. The whole working process of this approach can be divided into two sub-phases:- those are:- clique determination and candidate refinement. The following are steps of maximal clique generation process:

1. In the process of clique determination initially mining of each and every frequent mining pattern of size s is performed. s -item pattern graph is built for database.
2. Here in this step a major issue arises which is:- the mining problem due to huge maximal frequent patterns. It becomes very complicated to determine huge maximal cliques.
3. The pattern graph size is depends upon the number of frequent items present in database.
4. Determination of maximal cliques is completely independent of database size and pattern set size. This enhances the overall processing time of the algorithm.

In this case interactions among different sets of vertices of graphs are detected those lie in the same label. Pattern is represented as a sub-graph which is constructed by using many core groups of strongly interlinked vertices. All considered vertices must lie in a particular same label. Pattern is defined as a group of cliques which obey certain constraints for least number of individual cliques[6]. These cliques are of least sizes and present on least number of labels by all vertices. Such types of patterns are known as Maximal Homogeneous Clique Set. Additionally, it can be stated that the above proposed technique can be implemented for very large real graphs. Also, scalability is analysed with respect to various graph characteristics for synthetic datasets. Most traditional algorithms operate with restricted transaction records on transactional databases. Most real-time applications, however, use the method of discretization to convert real data into transactional format. But the function of discretization can lead to data loss or a high rate of error for frequent pattern mining.

Large numbers of association rules are produced in large databases (both relevant and non-relevant). Support and trust are two significant measures for assessing interest. Some other approaches for the evaluation of interest are Chi-square, Laplace, and Correlation coefficient. By implementing subjective metrics such as lift and conditional dependency, it is possible to reduce the number of association rules.

The quality of rules, along with the number of association rules, is also an essential factor. This also causes the association rule algorithms to have a major impact. Association mining algorithms are first executed and their configuration of parameters then takes place. Hence, it is necessary to define the parameters before execution. It is considered that all association rules exceeding the basic statistical measures are valid. This selection is based on the prior knowledge and experience of the users in the domain field specified. Algorithms having minimum or no parameters at all can solve this problem.

Spatio-temporal data is currently increasing day by day. Let us take a simple instance of spatio-temporal data, data extracted from the GPS feed. Points are typically used to reflect the spatial extent of the trajectory. Trajectories in the region help represent moving objects[7]. The size, shape and location of objects change dynamically in the case of moving objects. Complex polygon-based geometries are responsible for building trajectories in the region. Most of the spatiotemporal datasets include the process of spatiotemporal overlap. Such overlaps, for instance, can be seen in solar events.

Here, the connections between several solar events are analyzed in this research strategy. In addition, it also thoroughly studies how these events will affect the earth. The effects of the Sun and its radiation are also analyzed using the STCOPs approach. Such radiation mostly affects space travel, power grids, and certain communication devices. In order to produce candidates, all pre-existing STCOP mining approaches implement the Apriori technique. A pruning mechanism is also part of this approach. Nevertheless, this method has a significant drawback, that is:- with the rise of event types, huge numbers of candidate patterns are also generated. To evaluate the interest of a huge set of candidates will be very expensive. There is, therefore, a need for such an approach that can determine STCOPs without depending on the process of generating candidates.

Let's consider a spatiotemporal transaction dataset in which spatiotemporal co-occurrence of multiple region trajectories is generally represented by each individual transaction. In order to produce a group of proto-patterns, another traditional FP-growth approach is used here. Original STCOPs are determined from proto-patterns in the subsequent stage with the help of significance and prevalence measures. The above is a filtering process that is in charge of discarding unnecessary patterns.

A top-down graph-based frequent pattern mining framework[8] is basically the concept of maximum cliques. It is possible to divide the entire working process of this approach into two sub-phases:- these are:- determination of the clique and refinement of the candidate. The following are steps in the process of maximum clique generation:

1. Initial mining of each and every frequent mining pattern of size is performed in the process of clique determination. For the database, an s-item pattern graph is built.
2. A major problem arises here in this step, which is:- the mining issue due to enormous maximum frequent patterns. The determination of huge maximum cliques becomes very complex.
3. The size of the pattern graph depends on the number of common items present in the database.
4. Determining maximum cliques is completely independent of the size of the database and the size of the pattern set. This improves the algorithm's overall processing time.

Interactions between different sets of vertices of graphs that lie on the same label are detected in this case. Pattern is portrayed as a sub-graph that is constructed by using strongly interlinked vertices from many core groups. In a specific identical label, all considered vertices must lie. Pattern is defined as a group of

cliques which obey certain constraints for least number of individual cliques. These cliques are of the least size and all vertices are present on the least number of labels. These kinds of patterns are referred to as the Maximal Homogeneous Clique Set.

To retrieve maximum homogeneous clique sets, a new approach is introduced here. This approach is also responsible for the recovery by means of the pruning mechanism. In order to demonstrate safety as well as correctness of the above approach[9], a series of experiments are carried out. In real datasets of two kinds, all structures relating to MHCSs are present.

In addition, for very large real graphs, it can be stated that the above suggested technique can be implemented. In relation to different graph characteristics for synthetic datasets, scalability is also analyzed. On the transaction dataset, a new approach was developed to resolve the rare pattern mining issue using the multiple support measure(MSM) algorithms. This algorithm is known as the multiple support-based fuzzy weight algorithm (FWMSM). This approach takes two important measures (support and significance) into account at a time when this approach is advanced compared to the traditional MSM algorithm previously developed. The measure of significance assigns various levels of importance to each item selected. Only a few attributes of all are quantitative in nature, while rest is not. Basically, the Fuzzy-based technique includes fuzzy set theory along with various attributes for computing complex instances. To evaluate the importance of items, fuzzy-based computation is therefore implemented. [10] has developed a new algorithm, which in nature is very scalable. Their suggested system is trie-based. They generate interesting patterns and correlations via interval support and are also evaluated between these patterns. It is possible to define representative association rules as the minimum set of rules from which all association rules are produced. All previously developed algorithms accept as input total frequent item sets. Developed a new framework to address all issues relating to representative association mining rules. [11] New methods for detecting association rules and Weighted Least association rules have been introduced (WELAR). Their strategy was called Weighted Support Association Rules (WSAR).

