SpatialHadoop: For OpenStreetMap Data

Kirandeep Kaur, MTech Student, GNDEC,

Sukhjit Singh Sehra, Assistant Professor, GNDEC,
Priyanka Arora, Assistant Professor, GNDEC
Sumeet Sehra, Assistant Professor, GNDEC

Abstract—With the change of time information related to geography and volunteered geography also changes. In this way extraction of spatial patterns from crowdsourced data has become most valuable for service suppliers. These patterns represent the spatial features of the co-related objects. The existing approaches used Dijkstras algorithm and Euclidean distance to find spatial patterns which can not compute accurately. Crowdsourced data is growing on daily basis through mobile phones, road networks and remote sensors. In order to process this type of large data set is also becoming difficult. In this research work we have proposed a system to process crowdsourced data taken from OpenStreetMap to mine the useful patterns using SpatialHadoop. SpatialHadoop has used Pigeon script, a spatial extension to Pig that is a high level language. These patterns will assist service providers to offer different sites based on facilities. In this extraction method, spatial data is loaded into the system and filtered for nodes, ways and relations. The filtered data is used for the mining process by using kNN joins. After this the evaluation of multiple resolution pruning filter with spatial datasets are generated using argument values. The different data sets have been checked using this methodology to extract the spatial patterns. This technique is compared with PostgreSQL and it is observed that SDM has provided more efficient results. The result obtained from experiment has shown the performance of our system that is better in comparison to the already existing systems in consideration of efficiency, speed and accuracy that rely on a network.

Index Terms—Spatial Data Mining, OSM, SpatialHadoop, Pigeon.

I. Introduction

SPATIALData Mining In the growing field of technology crowdsourced data is regarded as important as currency. Spatial Data Mining is the way to discover new useful and dif-ferent patterns from existing spatial data like OpenStreetMap. SDM has used to mine various spatial patterns such as traffic sensors, prediction models depending on place, spatial clusters, hotspots and associated patterns that includes the collection of different types of spatial events. Structure of SDM technique is given below in Figure 1:

Spatial data can be treated as network of GIS (Geographic Information Science) that is achieved from different resources using phones, telescopes and medical instruments. The am-plification of spatial data creates opportunity for specialized systems to manage big spatial data. Spatial data is increasing day by day and well supported by mapreduce systems like hadoop. Spatialhadoop uses its constructs in all layers of hadoop like language, storage and operations to provide its mapreduce functions.

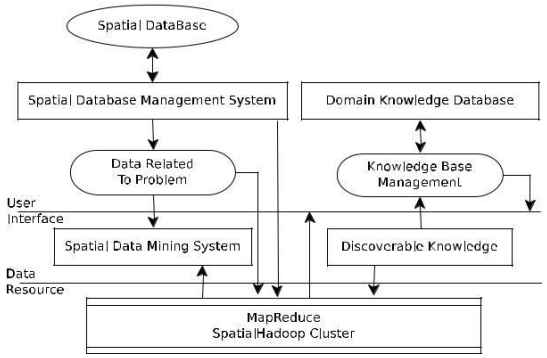


Fig. 1. Spatial Data Mining Architecture

A. Spatial Data and Big Data

Data is of two types: Spatial and non-Spatial. Spatial data is available for mining of useful and interesting patterns. Spatial data can be treated as network of GIS (Geographic Informa-tion Science) that is achieved from different resources using phones, telescopes and medical instruments. The amplification of crowdsourced data has created opportunity for specific systems to manage big spatial data. Spatial data is increasing day by day and well supported by mapreduce systems like hadoop. Spatialhadoop has used its events in the layers of hadoop like language-pigeon, storage-cluster and operations to provide its mapreduce functions. This mechanism is to apply on regular algorithms to the large sets of data to extract the relevant information.

The sets of data is correlated to business or market is famed as Big Data. This term has brought up the data sets or mixture of data to include its mass, variance & speed and compose them to manage through formal technologies and different tools like relational databases or perception, within the required time to make them useful. The Classical and Statistics approaches differentiate between the classical data mining and the spatial data mining technique. The base of business intelligence is to transfer the data that tends increased value to the enterprise.

