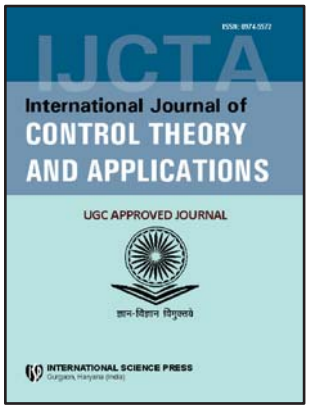

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 31 • 2017

A Novel Hadoop Based Multi-Way Join model for Co-locating Spatial Pattern Mining

V. Narendra Babu[a] and A. Suresh Babu[b]

[a]*Research Scholar, Department of Computer Science and Engineering, JNTUK, Kakinada-533001, Andhra Pradesh, India*
E-mail: narendrababu.v06@gmail.com
[b]*Associate Professor, Department of Computer Science and Engineering, JNTUA, Anantapuramu-515001, Andhra Pradesh, India*
E-mail: asureshjntu@gmail.com

Abstract: Mining spatial objects and its relations among the spatial events is one of the most essential researches in the spatial machine learning. In recent years, different pattern mining methods for spatial events analysis have been implemented; however none of them consider limited computing memory and computational speed. Also, most of the traditional spatial mining models often generate a large number of frequent spatial patterns which are difficult to analyze spatial events for decision making. Such constraints are necessary in the context of huge datasets for efficient discovery of spatial analysis. In this paper, we designed and implemented a Hadoop based Multi-way join model for frequent pattern discovery on the large spatial datasets. Hash-Join and Probabilistic Join operations are used in Mapper and Reducer phases for efficient event classification and pattern filtering process. Experimental results proved that proposed model has less error rate and time computations compared to traditional Hadoop based Multi-Join pattern discovery models.

Keywords: *Multi-way Join spatial mining, Spatial Pattern discover, Co-location patterns, Hadoop, Hash-Join, Probabilistic Join.*

1. INTRODUCTION

As the development of spatial technology and storage increases, it becomes difficult to organizations to process and extract essential patterns from the massive spatial data. So how to find essential patterns from these data becomes more and more complex. Because of the importance of spatial pattern discovery, the domain of finding frequent spatial relationships has been explored such as a join-less models, probabilistic prevalent models, join based models in uncertain spatial datasets. But some of these models may result in large patterns which are difficult to understand and use. Spatial frequent pattern may yield important results for many domain areas such as mobile commerce, event planning, transportations etc. For instance, location based services (LBS), various services provided by third party agencies with different geographical locations are interested to know what events are frequently occurred together in geo-spatial locations. Here the services are represented as spatial events, and the service points are represented as spatial objects or spatial points. Another example, the set {pharmacies, hospitals} may be an example because both the pharmacies and hospitals are frequently occurred together as a spatial pattern. Hence, spatial frequent pattern analysis is useful in geographic context and spatial mining context as shown in Figure 1[1].

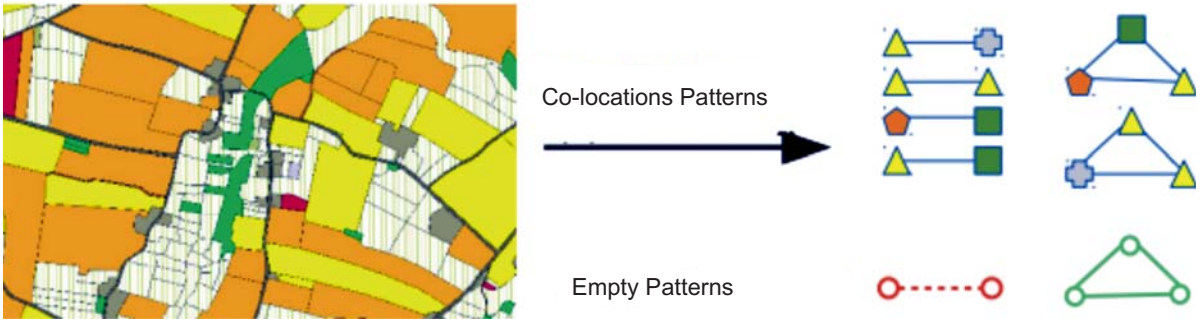

Figure 1: Spatial Neighbor Patterns

Spatial pattern discovery models which are based on association pattern analysis may be categorized into two groups. In the first group, association models are designed and developed on the small spatial datasets. In addition, association models require a significant amount of processing time to prune the candidate spatial objects. The models in the second group are applicable to spatial datasets using the Clique graph structure [2]. The basic spatial pattern mining architecture is summarized in the Figure 2. This architecture consists of a spatial data which is processed to discover the clique graph structures. From the clique structure, neighborhood relations are processed to find the *k*- frequent spatial patterns. Prevalence measure is used to prune the frequent candidate sets and its co-existing patterns.