Frequent patterns of spatial co-location are usually described as a collection of several spatial features with instances that are very closely located. For instance, schools and parks are mostly located close to each other. This can also be regarded as the city's frequent co-location pattern. With the rapid growth of mobile technology and the development of GPS-enabled mobiles, it is easy to collect a huge amount of spatial data in no time. It is sometimes very important to identify other living organisms that are closely located in a specific geographical region. There are a huge number of significant spatial co-location pattern mining approach applications. Here, let us take an example, all service providers have an objective to improve their services in a near region. As the amount of spatial data increases exponentially day by day, it would be very costly and time-consuming to mine all these data for them. In the case of spatial large data mining, all traditional mining algorithms that work very efficiently in the case of conventional data become inefficient. Many researchers have therefore attempted to develop an advanced and more efficient approach to spatial co-location mining in order to meet the growing needs of such spatial data.

The join-based approach, partial-join approach and join-less approach are some effective approaches. There is a similar joint-based approach to the Apriori algorithm. It wastes a huge amount of valuable time in the entire database scanning process. It also looks for the most suitable and interesting joining patterns. On the other hand, in order to divide them into two separate parts, the partial-join approach usually considers the distribution of objects. These portions are referred to as intraX and interX. There is only theoretical significance to this concept. But in the case of real world scenarios, dividing these objects is very complicated. In addition to the two approaches mentioned above, there is another approach known as the join-less approach. As the name indicates, no joining operation is required for this method. Instead, it checks frequently for several important structures.

Related Work

The full skyline can be found by finding next NN's (nearest neighbours) recursively [12] in the non-dominated regions. With an R-Tree indexed data set, BBS [13] offers minimum I / O costs. Yet R-Tree suffers from the curse of the dilemma of dimensionality. It has not demonstrated how high dimensional data can be treated. In FASTSKY, the efficiency of skyline computation with high dimensional data was improved. To index low-dimensional data and high-dimensional data on PODs (Partial Order Domains), a stratification technique is used along with two index structures, namely Stratified R-tree (SR-tree) and Stratified MinMaxtreaps (SM-treaps). Dynamic Query means the execution space of the query depends on the query itself. Query processing costs would be proportional to the number of peers, the key downside of solely decentralised architectures. Hybrid centralised indexing systems[14] will be one server and a centralised index framework for monitoring peer-to-peer interactions, but without server intervention, data sharing will take place between peers. This seems to be successful, but single-point failure is a big downside of this. In addition, this becomes an obstruction for the system, with a high query rate in the large network. If only super-peers are considered, they form a strictly distributed P2P structure. They exhibit some kind of ambiguity as the data obtained from various sources and from distributed locations, which comes with imprecise measurements[15]. Global conditions affect decision-making in each domain. Decisions made on the basis of the full data available are more precise than restricted data. Therefore, effective forecasting involves consideration of each and every piece of data. Most applications require the collection of data from distributed sites and the execution of user requests on the basis of a collective view of data from all local sites, a costly effort[16]. Communication between all remote sites and central servers is also expensive in many application domains due to the vast quantities of data stored at each local site and the latency faced by the network, as well as the economic costs associated with data sharing and communication. In addition, the handling of user requests is a time-consuming operation.

A group of algorithms running on a centralised CPU forms the first subclass. The Nearest Neighbor NN algorithm[17] uses R-Trees as an index structure. The second subclass groups the algorithms running on a distributed CPU while the

algorithms of the third subclass operate on a multi-core architecture. By considering the Euclidean distance, NN finds the nearest point p to the origin O , and then divides the space with respect to that point along each axis. If we have a 2-dimensional space, the two-axis projection will create a rectangle where the main heads are p and O . It will delete the points that do not belong to the rectangle. The processes of locating the nearest point to the origin and projection will be repeated more times, before the points of the skyline are discovered. The projection will create a hypercube if we are in front of a 3-dimensional space. For progressive applications, NN is well suited. It operates efficiently in a two-dimensional vacuum, but when the dimension exceeds four, it suffers from problems of duplication during elimination[18]. The BBS algorithm of Branch and Bound Skyline calculates nested spaces and executes NN recursively. It terminates when no point inside the corresponding space is obtained by the corresponding nearest neighbour query. BBS uses R-trees; thus, it suffers from problems with the same NN, such as the large number of nested spaces to be computed. [20] has combined BNL and presorting for SFS (Sorted First Skyline) to solve the problem of CPU occupation. Experiments have shown that memory access and CPU occupancy are decreased by this combination. A new framework has been proposed, which they call Skycube. In order to create the skyline according to their two (Bottom-Up Skycube) and (Top-Down Skycube) algorithms, this structure consists of calculating all possible variations of the lattice. Several algorithms have been developed for distributed applications in the second sub-family. For example, [21] is an algorithm which uses views to compute the skyline in a distributed environment. These store the pre-calculated skyline outcome of an input subspace. The MapReduce Grid Partitioning based Multiple Reducer Skyline Computation MR-GPMRS algorithm was introduced to exploit new platforms. This study presents the effects of parallelization on skyline computing on various datasets by varying the dimensionality, cardinality and execution buffers. The algorithms published in [22] were compared with MR-GPMRS. [23] provides a MapReduce structure for the computation of independent regions after creating a query's convex hull. Imprecise data still remains when mining unknown data for user-generated queries. The final skyline is the union of all skyline regions. Such identification inaccuracy is usually mitigated using the technique of skyline querying, which produces skyline points relevant to user end question. The technique of querying the skyline is intended to provide better points for the skyline, even when the database is unclear or incomplete. If the technique of querying the skyline fails to load the skyline points, it correctly results in meaningless skyline results. The customised skyline queries are provided to classify interesting skyline points based on user preference and retrieval scale. In addition, as a dynamic search over subspaces based on user preferences, ranking of personalised skylines was performed. In order to minimise storage space, an algorithm was proposed to submit the compressed structure and to study the construction of interleave dynamic compressed skycube for particular scenarios. In addition, two algorithms are proposed to provide support for user preference based on equivalence preference and to increase model performance. A skyline algorithm has been proposed to minimise the overhead in the processing of the tuples on fully and partially ordered domains. The proposed method uses the FAST-SKY algorithm to solve the problem of overhead and weak comparison of dominance arising from high dimensionality, because the data sets are huge. Topological sorting order and new index structures are ensured by the

