International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 1 (2016) pp 586-591

Parallel Co-location Pattern Mining Discovery:Constraint Neighborhood Approach

Eman M. Refaye,Osman Hegazy,

Department of Information Systems,
Faculty of computers and Information, Cairo University,
Cairo, Egypt.
eman_refaye@yahoo.com

Professor,
Department ofInformation Systems
Faculty of computers and Information,Cairo University,
Cairo, Egypt.
osman.hegazy@gmail.com

Abstract-Spatial data mining become one of the important areas because of the rapid evolution in technology which leads in big spatial data. Co-locations pattern mining isan interesting and important issue in spatial data mining area which discovers the subsets of features whose events are frequently located together in geographic space. Spatial proximity is the important concept to determine the colocation patterns from massive data. The computation of co-location pattern discovery is very expensive with big data volume and nearby existence of neighborhoods. So there is number of spatial co-location mining algorithms have been developed to overcome these drawbacks. In this paper, a new a colocation pattern mining framework has been proposed that benefits from the power of parallel processing, in particular, the MapReduce framework to achieve higher spatial mining processing efficiency. MapReduce model have been proven to be an efficient frameworksolution for big data processing on clusters of commodity machines, and for big data analysis and many applications. The experimental result of the proposed framework shows scalable and efficient computational performance.

Keywords: Spatial Data Mining, Co-location Pattern, Hadoop, MapReduce, Constraint Neighborhood.

Introduction

Data mining in general is searching for hidden and interesting patterns that may exist in generic data. Spatial data mining in particular is discovering the interesting relationships and characteristics that may exist implicitly in spatial data [1].Spatial data mining is a new and rapidly developing area of data mining, concerned with the identification of interesting spatial patterns from data stored in spatial datasets and geographic information systems. GIS are used in various areas such as environmental impact assessment, urban planning, cartography, criminology, traffic analysis, etc. Collection of data is enabled byglobal positioning systems (GPS) and sensor networks, while computer storage technology enables the storage of enormous quantitiesof collected data. These advanced technologies are the reason for theexistence of a growing number of spatial datasets. The size of spatial datasets and thecomplexity ofdealing with spatial attributes require the use of specialized data mining techniques [2].

Spatial co-location pattern mining, is one of the important area in spatial data mining, has been researched in spatial data mining techniques. Spatial co-location pattern describes "a set of spatial events which are frequently observed together in a spatial proximity" [3].

Also a co-location pattern determines what these co-located objects, each typically occur in geographical proximity. Identified co-location patterns are interesting and helpful for many applications such as location-based services, public health and climatology, Disease control, Transportation, Business, Social Science, Geology, and Mobile Computing [3], [4].

Co-location rules are models to find the presence of Boolean spatial features in the neighborhood of instances of other Boolean spatial features. Mining Co-location pattern is the process to identify co-location patterns from big spatial datasets with a number of Boolean features. The spatial co-location rule discovery problem looks like the association rule mining problem, but, in fact, it is very different from the association rule mining problem. These differences have been made because of the lack of transactions. It uses spatial predicate as item types. So using of co-location pattern mining enhancing the efficiency of detecting interesting patterns from the very big spatial data. Co-location patterns are discovered by using neighborhood definition and spatial joins. These definitions and algorithms will be discussed the detection of co-location pattern from the large spatial datasets [5].

Big data is one of the hotspot in technological area and brings not only large amounts of data but also various data types that would not been considered [10].The evolution of location sensing, mobile computing, and scientific simulation is generating huge quantities of rich spatial data. Finding the solution that is able to translate the plentiful amount of spatial data that surrounds us into meaningful and useful information has led to the rise of spatial data mining [20]. So Spatial data mining has been popularly studied for detecting a specific association relationships between a set of spatial attributes and some of them may be non-spatial attributes. But dealing with large-scale spatial data mining isn't easy because of complex spatial data types, neighbor relationships [19].