B. Spatial Patterns

Spatial patterns are analyzed depending upon the neighbor distance [1]. A spatial pattern provides details about certain

978-1-5090-4715-4/17/$31.00 ©2017 IEEE

areas based on interest of users in terms of OSM road diversity. A spatial neighborhood pattern has included a arrangement of features that are placed nearby in spatial closeness. This pat-tern mining has also put in discovering new spatial habituation of different osm objects [2]. This spatial habituation is a trend of discovered objects which are situated close to others in the geographic region that produce a high degree of resemblance or non-resemblance between them. These computed instances in space depend on their spatial neighborhood definition using threshold. In terms of closeness, they has been defined by various types of distant metrics.

C. SpatialHadoop

SpatialHadoop is an open source platform which helps in co-location pattern mining to carry out spatial operations. A spatial hadoop cluster has one node as master that divides a map-reduce work into small jobs and transmit by slave nodes. SpatialHadoop has assembled in Hadoop that proceed to spatial design in the origin of Hadoop and make effective with processing of large data sets [3]. It includes three type of users to communicate with spatial hadoop. They can be casual users, developers and system admins given below in Figure 2.

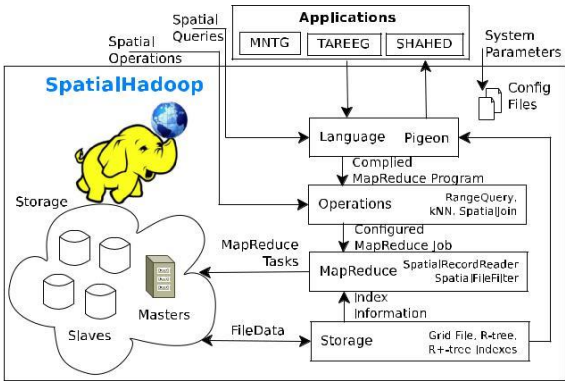


Fig. 2. Spatial Hadoop Architecture [4]

SpatialHadoop is efficient tool in distributed environment. This is based on a computational prototype known as MapRe-duce which is dependent on map and reduce functions. Com-putation operations can be divided into fragments and its distribution among the cluster nodes is done in map phase. A reduce phase which merges all given values related to the same key. Spatialhadoop has four layers-

1) First layer of spatialhadoop is called Language layer. Spa-tial hadoop does not have its own language. It provides pigeon extension to pig which is high level language. Pig-Latin is taken from Latin word. This has compatibility with Open Geospatial Consortium (OGC) standard and make it easy to use specially for non-technical users, fa-miliar to use PostGIS tools. Pig has support for significant parallelization and capability to process huge data sets.
2) The Storage layer has three spatial indexes, Grid File, R-tree and R+-tree. These all are implemented inside the Hadoop Distributed File System (HDFS). Indexes are organized in two-layers global index and multiple local indexes to organize records inside each node.
3) MapReduce layer consist two new segments Spatial-FileSplitter and SpatialRecordReader which allow spatial operations to access the constructed indexes.
4) Operation layer encapsulates the spatial operations given by SpatialHadoop. It has three basic spatial operations-range query, kNN and spatial join.

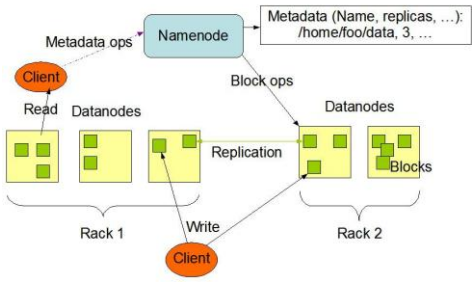


Fig. 3. Hadoop Distributed File System

D. Pigeon

Pig is high level language, developed by Yahoo Research in 2006. Its structure is divided into two different layers. An infrastructure layer having built in compiler that can spawn MapReduce programs and next one language layer called Pig Latin consists text-processing language [5]. Pigeon extension is added to Pig for spatial queries and operations in distributed cluster of spatialhadoop as shown in Figure 4. Pigeon includes lesser number of code lines as compared to any other language. 10 lines of Pig code is equal to 200 lines of JAVA. It is similar to SQL but specially designed for Hadoop and SpatialHadoop cluster.