assessment of the radical skyline carried out in their work. With index data using the stratification technique and two new index structures suggested, a partially ordered domain has been adopted. This includes: stratified R-trees for low and high-dimensional results, and stratified MinMax heaps, respectively. Thus, the quick comparison of dominance and reduction of dimensionality is accomplished with reporting queries rather than dominance queries. The multi-objective optimization of machine learning has been investigated. Using sub-space skyline queries, this method retrieves the best skyline points in complete data space. The operator of the skyline commutes with a selection operator which expresses the user's preference for the skylines.

[24] developed an advanced and efficient framework. Utilities are initially integrated into the aforementioned framework in this piece of research work. In the above proposed technique, a number of concepts relating to high utility co-location patterns are described. In order to resolve all of them effectively, they identified all the problems related to high utility co-locations and presented a new advanced technique. In the above mentioned case, the downward closure property can not be introduced. Hence, in order to prune candidates, an extended pattern utility ratio is presented. Experimental evaluation proves the efficiency of the EPA algorithm proposed above. When a pattern c has low utility co-location, the pruning of its supersets depends directly on EPUR in the EPA approach (c). Hence, c with low co-location will result in high vss value and also high EPUR value (c). The experimental results show that EPA is a technique for size-order traversal. When it moves from low size patterns to high size patterns, the pruning effect becomes less obvious. In order to address the above-mentioned problem, further research work must be carried out in the future. More emphasis on low utility patterns should be given.

Y. An incremental topological spatial association rule mining approach, integrated with a clustering method, was proposed by [25]. The above approach accepts and implements geographical datasets in a probabilistic way. Finally, spatial clustering is carried out according to the rules of spatial association that have been generated. The evaluation process shows that the above proposed approach achieves 83.14 percent accuracy, which is much better than that of other traditional approaches. Here, in order to filter the generated spatial rules, the probability threshold and relationship threshold play a vital role. For their experiment, they considered three different datasets.

[26] introduced a new method in order to mine approximate patterns in multi-graph collection. They referred to their newly established strategy as AMgMiner. In the area of data modeling using multi-graphs, a large number of studies have been carried out. Multi-graphs for many practical problems are regarded as an efficient solution. These methods are effective for addressing problems of classification. Multi-graphs also play a very important role in the pattern mining process. There is a single approach to date that supports pattern mining using multi-graph methodology. It never directly evaluates approximate patterns from multi-graphs. Therefore, the technique mentioned is not efficient and with time it has become obsolete. In this research work, emphasis is placed on developing a pattern mining approach that, with the aid of multi-graphs, can directly mine patterns. The traditional canonical form is slightly modified and extended here,

based on the depth-first search technique. In the evaluation phase, a series of experiments are conducted on different public standard databases. In performance terms, the presented algorithm outperforms the pre-existing algorithm. During run time, the performance is analysed. Mining of maximum or closed patterns can be effectively done in the future by extending this presented method.

Indirect association is a category of infrequent pattern that is responsible for reducing the number of infrequent patterns that are uninteresting. In our survey, we have described various schemes for association mining (including conventional schemes as well as hybrid schemes). By taking into account temporal characteristics and bit vectors, many researchers have attempted to improve temporal efficiency. Later, in order to solve the problems of the pre-existing ones and also to improve the effectiveness of those algorithms, other advanced approaches are introduced. Several attempts were made to enhance the scans and memory consumption. They presented a new technique for the mining of association rules. Multi-time interval sequential patterns have been developed which are responsible for detecting the time intervals between item sets. In order to resolve the issues of the traditional techniques, two algorithms were introduced.

[27] introduced an updated technique for the rules relating to mining temporal association. Many essential rules may be discarded due to very high minimum support, according to the traditional association rule mining scheme. But a lot of trivial rules are exploited in the case of very low minimum support. To isolate useful rules from the rest is very difficult and complex. Mining object sets are only frequent in certain selected regions in the temporal association. An effective temporal association mining scheme has been introduced that automates the process of discovery of association rules. For this scheme, no prior domain knowledge is required. The major components of their theory are semi-intervals and partially-ordained patterns. From sequential patterns and itemset mining, it integrates various algorithms and gives rise to improved semi-interval patterns. The traditional Apriori algorithm was modified and an extended version was proposed, i.e., Mining Periodic Temporary Association Rules (MPTAR). By implementing time expressions in association rules, some problems are detected. They studied and identified these issues and introduced enhanced temporal association mining[28].

A combination of $2^{|I|}$ Item Sets is the search space. In the case of a larger 'I', generating and evaluating support values within a particular time limit is quite impossible or difficult. Let us consider a simple example in which 'I' represents thousands of objects and exceeds the number of atoms in the universe by the number of object sets. We concentrate on generating datasets that participate at least in a single transaction instead of generating such large numbers of itemsets. Research is carried out to optimize the process in order to solve this problem involving a large number of object sets. The process of optimization must contain subsets along with a limitation of maximum size.