One of the drawbacks of co-location pattern mining that is the wasting of cost of time to hold a vast number of candidate pattern sets. Also single processor's memory and CPU resources are very limited, which make an inefficient performance of co-location mining algorithms. Furthermore, because of growth of information, enterprises have to deal with growing amount of spatial data. So, the solution to this problem is parallel and distributed computing [6].

MapReduce offers a simple programming model for parallel data analysis. It is one of the most popular system built based on these ideas is Google's MapReduce and its open-source implementation, Hadoop. It simplifies parallel data processing by abstracting the details of data partitioning, node communication and fault tolerance [7].

This paper will represent the following: background of co-location pattern mining algorithms, MapReducemodel and Hadoop framework, related work, the proposed framework, performance evaluation and conclusion.

Background

A. Spatial Data Mining Approaches

Spatial data mining is a rising exploration field devoted to the advancement and utilization of novel computational procedures for the examination of big spatial datasets. It envelops methods for finding valuable spatial associations and patterns that are not stored in spatial datasets. Generally these procedures need to manage complex features with spatial data properties. The properties and relationships that have been contained in spatial data are different from transactional data. For instance, transactional data are stored in discrete space of numeric and categorical data rather than spatial data which are stored in continuous space. Transactions in transactional data are independent of each other unlike the spatial data share a variety of spatial relationships among each other.

The complexity of spatial data type and implicit relationships among spatial objects makes the process of discovering spatial patterns from spatial data is more difficult compared to the process discovering patterns from traditional data.Different approaches have been developed for knowledge discovery from spatial data such as Spatial Classification, Spatial association rule mining, Spatial Clustering [8].

i. Spatial Classification

spatial classification is known as that attributes can be grouped with respect to their values into categories also the attribute values of objects of neighbors may also be related to the membership of objects so that they have to be considered as well Assigning an object to a class from a given set of classes based on the attribute values of the object is the main objective of classification [8].

ii. Spatial Association

Identifying the regularities between the items in the large transactional datasets is known as association rule mining. Confidence and support are measures that indicate the strength of the frequency of the association rule [8], [20].

iii. Spatial Clustering

The task of collecting the objects of a datasets into meaningful detectable subclasses that is clusters is known as clustering so that the members of the same cluster are as similar as possible whereas the members of different clusters differ as much as possible from each other. Clustering algorithms can beCategorized into Five types they are Model-Based Clustering Methods, Partitioning algorithms, Hierarchical Algorithms ,Density based clustering and Grid-Based Methods [8].

B. Co-location Pattern Mining

There is a similarity between Spatial co-location pattern mining and association mining. A rule of the form "A→B" is a spatial association rule, where A and B are sets of predicates and some of which are spatial ones.For big datasets many relationships may exist but some may occur rarely or may not hold in most cases [3].

A set of instances S, a set of spatial features F, and a spatial neighbor relationship R over S. Spatial neighbor relationship R may be one of the following categories: directional, topological and distance. R could be distance relationships (e.g. Euclidean distance metric) and topological relationships (e.g. linked, intersection), and mixed relationships (e.g. the shortest distance of two points on a map) [9].

A co-location C = {A,B,C,…..} is a subset of spatial features C⊆ F whose instance objects are frequently observed in a nearby area according . A co-location instance I ⊆ S of a co-location C is defined as a set of objects which includes all features types in C and forms a clique under the neighbor relationship R. i.e., {A.2, B.4, C.2} is a co-location instance of {A, B,C}. The prevalence of co-locations is often measured by participation index [10], [11].