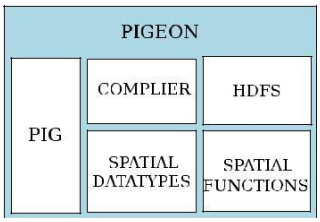


Fig. 4. Overview of Pig [5]

SpatialHadoop has used the keywords join and filter. This framework has spatial predicate for performing range query and spatial join [6], respectively. Spatial data types has ex-plained the schema of an incoming file during loading. The primitive functions such as MBR, distance and overlap, are applied to spatial characteristic for measuring the distant having main centroid of two shapes and provision to check that they overlap each other or not, and compute the shape of a polygon [7].

II. OpenStreetMap Data Sets

OpenStreetMap has been founded in 2004. This is free editable data. This provides high resolution images and trans-parency of data to copy, edit or its distribution [8]. OSM has created a unique curriculum that is focused on different topics

like mapping, crowd sourcing and open source technologies [9]. It can be used for single directional and unidirectional Road and Street mapping along with their dimensions, direc-tions for the particular area. Further OpenStreetMap data has many open problems that would create valuable research work [10]. These terms require investigation and some areas to be worked on by researchers as tasks that is different from the base of open source expansion progress proceeding in any community. Maps are visual symbols of our world [11]. OSM is also used to develop the 3D models of buildings. This can be of two types:

Vector data model This requires points, lines, and poly-gons to present world data. These models store data that is limited to boundaries such as country borders, streets and land parcels.

Raster data model In this, world surface distributed with a cell of regular grids. These are helpful for storage of data that continuously differs, as in an aerial photograph, a image of space station or elevation surface.

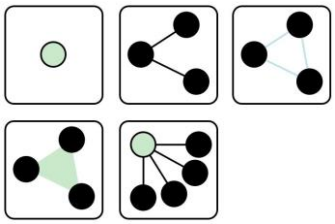


Fig. 5. OSM Elements

A. Elements of OSM Data

Three types of elements are defined by OpenStreetMap data refering to Figure 5 as shown above.

1) Nodes: Node is base in OpenStreetMap data and consist points illustrated by latitude and longitude. Every node in OSM has unique id and coordinate points that define point of interest. Ex. Park bench. Syntax is:
   Node <Node_id , Latitude , Longitude>
2) Ways: It defines lines known as polylines to present rivers or lakes and boundaries through polygons defined as closed ways. These ways make connection between 2 and 2000 nodes to form polyline. Ex. Rectangle and Polygon.
3) Relations: Relation is data structure that forms the re-lation between data elements and it defines routes and multipolygon. Spatial relations can be of three types: topological, distance and direction that express the more complex neighbourhood.

B. Tags

OSM tag constructed of two keywords one is key and second is value. It contains strings up to 255 characters which are Unicode. The most common attributes being used in OSM are described in Table I.

Example: highway=residential leisure=park describes the way to define connection between people houses.

TABLE I
ATTRIBUTES OF OSM TAGS

Name	Value	Description
Id	Integer	Each element has unique Id in integer format
User	String	Last modified user name
Timestamp	Time and Date Format	Time of last modification
Version	Integer	A version of an object is 1 for a new version of the ob-ject and that is incremented by the server
Visiblity	False/True	When the object has moved and defines visiblity=false

There is no particular language of OSM tags. In OSM, various locations are represented using tags by combination of nodes, ways and relations.It includes many standards being set up by the community centers that commonly used.