For their research, [29] studied and analyzed sparse transaction databases. For the processing of dense databases, the proposed algorithm needs sufficient

memory space. Some techniques are created to solve this problem, resulting in a directed search strategy using the search space given. Set of candidate itemsets and their support values for directed search method results. The size of the total candidate item set must be within the size limit of the main memory space available. In order to decrease the time consumption due to support evaluation, algorithms should generate fewer item sets. Once the method is optimized, only frequent item sets are generated and evaluated. But the above process is very challenging in practice. With the extension of the object set, the support value decreases in most algorithms. Let us consider market basket analysis problem with an association rule $X \rightarrow Y$, where X and Y are represented as items that customer wants to purchase together. There are also items that customers do not want to buy together, these rules are referred to as negative rules of association. Sellers can consider negative association rules in order to study customer purchase patterns. To grow their business with more profits, the sellers also develop different marketing strategies.

In the following period, a new type of mining association rule was introduced and named as the mining substitution rule[30]. The entire process of mining substitution rule is divided into 2 phases, which are:- detection of specific item sets and generation of substitution rule. In the previous part of substitution rule mining, they carried out the Chi-square test together with the correlation coefficient. Different researchers also propose some other techniques to mine negative rules of association. This process is further broken down into three phases, which are:

Detection of positive rules for association.

With the assistance of previously generated rules along with domain knowledge, negative association rules are formed.

- Duplicate rules are discarded by the process of pruning.

Negative association rules are a combination of negative rules on the left hand side and the right hand side. For years, several positive and negative association rule mining algorithms have been created. The problems due to both positive and negative mining association are broken down into:- 1) Detection of both frequent and rare item sets. 2) Discovery of favorable and negative rules of association.

In the basic process of association rule mining, a support-confidence framework is implemented. Some scientists enhance this basic framework and a modified framework is developed which is known as the framework for support-trust-correlation. In the case of association rule mining, there are some problems usually noticed, such as mining of generalized association rule, multi-level association rule, quantitative association rule, fuzzy association rule, periodic association rule and restricted association rule. Among these, some problems are due to frequent patterns, while others are due to rare patterns. Many approaches been developed in order to detect any interesting patterns hidden inside infrequent itemsets.

Another new approach is introduced to address the above-mentioned issues of traditional association rule mining, which is called indirect association rule mining. This approach is responsible for significantly reducing the number of unnecessary, uncommon patterns. Basically, this technique detects negative rules of association without considering negative object sets and domain knowledge. In the traditional support-confidence framework, the inclusion of

correlation measures is extended. This method works in two steps, which are:-1) generation of the association rule and 2) classification. In the initial step, numbers of rules of class association are generated and a correlation of statistical measure is implemented. This correlation value defines, i.e., either positive or negative, the class association rules category. The rules generated are arranged according to their support and trust in the second step. Later, an approach-based mining method for classification will be developed. It is also carried out in two phases, namely the generation and classification of association rules. In the first step, the association rules are produced according to a vertical mining approach from the training dataset. In order to build the classifier, the resulting rules are implemented in the classification algorithm.

Proposed Methodology

Parallel processing will now play a significant role in most of the daily processing aided by computer systems. The phrase "skyline query processing" is not uncommon. Large scale computing by using MapReduce will take a big role in skyline computation. MapReduce could not scale properly due to including unnecessary information. Our proposed Map Reduce - Pre and Post Filter-based Skyline Computation (MR-PPFS). This is to limit the number of search points for MapReduce computation. Due to the growing spatial datasets in the skyline, the computational memory and efficiency is also becoming increasing, when comparing with the spatial datasets. Segmentation has drawn huge attention in recent years due to its space and time complexities over unlimited data. Many methods for analyzing the sequential data have been developed in the recent years. Spatial pattern prediction is really troublesome due to time and cost. Another limitation with the sequential spatial pattern mining models is the produced duplicate vector space for each event vector. The proposed method is implemented using MapReduce on large spatial uncertain datasets. A specific type of k-nearest neighbor mapping class is applied with the hadoop framework.