The participation index PI(C) of a co-location C = $\{f_1,f_2,\ldots..f_k\}$ is defined as:PI(C)=min e i∈ C{PR(C,f_i)}, where $1 \le i \le k$, and PR(C,f_i)is the participation ratio of event type fi in the co-location C that is the fraction of objects of event fi in the neighborhood of instances of co-location C−{f_i},

$$PR(C,f_i)= \frac{\text{number of distinct objects of fi in instances of C}}{\text{number of objects of fi}} \quad (1)$$

The prevalence measure indicates wherever an event in C is observed, with a probability of at least PI(C), all other events in C can be observed in its neighborhood. If the participation index of an event set is greater than a user-specified minimum prevalence threshold min prev., the event set is called a co-location or co-located event set [10].

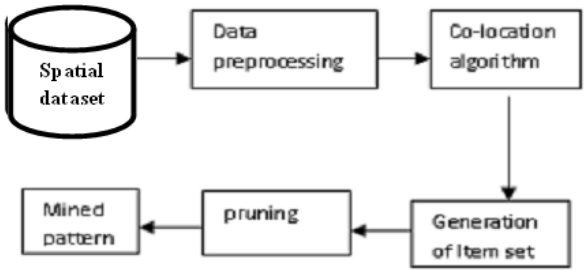

Fig.1. General Architecture of co-location mining algorithm[12]

C. Hadoop Framework

Hadoop is open source software that runs on a cluster of machines. Hadoop supplies both distributed processing and distributed storage for very large data sets [13].It consists of two models: Hadoop Distributed File System (HDFS) which

is the distributed storage model which designed after Google File System (GFS) and MapReduce; the programing model. Now it supports additional models and systems such as: HBase; a distributed column-oriented database, Hive; a data warehouse system, Avro; a data serialization system, Chukwa; a data collection system, ZooKeeper; a high-performance coordination service for distributed application, and Pig; a high level data-flow language [14].

HDFS - Hadoop Distributed File System HDFS, gives the programmer unlimited storage, HDFS implementation is modeled after GFS, Google Distributed File system .However; here are additional advantages of HDFS.

• Horizontal scalability. Thousands of servers holding petabytes of data. When you need even more storage, you don't switch to more expensive solutions, but add servers instead.

• Commodity hardware. HDFS is designed with relatively cheap commodity hardware in mind. HDFS is self-healing and replicating.

• Fault tolerance. Every member of the Hadoop knows how to deal with hardware failures. If you have 10 thousand servers, then you will see one server fail every day, on average. HDFS foresees that by replicating the data, by default three times, on different data node servers. Thus, if one data node fails, the other two can be used to restore the third one in a different place [13].

TABLE.1 HadoopRoles [13]

Job type	**Job Function**	**Skills**
Hadoop Developer	Develops MapReduce jobs, designs data warehouses	Java,Scripting,Linux
Hadoop Admin	Manages Hadoop cluster, designs data pipelines	Linux administration, Network Management, Experience in managing large cluster of machines
Data Scientist	Data mining and figuring out hidden knowledge in data	Math, data mining algorithms
Business Analyst	Analyzes data	Pig, Hive, SQL superman, familiarity with other BI tools

HadoopMapReduce is a software framework used for writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner [15].

D. MapReduce Model

MapReduce is a programming model for expressing distributed computations on massive amounts of data and an execution framework for large-scale data processing on clusters of commodity servers. MapReduce simplifies parallel processing by abstracting away the complexities involved in working with distributed systems, such as computational parallelization, work distribution, and dealing with unreliable hardware and software. The MapReduce model abstraction is presented in Fig2. A MapReduce job is executed in two main phases of user defined data transformation functions, namely, map and reduce. When a job is launched, the input data is split into physical blocks and distributed among nodes in the cluster. Such division and distribution of data is called sharding, and each part is called a shard. Each block in the shard is viewed as a list of key−value pairs. In the first phase, the key−value pairs are processed by a mapper, and are provided individually to the map function. The output of the map function is another set of intermediate key − value pairs [10].

A MapReduce job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce tasks. Typically both the input and the output of the job are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks [15]. .MapReduce takes care of distributed computing. It reads the data, usually from its storage, the Hadoop Distributed File System (HDFS), in an optimal way. However, it can read the data from other places too, including mounted local file systems, the web, and databases. It divides the computations between different computers (servers, or nodes). It is also fault-tolerant [13].