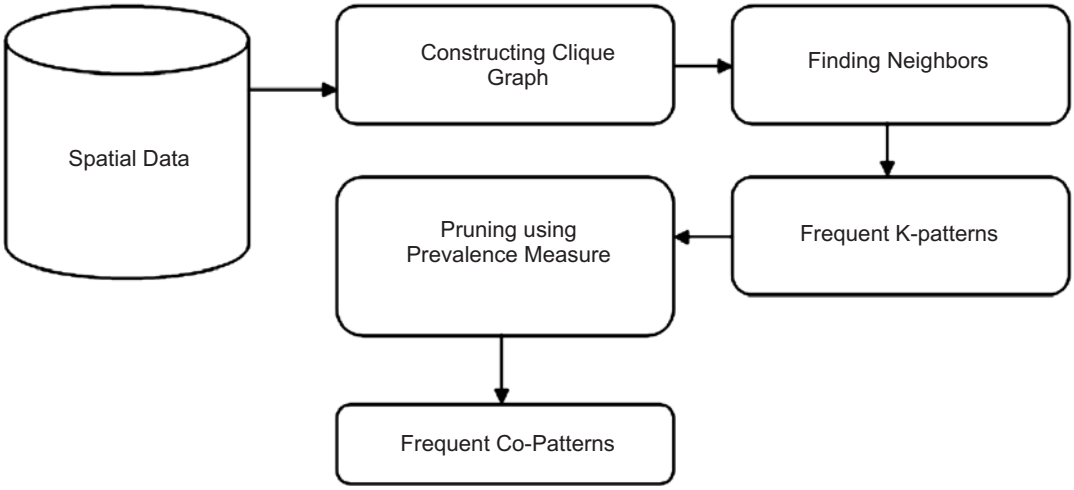

Figure 2: Traditional Spatial Pattern Mining Model

Although a large number of pattern mining models on spatial databases have been proposed in the literature, the computational memory involved in processing spatial objects on huge datasets is inherently time consuming. Also, due to the exponential growth of spatial data emphasizes the need for implementing computational efficient models for analyzing the spatial objects. As a solution, Hadoop based parallel processing models are becoming more important to deal with the large amounts of spatial datasets. Hadoop based spatial mining is used to discover frequent pattern analysis in many applications.

Geographical spatial data is represented by three topological structures- spatial point representation (x, y), spatial line representation $(x_i, y_j$ where $i, j = 1 \ldots m, n)$ and the polygon representation . The spatial analysis of discrete points has been analyzed using distance based methods and raster methods. The participation index is a measure used to find the co-location pattern in the traditional models of pattern mining. But the occurrences of nearest neighbor relationships are located in the specified region or location [3].

The Voronoi structure representation has been used as an alternative to process proximity amongst spatial objects, overcoming the problems of conventional data spatial adjacency methods. Figure 3 shows that the response time is increasing dramatically when number of dimensions n increasing from 2 to 10. Figure 4 shows the response of spatial patterns with respect to the number of spatial objects.