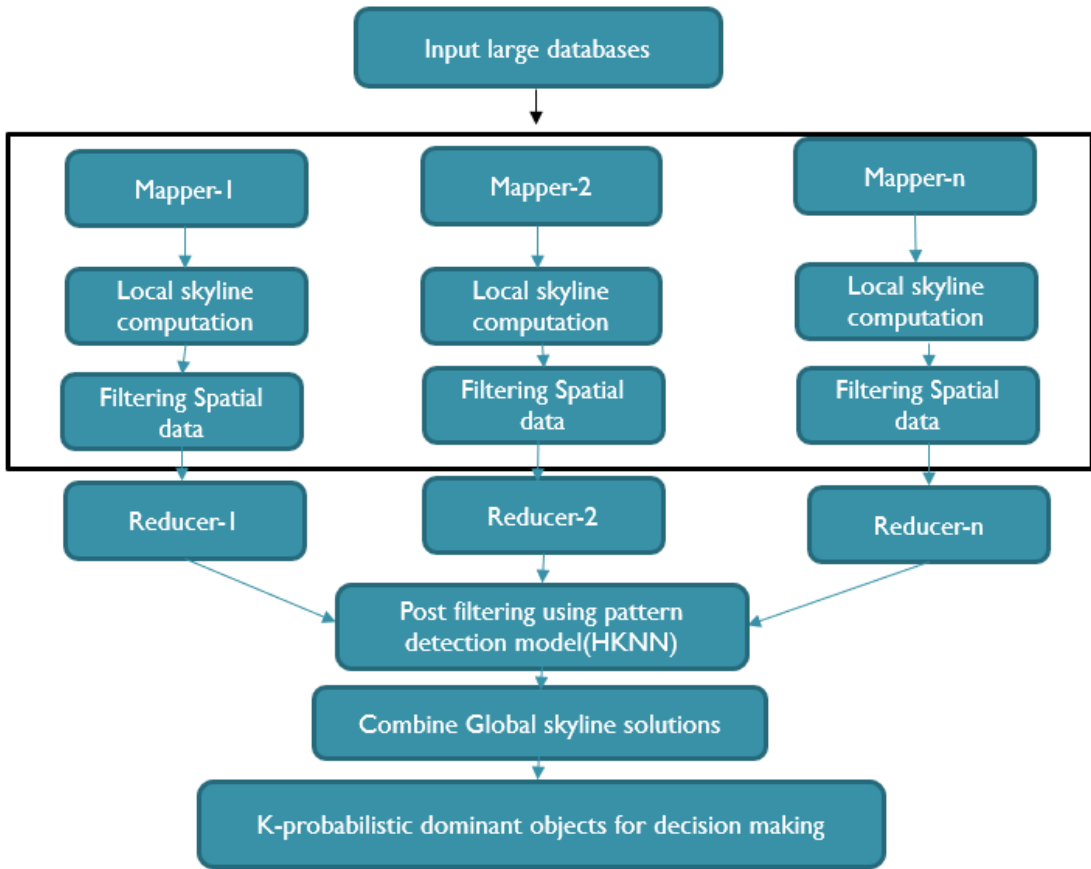


Figure 1: A Hybrid local and global filtering based skyline computation model

In this work, a hybrid parallel local and global skyline computational model is developed using the traditional GPMRS [31] model.

Algorithm:

1: Skyline bit string representation in the grid.:

In the proposed model, each point the input data is represented in bitstring for skyline computation model. The bitstring representation of the skyline point is given as

$$\text{Bit_String}_R[i] = 1, \text{ if } Gp_i \neq \text{null} \\ = 0, \text{ else } \text{----}(1)$$

In this work, each input data is partitioning into k disjoint sets for map reduce framework. Here, each disjoint set is given to the MapReduce framework for local and global skyline computation.

Step 2: Apply GPMRS model to each mapper and reducer framework for local and global skyline computation. Here, in this work, multiple mappers and multiple

reducers are taken to filter the essential key skyline objects using the probabilistic kernel density estimation model.

Probabilistic kernel density clustering algorithm

Input: Spatial datasets, K objects, KNN.

1. Local skyline method for the initial data filtering in the multiple mappers and multiple reducers.
2. Using the KNN model, finding the k nearest local skyline objects in order to filter the large number of candidate skyline objects for the global skyline computation.
3. To each instance in the local skyline objects
do
 For each instance O_i in the KNN objects KNN[]
 Do
 For each instance O_j in IPG[] // where $i \neq j$
 Computing the Chebyshev distance N_m^k on the KNN objects.
 Done
 To each Chebyshev distance objects in the local skyline modelling, find the k nearest objects in the sorted as
 $N_m^k [] = \text{TopKNN}(k)$;
5. Apply local density estimation probability on the filtered local skyline objects.
6. To each reducer in the MR framework
7. do
8. Find the nearest density objects using the proposed probabilistic KNN method.

$$Dist_c = mean^K + \kappa \cdot \sqrt{\frac{1}{N-1} \sum_{i=1}^N (\phi_i^K - mean^K)^2}$$

Here, N is total number of filtered knn objects

ϕ_i^K is the average of the kth nearest neighbour to instance i.

$\phi_i^K = \max_j \{KNN_i(Dist_{ij})\}$, and

$mean^K$ is the average of ϕ_i^K ,

computed as $mean^K = \frac{1}{N} \sum_{i=1}^N \phi_i^K$

Prior estimation probability $= \kappa = \text{Prob}(I, C_k) = \text{Max}\{\text{Prob}(I/C_k); k: k\text{-nearest objects}\}$

Proposed local density estimation is given as

$$PLDE(v_i) = \frac{1}{\eta} e\left(-\frac{\|v_{ij} - Dist_c\|^2}{2\sigma^2}\right) \sum_{j \in KNN_i} \kappa \cdot e\left(-\frac{v_{ij}^2}{Dist_c^2}\right)$$

nearest neighbor objects using the local kernel density estimation.

Done

Done

In the first step, the filtered data objects and its labels are taken as input for KNN approach. We perform step 2 by applying kernel-based clustering. In the third

6. Filter all the k-

step, a local kernel density is used to predict the nearest severity based density clusters. Here different groups are predicted using the probabilities of severity. Lastly, in this step, local density clusters are merged with each other based on probabilistic measure. In the final step, all the density clusters are visualized on the map.

Experimental Results

Experimental results are performed on different synthetic datasets in order to check the performance of proposed model to the conventional models. In this experimental study, python, java and spark are used to evaluate the results on different dimensions. The following experiments performed by considering two types of data sets they are Independent data and anti-correlated data. The sample parameters used in this experiment are Data cardinality and dimensions are considered in this experiment and sampling rates are 0.1%, 0.5%, 1%.