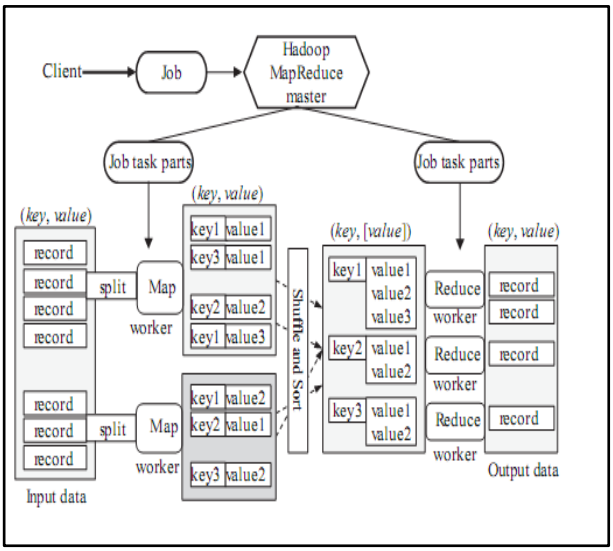

Fig.2. MapReduce Program and Execution Models [10]

Related Work

Candidate co-location pattern instances have been discovered using several proposed approaches. In [17] the authors of this approach groups neighboring instances with a non-overlapping instance grouping constraint arbitrarily.

Different instance sets by the order of grouping can be obtained from the disjoint grouping method [16].

Other works on spatial co-location mining have presented different approaches for identifying co-location instances. According to [18] presented these contributions such as, some uses space partitioning and non-overlap grouping scheme is used for finding neighboring objects for a frequent neighboring feature set. However, there are number of missed co-location instances across partition areas and incorrect results generated from the distinct space partitioning approach.

Other proposed a join-based algorithm for an instance co-location mining which is similar to apriori_gen. With the increasing of co-location patterns and their instance, The join-based approach is considered very computationally expensive .Reducing the number of expensive spatial join operations in finding co-location instances is very important which is proposed in a partial join approach . However, the performance depends on the distribution of the spatial dataset [18]. Joinless algorithms have been proposed, but methods including the joinless algorithm do not discover both single self co-location patterns and complex self co-location patterns [10].

In [3] a novel constraint neighborhood based approach had been proposed to find co-location patterns. This approach can discover different co-location patterns such as star and clique (as shown in fig. 3), including single and complex self co-locations. Based on the constraint neighborhood idea, this method neither needs to perform spatial or instance neither joins nor checks for cliques to find co-location instances.

The constraint neighbor co-location approach discovered the colocation patterns across two algorithms:
Algorithm1:Starts with determining the constraint neighbors by scanning the spatial object dataset to of each object, and then builds a set of single feature (size-1) co- location candidates.
Algorithm 2: It generates size-k pattern candidates based on size- (k−1) prevalent patterns by applying the level-wise approach. Also checks whether all subsets of the new candidate are prevalent. The algorithm uses the participation index that has the anti-monotonic property to measure the prevalence of new candidates [3].

Comparing the advantages of this approach, to others, this algorithm neither has to perform instance joins nor checks for cliques co-location patterns. Thus creating to an important performance gained form co-location patterns discovered of mining process. But this approach finds co-location patterns with more reads in the candidate set generation process.

Authors in [19],[10] they used partition strategy edges in the neighbor graph to divide the search space of co-location patterns. Dividing the edges of the neighbor graph is done according to this rule such that each vertex v (a spatial object) keeps the relationship edge with other vertex uwhen v.type<u.type. There is a total ordering of the event type was assumed.