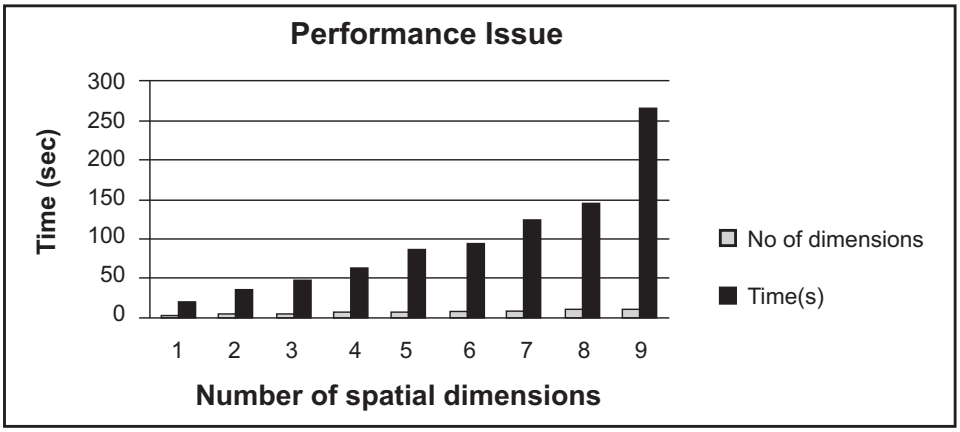

Figure 3: Response time of dimensions vs. Time

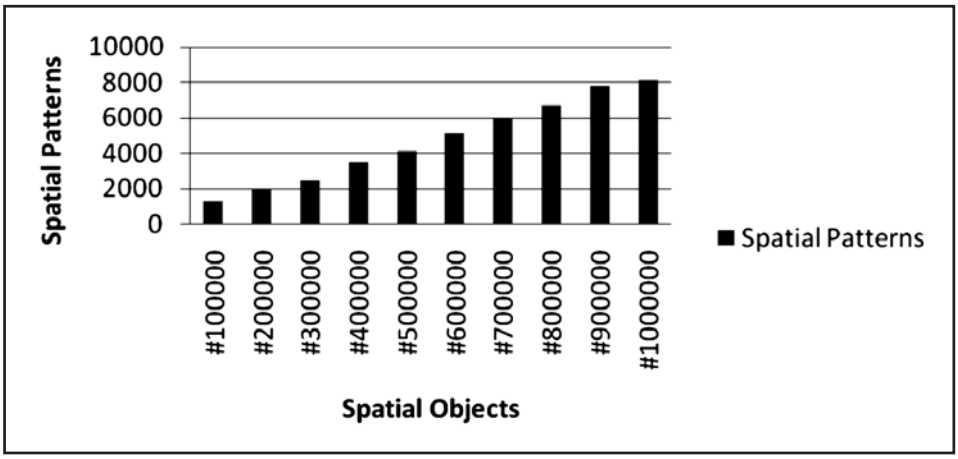

Figure 4: Response time of spatial objects vs Patterns

Traditionally, frequent spatial patterns are discovered by using event vector model, window vector model and feature raster mode. The conditional probability measure and the prevalence index are called pattern evaluation measures, which are used to find frequent co-location rules from the spatial objects.

In this paper, we have proposed a novel parallel processing model to achieve high processing efficiency for spatial pattern analysis. We take multiple large spatial datasets to discover the relationships among the objects and redefine the integrated Multi-Way spatial join operation using Hash-Join and Probabilistic Join in the Map-Reduce framework. Our proposed Multi-Way spatial join operation filters the candidate patterns and spatial relationships to reduce the computational time.

This paper is organized as follows: Some related work on the spatial pattern mining models and Multi-way join models are presented in section 2. Section 3 describes a novel hadoop based multi-way join using spatial frequent pattern model. Section 4 describes the experimental results and discussions. The last section presents our conclusion and future scope.

2. RELATED WORK

Grid based index is one of the most popular data structures for spatial pattern mining process. It partitions the entire grid space into equijoin cells. To provide fast spatial access, all the spatial objects are mapped to string join operation for spatial pruning. [4] proposed two measures that can be calculated without a transaction-type dataset. They studied the problem of similarity join, they proposed spatio textual similarity join operation using hadoop framework. The major problems of spatio textual join operation is how to choose methods related to data pre-partitioning, data feature selection and pattern filtering in hadoop framework[5]. [6] Proposed MRSimJoin, a hadoop based technique to efficiently solve the string similarity join problem. [7] implemented superset's theta-joins and equi-joins by using MapReduce. Based on MapReduce, the realization and corresponding preprocessing optimization of several parallel join algorithms were comprehensively studied [8], which were mainly focused on asymmetric tile replication join.

2.1. Spatial Pattern Models

A novel measure called the maximum participation index ratio is proposed for extracting frequent co-location patterns with rare spatial objects or events[9]. Probabilistic prevalent ratio is proposed in [10] to find frequently occurred patterns in the uncertain dataset or noisy data. Utility mining algorithm is proposed in UMining model to estimate the pattern pruning to optimize search space [11]. They mainly attributed the problem of computing multi-way spatial patterns on Map-Reduce framework. Specifically they proposed a Controlled-Replicate framework which can reduce the communication among cluster nodes. [12] Presented $\in$-Controlled-Replicate an improved Controlled-Replicate procedure for processing multi-way spatial queries on hadoop framework. They also proposed two-way spatial join based on Map- Reduce. However these algorithms do not consider optimization problem of filtering stage, which leads to a lot of useless calculation operation, and therefore greatly increased the data processing cost.