Table1
Performance of proposed model to conventional models on different anti-correlation data samples When threshold $\phi=0.5$

TestData	AGPMRS	GPMRS	MR-GPMRS	ParallelGPMRS	Proposed
#1	5457.73	5602.01	5256.92	5905.25	4150.31
#2	5522.57	5808.21	5512.65	5031.87	3889.65
#3	5267.81	5034.47	5152.21	5671.64	4103.64
#4	4654.16	4590.52	5743.77	5467.65	4034.57
#5	4907.8	4697.92	5697.26	4590.71	4006.92
#6	5255.06	5783.48	4806.17	5753.15	4194.32
#7	5576.42	5187.76	5584.53	5680.31	3985.07
#8	5463.97	5509.66	5200.4	5330.78	4004.18
#9	5398.41	5008.92	4854.44	5109.32	3930.54
#10	4675.31	5708.5	5875.32	4570.44	4040.29
#11	5683.32	5489.67	5400.73	4707.45	3889.79
#12	5794.14	5892.72	5789.68	5277.21	3821.02
#13	5130.36	4955.01	5246.21	4741.8	4220.31
#14	5277.45	5652.51	4740.57	5285.34	4151.39
#15	5053.29	4992.08	5737.81	5882.44	4223.02

Table2
Performance of proposed model to conventional models on different anti-correlation data samples
When threshold $\phi=0.7$

TestData	AGPMRS	GPMRS	MR-GPMRS	ParallelGPMRS	Proposed
#1	5255.59	5552.48	5739.05	5796.98	3853.59
#2	5188.98	5922.72	5728.26	5332.06	3950.78
#3	5465.57	4743.73	5330.89	5787.95	4107.79
#4	5395.68	5317.58	5122.1	5122.45	3847.42
#5	4581.96	5501.57	5097.4	5707.32	4143.73

#6	4653.26	5111.33	5553.15	5275.48	3847.41
#7	4750.36	5291.44	4871.21	4970.34	3844.06
#8	5678.76	5518.11	5099.13	5728.01	3853.61
#9	5476.22	4617.5	5616.15	5351.27	4013.81
#10	5240.38	5743.09	5253.83	5426.25	4200.27
#11	5556.68	5043.22	4715.02	5163.9	3897.82
#12	5120.32	5378.32	4928.66	5709.21	3799.88
#13	4570.56	4564.81	4913.83	5330.53	3810.43
#14	4683.93	5631.18	4599.53	5581.28	4187.64
#15	5342.86	5076.49	4980.97	4822.49	4245.36

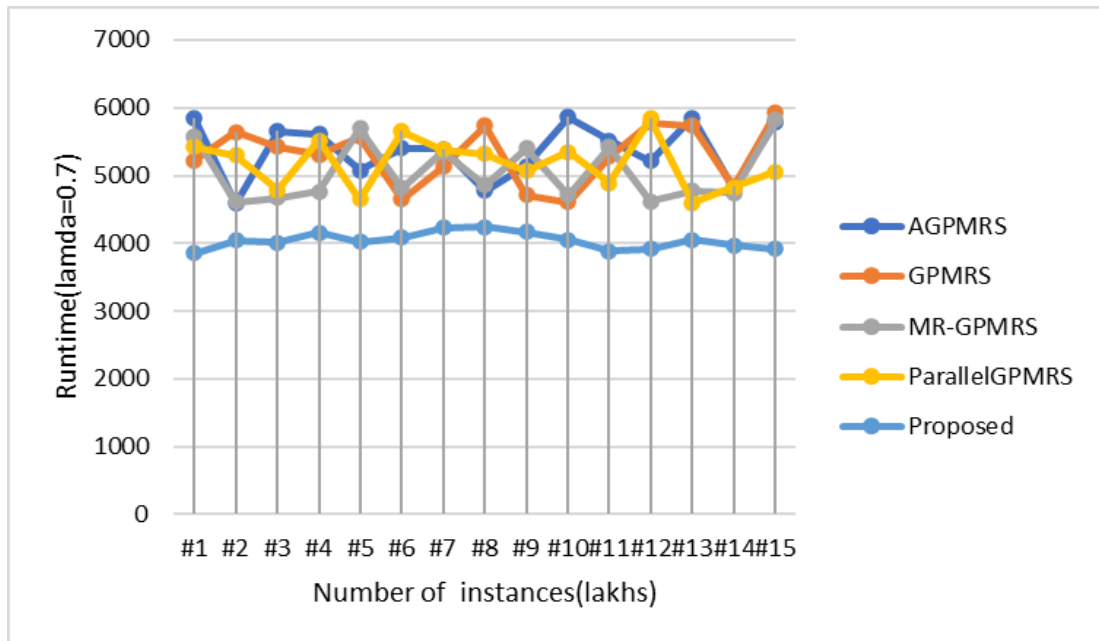


Figure 2: Performance of proposed model to conventional models runtime (ms) on different anti-correlation data samples
When threshold $\phi=0.8$

Table 3:

Performance of proposed model to conventional models local skyline patterns count on different anti-correlation data samples

TestData	AGPMRS	GPMRS	MR-GPMRS	ParallelGPMRS	Proposed
#1	20334	11735	21072	13188	29256
#2	22143	15635	16544	10690	26578
#3	17977	11498	17887	14858	29769
#4	11267	12477	19834	13798	29614
#5	10706	19568	13386	14714	27749
#6	13320	15893	14324	17121	28893
#7	18316	18646	17336	10275	29380
#8	22271	21100	18217	18174	25863

#9	23570	15000	13250	13549	25061
#10	11644	12693	13353	11547	26848
#11	14015	17529	12325	20362	26773
#12	15167	15064	17419	11624	25519
#13	12622	18761	19716	18376	29503
#14	18197	10373	14146	23863	27306
#15	16177	12112	20599	10818	25653