The partition strategy is assumed to divide neighbor relations without duplicating or missing any relationships needed for co-location patterns. Also these algorithms have been implemented using the power of parallel processing, in particular, the MapReduce framework to achieve higher spatial mining processing efficiency. But these models do not discover both single self co-location patterns and complex self co-location patterns also need more computations in checks for cliques co-location instances.

In this paper a new implementation of parallel Co-location Pattern Mining algorithm based on Constraint Neighborhood Approach using Hadoop-MapReduce model to overcome the problems found in the previous algorithms.

Proposed System

In this section, we present our parallel colocation pattern mining model based on constraint neighborhood approach and MapReduce model, Which starts by Identifying spatial input dataset reordering the instances records according their types then according to ID'S. COUNT find total number of instances for each feature type Find CN for each object in the dataset according to spatial relationship between objects such that $R(o_i,o_j)$ computed according to constraint neighbor approach. Find list of constraint neighbor for each object.

Eliminate /remove object that are irrelevant. If object has no neighbors and doesn't included in a list of constraint neighbor. Finally generate colocation patterns. Our proposed system find prevalent co-location patterns and implemented across three map/reduce jobs.

A. Phase 1: Preparation of Spatial Files

Our Spatial Dataset obtained from more than one file. So there is need to combine these files into one file using MapReduce model. First the map job assigns each object o_i to its corresponding feature type fi . Then applying the reduce job for sorting these objects according to their features types and within the same type according to the objects ID'S. Then counting and saving the number of object instances per feature type for future prevalence calculation. The ordering task used for eliminating the duplications and missed instances. If we have D spatial dataset consists of n objects D = {o1,o2, ..., on} , and F set of m features F = $\{f_1,f_2, ..., f_m\}$, (m << n), i.e., $f_1< f_2<... <f_m$, $id_1< id_2< ... <id_n$.

```
1: procedure MAPPER(key,value=o)
2:F( o_i) ← o_i ;
3:emit( F(o_i), o_i) ;
4:end procedure

5:procedure REDUCER(key= , value=[ o])
6:objectSet ← o_i ;
7:sort(objectSet) ;
8:count ← sum(value) ;
9:save( objectSet, count) ;
10:emit(F( o_i), o_i ) ;
11:end procedure
```

Fig.3. Preparation of Spatial Files

B. Phase 2: Generate the List of Constrain Neighbors (CN) For Each Object

In this phase the main job is to generate the list of neighbors of each object according to the definition of the constraint neighborhood approach CN by check each object with other to find the constraint neighbor list according to the following:For clique colocation:CCN($\{o_i\}$):= sort<($\{o_j | (o_i,o_j) \in R \wedge ((o_i.type<o_j.type) \vee (o_i.type = o_j.type \wedge o_i.id < o_j.id))$, $(j \neq i)\}$). For star colocation patterns:SCN ($\{o_i\}$):= sort<($\{o_j | (o_i,o_j) \in R, (j \neq i)\}$).Then builds a set of single-feature (size-1) co-location candidates.

```
1: procedure MAPPER(key=oi, value= oj)
2:findN( oi);
3:emit(oi, oj) ;
4:end procedure
5:procedure REDUCER(key=oi, value=oj)
6:checkCN(oi, oj ) ;
7:GroupCN(oi);
8:emit(oi, CN);
9:end procedure
```