2.2. Maximal Clique Model

This model finds cliques that do not have any of their participation index measure in the neighborhood cliques. This model violates the maximal clique definition, which is a group of events having a common distance among them. For instance, clique-1 has three instances {X1, Y1, Z1} and clique-2 has {X1, Y2, R1}, instance {X1} is a common instance between the two neighbor cliques, which is possible. The main limitations of this model are: it allows some redundancy to form complex patterns and takes more time to form relational patterns.

2.3. Spatial classification models

Classification is a supervised classification which needs training spatial objects to configure the classifier, a validation data known as test data to determine the accuracy of the trained model. Spatial pattern discover models based on classification methods include, artificial neural networks, decision trees, support vector machines, *k*-nearest neighbor etc are used to classify the spatial events. Also, spatial prediction models such as regression model consider the dependent variable or the independent spatial objects of nearest neighbours in predicting the specific spatial location in pattern discovery. But these classification models are not extended to attributes of neighbouring objects or patterns and its relationships. Also, these models are not applicable to outlier detection on huge spatial data.

2.4. Spatial Association Pattern Mining

Spatial association pattern mining was used to discover interesting patterns between spatial events in large spatial databases. A spatial association pattern is represented as AàB, where both A and B are subsets of spatial objects set and $A \cap B = \phi$. Many possible spatial predicates (E.g., far spatial objects, closest objects, overlap , co-locating etc) can be used to evaluate the spatial association mining. It is practically expensive and time consuming process to predict various spatial patterns from a large spatial candidate sets. Another major limitation with spatial pattern mining model is that a large number of spatial candidate sets are generated with duplicate events.