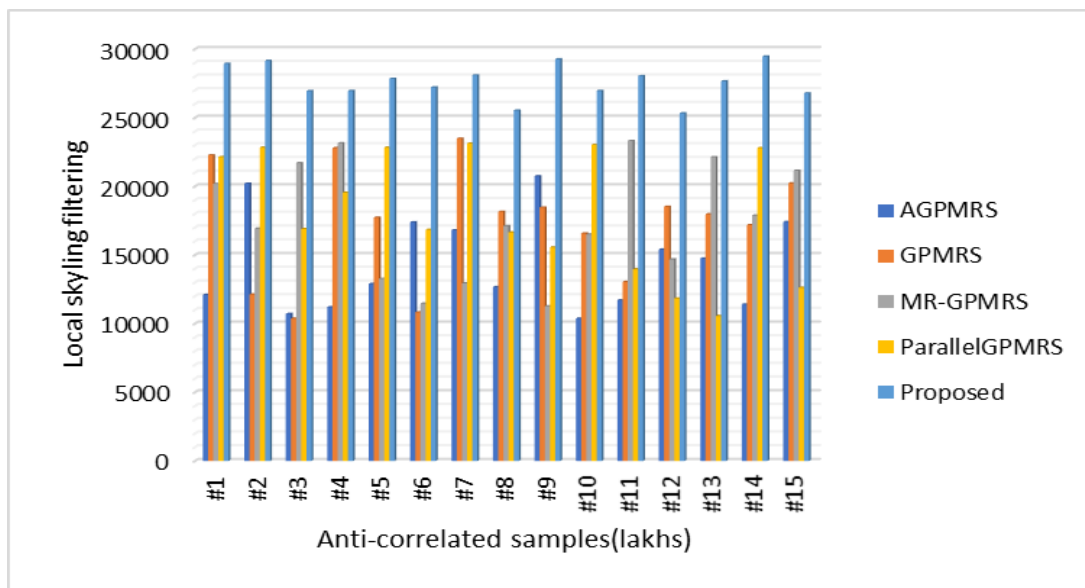


Figure 3: Performance of proposed model to conventional models local skyline KNN patterns count on different anti-correlation data samples

Table 3

Performance of proposed model to conventional models global skyline patterns count on different anti-correlation data samples

TestData	AGPMRS	GPMRS	MR-GPMRS	ParallelGPMRS	Proposed
#1	18641	16579	22070	18141	26475
#2	19770	20162	22724	15526	25858
#3	19822	10520	11322	12166	28258
#4	19209	17388	20092	20024	26472
#5	16929	23417	14488	14055	27546
#6	21465	17545	13822	15945	28271
#7	10428	20243	23282	21507	27697
#8	14102	19731	12692	13685	26120
#9	16364	19718	18064	10679	26429
#10	21648	17887	18963	12805	27300
#11	22194	11029	22486	20327	29479
#12	20354	15536	17310	19808	26146
#13	15373	11784	22215	10602	29057

#14	12456	22603	21592	16390	25086
#15	23885	22524	18127	17463	25951

Conclusion

Skyline processing become time consumed when growing data space. For this we proposed pre and post data filtering method that reduces skyline data space, and leads to reducing skyline processing time. A Novel Filtered based MapReduced Grid partitioning on high dimensional spatial datasets. This is an extension of GPMRS model using hadoop framework. This F-GPMRS is implemented in three phases shown in figure1. This F-GPMRS method could remove more candidate sets before skyline calculation and save more execution time in entire processing. In the future work, this work can be extended to optimize the performance of multi-dimensional real-time big databases with data size (>5GB) by using high end cloud computing environment.

References

- [1] B. Yin, K. Gu, X. Wei, S. Zhou, and Y. Liu, "A cost-efficient framework for finding prospective customers based on reverse skyline queries," *Knowledge-Based Systems*, vol. 152, pp. 117–135, Jul. 2018, doi: 10.1016/j.knosys.2018.04.011.
- [2] I. Kertiou et al., "A dynamic skyline technique for a context-aware selection of the best sensors in an IoT architecture," *Ad Hoc Networks*, vol. 81, pp. 183–196, Dec. 2018, doi: 10.1016/j.adhoc.2018.08.011.
- [3] H. Jung, H. Han, H. Y. Yeom, and S. Kang, "A fast and progressive algorithm for skyline queries with totally- and partially-ordered domains," *Journal of Systems and Software*, vol. 83, no. 3, pp. 429–445, Mar. 2010, doi: 10.1016/j.jss.2009.09.032.
- [4] K. Alami and S. Maabout, "A framework for multidimensional skyline queries over streaming data," *Data & Knowledge Engineering*, vol. 127, p. 101792, May 2020, doi: 10.1016/j.datak.2020.101792.
- [5] Y. Gulzar, A. A. Alwan, A. Z. Abualkishik, and A. Mehmood, "A Model for Computing Skyline Data Items in Cloud Incomplete Databases," *Procedia Computer Science*, vol. 170, pp. 249–256, Jan. 2020, doi: 10.1016/j.procs.2020.03.037.
- [6] L. Wei, W.-C. Oon, W. Zhu, and A. Lim, "A skyline heuristic for the 2D rectangular packing and strip packing problems," *European Journal of Operational Research*, vol. 215, no. 2, pp. 337–346, Dec. 2011, doi: 10.1016/j.ejor.2011.06.022.
- [7] H. Han, H. Jung, H. Eom, and H. Y. Yeom, "An efficient skyline framework for matchmaking applications," *Journal of Network and Computer Applications*, vol. 34, no. 1, pp. 102–115, Jan. 2011, doi: 10.1016/j.jnca.2010.08.011.
- [8] A. K. Pujari, V. Padmanabhan, and V. R. Kagita, "Bounds on skyline probability for databases with uncertain preferences," *International Journal of Approximate Reasoning*, vol. 80, pp. 199–213, Jan. 2017, doi: 10.1016/j.ijar.2016.09.004.
- [9] B. Bhattacharya, A. Bishnu, O. Cheong, S. Das, A. Karmakar, and J. Snoeyink, "Computation of spatial skyline points," *Computational*