Fig.4. Generate the List of Constrain Neighbors (CN) For Each Object

C. *Phase 3: Colocation Patterns Generation*

In this phase the algorithm applies the level-wise approach to generate size-k pattern candidates from size-(k−1) prevalent patterns and checks whether all subsets of the new candidate are prevalent. The pattern instances of the new candidate are discovered. The algorithm uses the participation index that has the anti-monotonic property to measure the prevalence of new candidates.

```
1:  procedure MAPPER (key=oi, value=CN)
2:  k := 2;L1 := CN;
3:  while Lk-1 ≠ ϕ do

4:        foreach p, q ∈ Lk-1 s.t.
5:        p =[c1, ..., ck-2,ck-1],
6:        q =[c1, ..., ck-2,ck], ck∈p.CN
7:        DO
8:          c =[c1, ...ck-1,ck];
9:            if checkSubsets(c, Lk-1) then
10:             foreach I ∈ p.instances do
11:           foreach
12:                       (o ∈ I.CN) and (o.type = ck) do
13:             newI.objs := I.objs ∪{o};
14:                   newI.CN :=getCN(I.CN,o);
15:                   c.instances.add(newI);
16:               end for
17:           end for
18:         end if
19: emit(c,c.instaces);
20: procedure REDUCER(key=c, value=c.instances)
21:    if Prevalence(c) ≥ ρ then
22:    Ck.add(c);
23: end procedure
```

Fig.5. Generate Colocation Patterns Discovery

Experiment and Results

In this section, we present the results of our experimental evaluation of the proposed model constraint neighborhood based on MapReduce to mine co-location patterns. We show the effectiveness of our approach to find co-location patterns by comparing the patterns discovered to those of the constraint neighborhood approach represented in [3].

All experiments have been performed on a single machine that contains: Windows 7 64-bit, Eclipse Java EE IDE for web developers version Mars Release (4.5.0), Apache Hadoop 2.3.0 run on the stand-alone mode,ESRI geometry API, Spatial SDK Hadoop. All algorithms have been implemented in Java: JDK version 4.5.0. Experimental dataset used is spatial data about Leeds city that contains multiple features such as schools, accidents, hotels, traffic signals and so on. The dataset sample contains 6473records withdistinct feature109. The following table demonstrates the performance of the proposed algorithm compared by constraint neighborhood colocation miner using different prevalence threshold.

TABLE.2 ExecutionTime VS Prevalence Threshold

No. of Records 6473 and 109 Unique Features									
Prevalence	0.01	0.015	0.020	0.025	0.030	0.035	0.040	0.045	0.050
MapReduce CN Co-Location	59.30	59.16	58.50	59.90	60.90	58.70	58.88	58.34	58.38
Constraint Neighbor	315	312	304	301	294	288.16	286.30	279.20	279

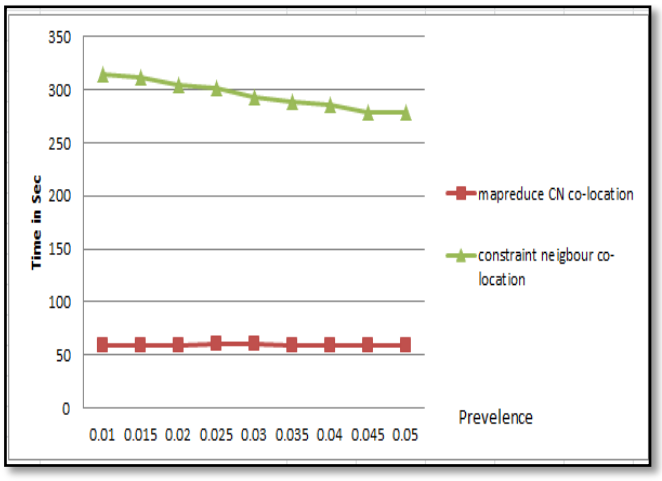

Fig.6. Performance Evaluation of the Proposed Colocation Miner

Conclusion

In this paper an efficient parallel co-location pattern mining approach proposed that effectively discovers colocation patterns and self co-location patterns based on constraint neighborhood approach. Also the drawbacks in previous approaches have been enhanced by using hadoop–MapReduce model which enable us with the parallel and distributed processing manner.

References

1. BanalataSarangi and LaxmanSahoo,"A Survey on Spatial Association Rule Mining Technique and Algorithms for mining spatial data,"International Journal of Scientific & Engineering Research, Vol. 4, Issue 12, pp.1664- 1670, 2013.