3. PROPOSED MULTI-WAY SPATIAL JOIN FOR CO-LOCATION PATTERN MINING

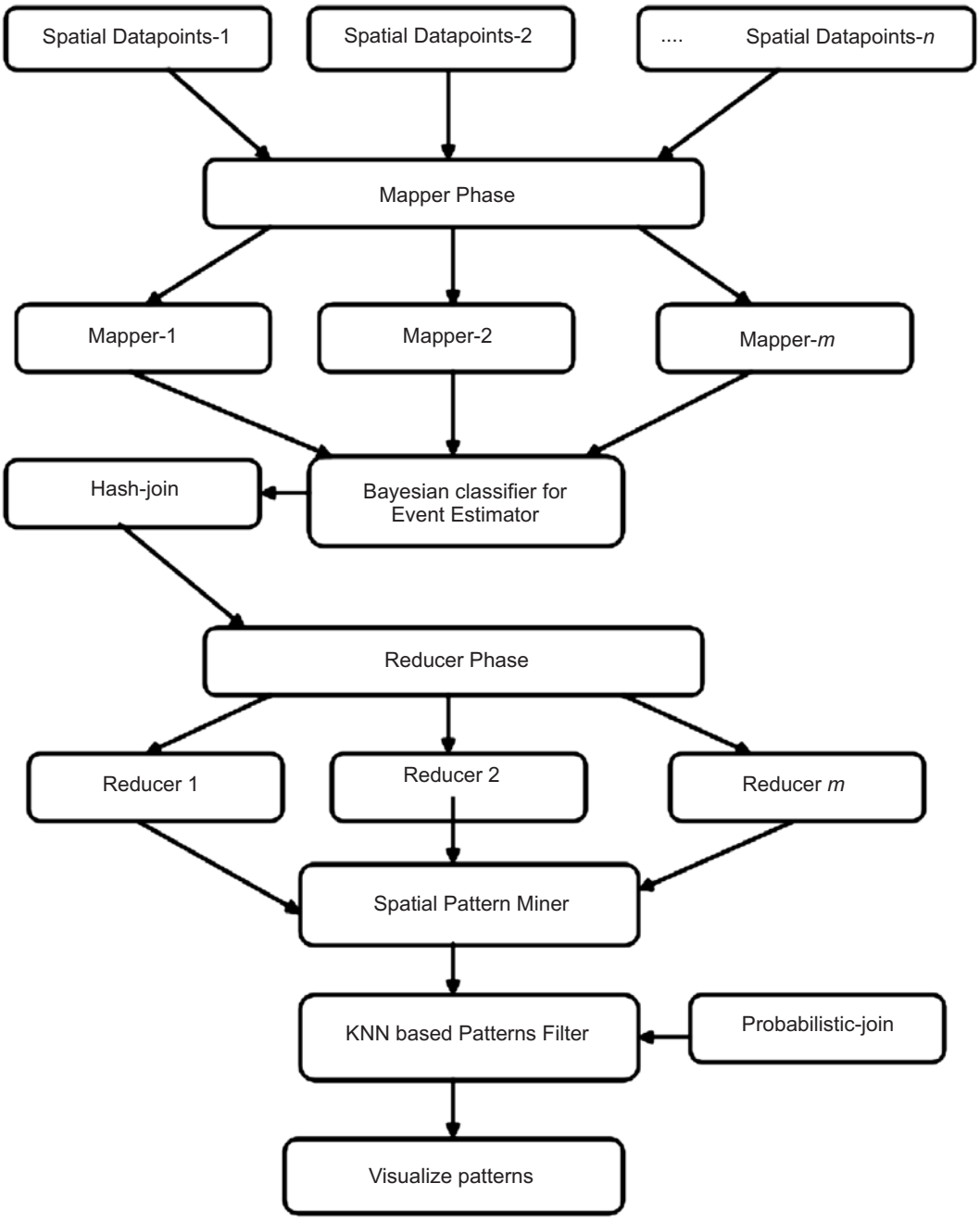

Figure 5: Proposed Architecture

Frequent spatial relationship can be defined with join relationships such as topological relations (*e.g.*, nearest or within region), (Euclidean distance), and regional relationship (*e.g.*, Asia, America). This work uses a probabilistic similarity join on each spatial dimension.

Figure 5, describes the proposed architecture on the spatial data. A multi-dimensional spatial datasets are taken as input to find the interesting spatial patterns from multiple spatial datasets.

Each spatial dataset is given as input to Mapper task to estimate the event class of the missing spatial object. In the reducer phase, spatial pattern mining model was implemented using probabilistic join and KNN based filter methods. Finally frequent patterns are discovered according to frequent usage basis.

Mapper : Spatial Event Classification

Input: N-dimensional Spatial Objects with Event, $<X_i, Y_j, O_1, O_2.. O_n, E_m>$, and distributed datasets D[1], D[2]..D[*n*].

For each Dataset D[*i*] ,where $i = 1,2 \ldots n$ // multiple dataset

Do

For each Data Object $P < X_i, Y_j, O_1, O_2.. O_n, E_m >$ in D[*i*]

do

if(D[*i*].*x*! = null and D[*i*]. *y*! = null and D[*i*]. E! = null)

then

Insert DataPoint P < X, Y > into Hash Join Grid and return $< G\_id, X_i, Y_j, O_1, O_2.. O_n, E_m >$

End if

Else if(D[*i*].*x*! = null and D[*i*].*y*! = null and D[*i*].E == null)

then

Apply Spatial Naïve Bayesian for Event classification.

Split data objects with each dimension as $< X_i, Y_j, O_1 >, < X_i, Y_j, O_2 >, < X_i, Y_j, O_3 >, \ldots. < X_i, Y_j, O_n >$

For each unlabelled Event object

do

Computing prior probability and Joint probability as $Prob(E_m)$ and $Prob (P < X_i, Y_j, O_1, O_2.. O_n > / E_m)$ on the trained Event data.

$$\text{Prob}(P < X_i, Y_j, O_1, O_2.. O_n > / E_m) = P(E_m) \prod_{i=1}^{n} \text{Prob}(P < X_i, Y_j, O_K > / E_m)$$

Insert Data Point $P < X, Y, O_k >$ into HashJoin Grid and return $<G\_id, X_i, Y_j, O_1, O_2.. O_n, E_p >$;

Where E_p is the predicted event.

Done

Done

Description: In the Mapper phase, N-dimensional spatial objects are taken as input to estimate missing spatial event class label. Naïve Bayesian event classification model was implemented to find the missing event. For this, prior and joint probability measures are used to find the most probable value of the Event estimator. Finally, estimated events along with N-dimensions are inserted into Hash Join operation with unique Grid_Id.