- Geometry, vol. 93, p. 101698, Feb. 2021, doi: 10.1016/j.comgeo.2020.101698.
- [10] H. Lu, Y. Zhou, and J. Haustad, "Efficient and scalable continuous skyline monitoring in two-tier streaming settings," *Information Systems*, vol. 38, no. 1, pp. 68–81, Mar. 2013, doi: 10.1016/j.is.2012.05.005.
 - [11] M. Morse, J. M. Patel, and W. I. Grosky, "Efficient continuous skyline computation," *Information Sciences*, vol. 177, no. 17, pp. 3411–3437, Sep. 2007, doi: 10.1016/j.ins.2007.02.033.
 - [12] B. Yin, S. Zhou, Y. Lin, Y. Liu, and Y. Hu, "Efficient distributed skyline computation using dependency-based data partitioning," *Journal of Systems and Software*, vol. 93, pp. 69–83, Jul. 2014, doi: 10.1016/j.jss.2014.03.021.
 - [13] X. Lian and L. Chen, "Efficient processing of probabilistic group subspace skyline queries in uncertain databases," *Information Systems*, vol. 38, no. 3, pp. 265–285, May 2013, doi: 10.1016/j.is.2012.08.006.
 - [14] Z. Yang, X. Zhou, K. Li, G. Xiao, Y. Gao, and K. Li, "Efficient processing of top k group skyline queries," *Knowledge-Based Systems*, vol. 182, p. 104795, Oct. 2019, doi: 10.1016/j.knosys.2019.06.003.
 - [15] D. Pertesis and C. Doukeridis, "Efficient skyline query processing in SpatialHadoop," *Information Systems*, vol. 54, pp. 325–335, Dec. 2015, doi: 10.1016/j.is.2014.10.003.
 - [16] I.-F. Su, Y.-C. Chung, C. Lee, and Y.-Y. Lin, "Efficient skyline query processing in wireless sensor networks," *Journal of Parallel and Distributed Computing*, vol. 70, no. 6, pp. 680–698, Jun. 2010, doi: 10.1016/j.jpdc.2010.01.001.
 - [17] Y. Li, Z. Li, M. Dong, W. Qu, C. Ji, and J. Wu, "Efficient subspace skyline query based on user preference using MapReduce," *Ad Hoc Networks*, vol. 35, pp. 105–115, Dec. 2015, doi: 10.1016/j.adhoc.2015.07.006.
 - [18] J. Auerbach and P. Wan, "Forecasting the urban skyline with extreme value theory," *International Journal of Forecasting*, vol. 36, no. 3, pp. 814–828, Jul. 2020, doi: 10.1016/j.ijforecast.2019.09.004.
 - [19] X. Li, Y. Wang, X. Li, X. Wang, and J. Yu, "GDPS: An Efficient Approach for Skyline Queries over Distributed Uncertain Data," *Big Data Research*, vol. 1, pp. 23–36, Aug. 2014, doi: 10.1016/j.bdr.2014.07.003.
 - [20] H. Im and S. Park, "Group skyline computation," *Information Sciences*, vol. 188, pp. 151–169, Apr. 2012, doi: 10.1016/j.ins.2011.11.014.
 - [21] S. He and Z. Liu, "Image quality assessment based on adaptive multiple Skyline query," *Signal Processing: Image Communication*, vol. 80, p. 115676, Feb. 2020, doi: 10.1016/j.image.2019.115676.
 - [22] S. R. Sporzynski, "Implementation and use of high-performance parallel skyline matrix subroutines on the IBM vector facility," *Computing Systems in Engineering*, vol. 2, no. 2, pp. 203–215, Jan. 1991, doi: 10.1016/0956-0521(91)90021-V.
 - [23] P.-K. Eng, B. C. Ooi, and K.-L. Tan, "Indexing for progressive skyline computation," *Data & Knowledge Engineering*, vol. 46, no. 2, pp. 169–201, Aug. 2003, doi: 10.1016/S0169-023X(02)00208-2.
 - [24] G. Klammer, M. Dietrich, M. Farshad, L. Iselin, L. Nagy, and A. Schweizer, "Intraoperative Imaging of the Distal Radioulnar Joint Using a Modified Skyline View," *The Journal of Hand Surgery*, vol. 37, no. 3, pp. 503–508, Mar. 2012, doi: 10.1016/j.jhsa.2011.12.009.

- [25] Y. Zeng, G. Chen, K. Li, Y. Zhou, X. Zhou, and K. Li, "M-Skyline: Taking sunk cost and alternative recommendation in consideration for skyline query on uncertain data," *Knowledge-Based Systems*, vol. 163, pp. 204–213, Jan. 2019, doi: 10.1016/j.knosys.2018.08.024.
- [26] J.-L. Koh, C.-C. Chen, C.-Y. Chan, and A. L. P. Chen, "MapReduce skyline query processing with partitioning and distributed dominance tests," *Information Sciences*, vol. 375, pp. 114–137, Jan. 2017, doi: 10.1016/j.ins.2016.09.046.
- [27] T. Jiang, Y. Gao, B. Zhang, D. Lin, and Q. Li, "Monochromatic and bichromatic mutual skyline queries," *Expert Systems with Applications*, vol. 41, no. 4, Part 2, pp. 1885–1900, Mar. 2014, doi: 10.1016/j.eswa.2013.08.085.
- [28] W. Son, S. Hwang, and H.-K. Ahn, "MSSQ: Manhattan Spatial Skyline Queries," *Information Systems*, vol. 40, pp. 67–83, Mar. 2014, doi: 10.1016/j.is.2013.10.001.
- [29] M. Fort, J. A. Sellarès, and N. Valladares, "Nearest and farthest spatial skyline queries under multiplicative weighted Euclidean distances," *Knowledge-Based Systems*, vol. 192, p. 105299, Mar. 2020, doi: 10.1016/j.knosys.2019.105299.
- [30] Y. Gao, Q. Liu, B. Zheng, and G. Chen, "On efficient reverse skyline query processing," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3237–3249, Jun. 2014, doi: 10.1016/j.eswa.2013.11.012.
- [31] V. Sharma et al., "Panorama Public: A Public Repository for Quantitative Data Sets Processed in Skyline*," *Molecular & Cellular Proteomics*, vol. 17, no. 6, pp. 1239–1244, Jun. 2018, doi: 10.1074/mcp.RA117.000543.