2. N. Lavrac, D. Jesenovec, N. Trdin, and N. M. Kosta, "Mining spatiotemporal data of traffic accidents and spatial pattern visualization," Metodoloskizvezki, Vol. 5, No. 1, pp. 45-63, 2008.

3. Tran Van Canh and Michael Gertz, "A Constraint neighborhood based approach for co-location pattern mining," Fourth International Conference on Knowledge and System Engineering, pp. 128-135, 2012.

4. H. Yang, S. Parthasarathy and S. Mehta., "A generalized framework for mining spatio-temporal patterns in scientific data," In Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 716–721, 2005.

5. VenkatesanMeenakshiSundarama, Arunkumarthnagavelub and PrabhavathyPaneerc,"Discovering Co-location Patterns from Spatial Domain using a Delaunay Approach",International Conference On Modelling Optimization And Computing, Vol. 38, pp. 2832–2845, 2012

6. SonaliSatijaand RajenderNath,"Performance Improvement of Apriori Algorithm Using Hadoop,"International Journal of Advanced Research in Computer Science and Software Engineering , Vol. 5, Issue 6, pp. 765-768, June 2015.

7. Gy˝ oz˝ o Gid´ ofalvi, "Spatio–Temporal Data Mining for Location–Based Services," Ph.D. Thesis, Faculties of Engineering, Science and Medicine at Aalborg University, Denmark, 2007.

8. G.Kiran Kumar, P.Premchand and T.VenuGopal, "Mining Of Spatial Co-location Pattern from Spatial Datasets,",International Journal of Computer Applications, Vol. 42– No.21, pp. 25-30, March 2012.

9. Lizhen Wang, YuzhenBaoa and ZhongyuLub,"Efficient Discovery of Spatial Co-Location Patterns Using the iCPI-tree," The Open Information Systems Journal, Vol. 3, pp. 69-80, 2009.

10. Jin SoungYoo, Douglas Boulware and David Kimmey,"A Parallel Spatial Co-location Mining Algorithm Based on MapReduce," IEEE International Congress on Big Data,Anchorage, AK,pp. 25 – 31, 2014.

11. S. Shekhar and Y. Huang, "Co-location Rules Mining: A Summary of Results," in Proceedings of International Symposium on Spatio and Temporal Database, pp. 236–256, 2001.

12. Rushirajsinh L. Zala, Brijesh B. Mehta and Mahipalsinh R. Zala, "A Survey on Spatial Co-location Patterns Discovery from Spatial Datasets," International Journal of Computer Trends and Technology (IJCTT), Vol. 7, No. 3, pp.137-142, 2014.

13. [https://github.com/hadoop-illuminated/hadoop-book] Accessed on 10/9/2015

14. Othman Yahya, Osman Hegazy and EhabEzat,"An Efficient Implementation Of Apriori Algorithm Based On Hadoop-Mapreduce Model,"International Journal of Reviews in Computing, Vol. 12,pp.59-67, 2012.

15. [https://hadoop.apache.org/docs/r1.0.4/mapred_tutorial.pdf] accessed on 2/8/2015

16. Jin SoungYoo and ShashiShekhar,"A Partial Join Approach for Mining Co-location Patterns," GIS'04, Washington, DC, USA, ACM, November 12–13, 2004

17. .

18. Y. Morimoto,"Mining Frequent Neighboring Class Sets in Spatial Databases," In Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2001.

19. Jin SoungYoo ,ShashiShekhar and Mete Celik, "A Join-less Approach for Co-location Pattern Mining: A Summary of Results,"5th IEEE International Conference on Data Mining, 2005.

20. Jin SoungYoo and DouglasBoulware, "A framework of spatial co-location mining on MapReduce," IEEE International Conference on Big Data, pp.44., 2013.

21. S. Park and J. S. Yoo, "Spatial Association Mining for Focal Events in Cloud Computing," In Proceedings of the International Conference on Advances in Big Data Analytics, pp.157-164, 2014.