Reducer Phase:

Co-locating Pattern Miner < G_id,value = [Objects])

Construct Clique(GridObject[]);

Assign GridObjects[] to GridPoints[];

RPoint = ϕ // relevant points set

For *i* = 1…GridObject[].length

While (GridObject[*i*].*x* – GridObject[*j*].*x*) > λ)

Remove(GridObject[*j*],GridPoints[]);

$j = j + 1$;
end while
For i = 1…GridObject[].length
White ((GridObject[i]. O_k == GridObject[i].O_k) or GridObject[j].E_m < GriodObject[j].E_m)
do
Remove(GridObject[i],GridPoints[]);
Remove(GridObject[j],GridPoints[]);
$j = j + 1$;
i = i + 1;
end while
done
Apply Construct Probabilistic KNN Patterns (GridObject)
Generate K-nearest candidate spatial objects using Knn(GridObject[], E_m);.
η_{min} minimum weighted threshold
PRules ← ∅ ;
Find 1-Dimension frequent Spatial Objects as (It_1)
for (i = 2, It_{i-1} ! = ∅ , i + +)
do
JoinSet$_i$ ← ProbabilisticJoin (It_{i-1} , It_1) ;
done
For each K-Set $It_i \in$ JoinSet$_i$
do
w_i = ProbVals(GO[], It_i)
if $w_i \geq \eta_{min}$ then
$fps_m \leftarrow fps_m \cup \{It_i, E_m\}$
done
Display Pattern sets fps.
Apply Knn(GridObject[],E_m)
Select the top K values of spatial objects which are most similar to training cases.
GridObject[]: Labelled spatial objects.
E_m : m class events.
Procedure:
For each pair of spatial objects GO[i], GO[i] in GridObject[]
Do
For each event e_i in E_m
do
If $(GO[i], GO[j] \in E_m)$
Then

$$\text{Dist}(\text{GO}[i], \text{GO}[j]) = |\text{GO}[i].x - \text{GO}[j].x| + |\text{GO}[i].y - \text{GO}[j].y| + \sum_{k=1}^{|O_k|} |\text{GO}[i].O_k - \text{GO}[j].O_k|$$

Done

Select Top k spatial objects from Dist.

Probabilistic Similarity Join

Co-Located patterns: It is the subset of spatial features whose spatial instances are frequently(correlated) observed in a nearest location.

Co-Located Instance : It is the subset of spatial instances which includes all event(or feature) types will forms clique graph under the neighborhood relationship.

ProbabilisticJoin (GridObject It_{i-1}, GridObject It_1, GridObject[] GO[])

$$\text{Prob}(It_{i-1,}\, It_1) = \frac{\text{Prob}(It_{i-1} \cap It_1)|}{\text{Prob}(It_{i-1} \cup It_1)|}.\text{Prob}(\{It_{i-1} \cap It_1\}/\text{GO}[])$$

Description: In the reducer phase, clique graph is constructed using the spatial objects of all event types. From the clique graph, outlier points are removed using the threshold measure and event types. After eliminating the spatial outliers, probabilistic based KNN model was applied on each nearest spatial objects to find the most correlated spatial patterns. Here, the probabilistic join measure was used to find the most similar spatial dimensions among the nearest neighbor spatial objects. Finally Top-*k* spatial patterns are filtered from the large number of candidate sets.

4. PERFORMANCE ANALYSIS

In this experimental study, we have used multiple spatial datasets with different event types. We have implemented this model in Amazon AWS server with greater than two cluster nodes, one is master and others are slaves. We have implemented these models on Amazon cloud services with Linux as operating system. Also, the Map-Reduce framework was used for spatial pattern mining process. Finally, we have analyzed our proposed model with the traditional models in terms of pattern filtering and time complexity are concerned.

Table 1
Clique Neighbor graphs with varying Knn distance

#Spatial Size	*#Clusters-Nodes*	*K-NN Distance*	*Clique Neighbors*
#200000	3	3	25
#400000	4	5	37
#600000	5	8	124
#800000	6	12	213

Table 1, describes the spatial data size with different cluster nodes in Hadoop environment. As the K-NN distance increases proposed model minimizes the neighbor relations.

Figure 6, describes the spatial data size with different cluster nodes in Hadoop environment. As the K-NN distance increases proposed model minimizes the neighbor relations.

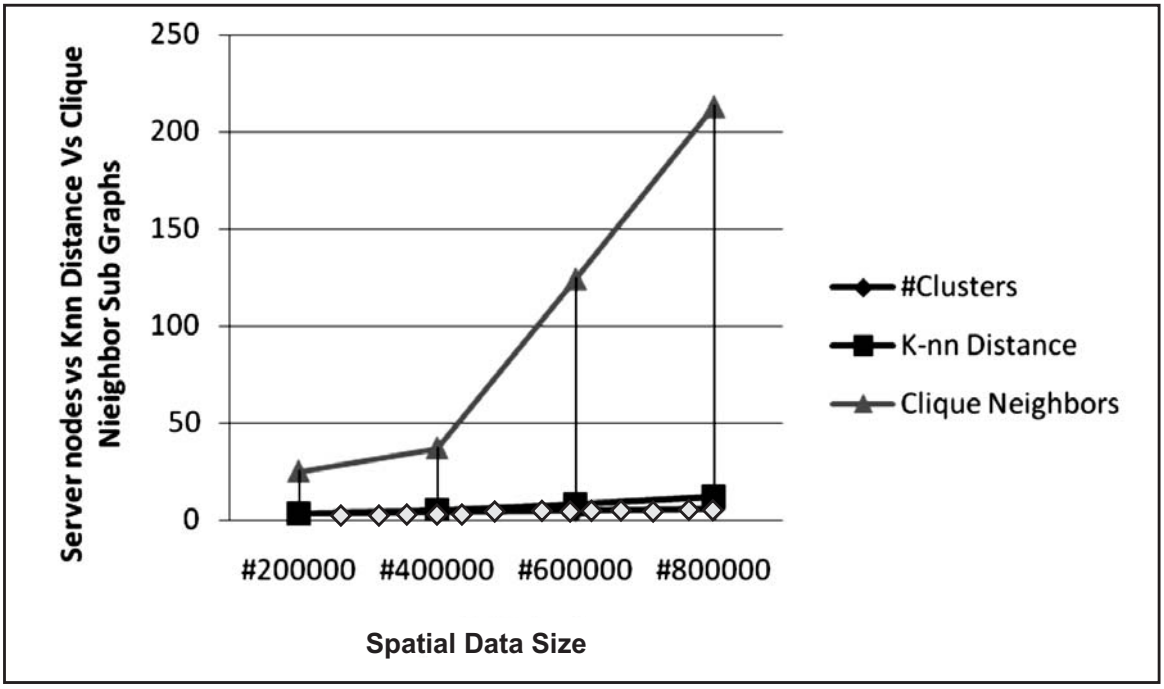

Figure 6 : Clique Neighbor graphs with varying Knn distance

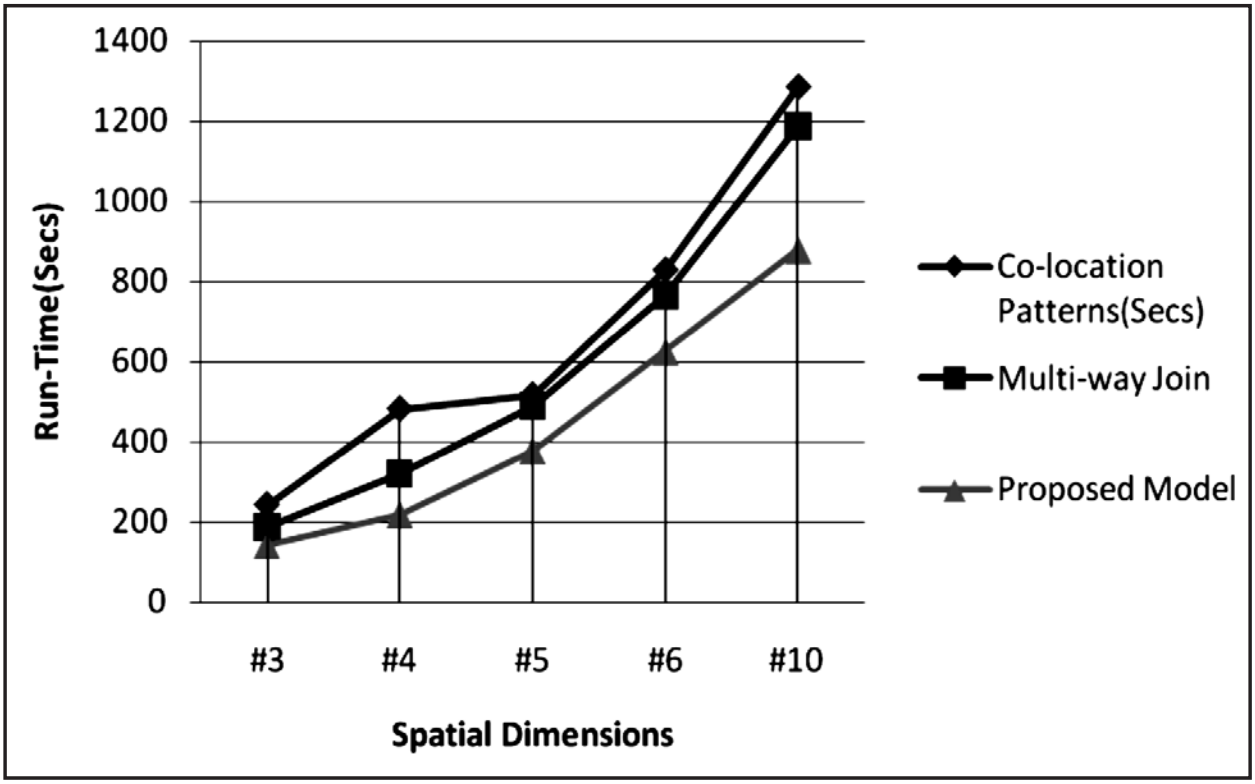

Figure 7: Runtime comparison of Proposed and Traditional models

Table 2
Run-time comparison of Proposed and Traditional models

Spatial Dimensions	*Co-location Patterns(secs)*	*Multi-way Join(secs)*	*Proposed Model(secs)*
3	242	187	143
4	482	321	219
5	517	491	378
6	829	765	625
10	1287	1189	879

Table 2, describes the runtime comparison of proposed multi-join model on high dimensional spatial data to the traditional models. As shown in the figure, as the size of the spatial dimensions increases, proposed model has less computational time compared to traditional models.

Figure 7, describes the runtime comparison of proposed multi-join model on high dimensional spatial data to the traditional models. As shown in the figure, as the size of the spatial dimensions increases, proposed model has less computational time compared to traditional models.

5. CONCLUSION

Multi-way join model for frequent pattern discovery on the large spatial datasets. Hash-Join and Probabilistic Join operations are used in Mapper and Reducer phases for efficient event classification and pattern filtering process. Experimental results proved that proposed model has less error rate and time computations compared to traditional hadoop based Multi-Join pattern discovery models. In future, this work can be extended to real-time web spatial mining to analyze the spatial objects clusters using the Hadoop framework.

REFERENCES

[1] https://raweb.inria.fr/rapportsactivite/RA2015/dream/dream.pdf

[2] Ghazi Al-Naymat,"Enumeration of maximal clique for mining spatial co-location patterns",Computer Systems and Applications, 2008. AICCSA 2008.

[3] Dasari, Naga Shailaja and Zubair Mohammad. "Maximal Clique Enumeration For Large Graphs On Hadoop Framework". Proceedings of the first workshop on Parallel programming for analytics applications - PPAA '14 (2014): n. pag. Web. 12 Oct. 2016.

[4] Yu, Minghe, Dong Deng, and Jianhua Feng. "String Similarity Search And Join: A Survey". Frontiers of Computer Science 10.3 (2015): 399-417. Web. 12 Oct. 2016.

[5] Zhang, Yu and Xiaofeng Meng. "Efficient Spatio-Textual Similarity Join Using Mapreduce".2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) (2014): n. pag. Web. 11 Oct. 2016.

[6] Silva, Y.N., Reed, J.M., Tsosie, L.M.: MapReduce-based similarity join for metric spaces. In: VLDB/Cloud-I (2012).

[7] Penar, Maciej and Artur Wilczek. "The Design Of The Efficient Theta-Join In Map-Reduce Environment". Communications in Computer and Information Science (2016): 204-215. Web. 12 Oct. 2016.

[8] Jin-Deog Kim, and Bong-Hee Hong. "Parallel Spatial Joins Using Grid Files". Proceedings Seventh International Conference on Parallel and Distributed Systems (Cat. No.PR00568) n. pag. Web. 12 Oct. 2016.

[9] Huang Y, Pei J, Xiong H. Mining co-location patterns with rare events from spatial data sets. Geoinformatica, 2006, 10(3): 239-260.

[10] Wang L, Wu P, Chen H. Finding probabilistic prevalent colocations in spatially uncertain data sets. Knowledge and Data Engineering, IEEE Transactions on, 2013, 25(4): 790-804.

[11] Yao H, Hamilton H J, Geng L. A unified framework for utility-based measures for mining itemsets//Proc. of ACM SIGKDD 2nd Workshop on Utility-Based Data Mining. 2006: 28-37.

[12] Jing Cheng, Yong Gao, and Jiajun Liu. "The Implementation Of Spatial Co-Location Mining And Its Application For Pois In Beijing". 2014 22nd International Conference on Geoinformatics (2014): n. pag. Web. 12 Oct. 2016.