III. K-NEAREST NEIGHBOR QUERIES(KNN)

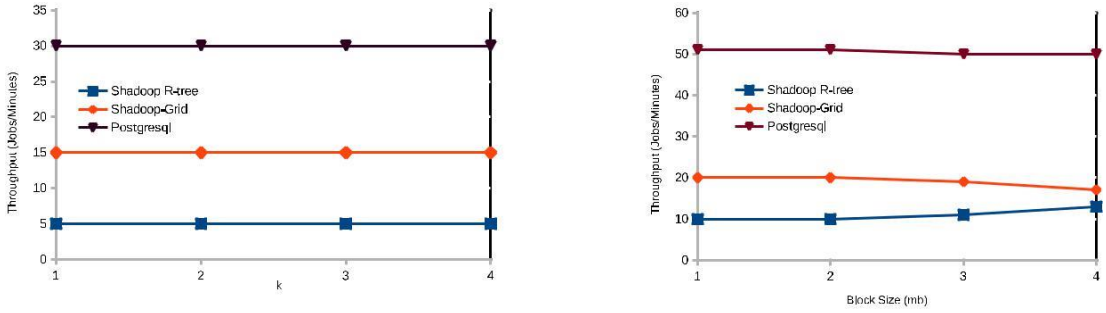
kNN is a classification algorithm, it gives fast performance in short span of time. In kNN each neighbor has its own importance to define a set of close neighbors. k nearest neighbor join is suitable for given dataset of defined object in other dataset. An operation is followed by SDM applications in kNN join.

Figure 6a has given effect of increasing k from 1 to 1000 on OSM dataset and Figure 6b shows the performance based on block size of data. For large volume of data, performance of kNN join on a cluster becomes effective by using MapReduce framework for these over clusters of machines [19].

IV. RELATED WORK

MapReduce framework supported for crowdsourced data efficiently in SpatialHadoop that is an extension of Hadoop. SpatialHadoop also has four layers known as language, in-dexing, query processing, and visualization [4]. Each layer has its own speciality. The language layer is a high level to support for all spatial data types and their operations for non-technical users. Second layer, the indexing layer has used with standard spatial indexes like grid, R-tree, to export up the spatial operations in Hadoop environment. Third the query processing layer has encapsulated the operations such as knn, range query, and spatial joins by spatialhadoop. And the last layer, the visualization layer has permitted permission to the programmer images that had defined very large datasets and easy to export spatial data [11]. Real systems such as MNTG, TAREEG, TAGHREED, and SHAHED has processed spatialhadoop as a main part for spatial data mining.

978-1-5090-4715-4/17/$31.00 ©2017 IEEE



(a) (b)

Fig. 6. kNN Performance

HadoopGIS is built on Hive, which is a data warehouse not useful for traditional MapReduce programs [15]. Flexibility is not supported by these systems because they are limited with their functions. Main drawback is the lack of integration with hadoop in these systems.

Big spatial data is animated the existing systems to use the functionality provided by these systems. In big spatial data, there are two aspects to understand [16]. It had proved that features of extraction had applied in many systems to make this easy to select the most suitable approach based on the system architecture. In case of non-spatial data, distributed systems are rising, they had required more facilities to extend the systems that support the spatial data.

Parallel data processing is a tool for MapReduce review that signed to serve the data and open source areas in perceptive of several technical prospects of the this framework [17]. In this approach, they had explained and characterized the MapReduce framework and talked over its integral advantages and disadvantages.

Apache developed hadoop which is known as for its efficiency towards large data sets. Hadoop has worked on mapreduce, a computational paradigm used for map and reduce the functions [18]. Map means division of operations into many fragments within cluster nodes. Map is resulted combined to reduce the output in reduce phase.

Pig is member of Hadoop ecosystem act as a procedural version of SQL [7]. Pig is very simple and high level language. This is user-friendly and flexible in its syntax and variable allocation. It has added extensibility for the framework using user-defined functions. It has processed for unstructured data to preprocess this in parallel form.This language is similar to the scripting languages like Perl.

Collection of data continuously growing that has caused im-feasible to produce this data by traditional methods [14]. It had required new techniques and tools that can provide sponsorship to user in translating the data into spatial knowledge with new research area called knowledge discovery process in databases.

V. Pattern Distribution

In this research work we have introduced MapReduce components to find routes, directions from crowdsourced data. This work has required the deployment of toolkit like SpatialHadoop with Pig language to mine the patterns. Pig is implemented using KNN algorithm that compiled Pig Latin to make a place between the procedural style of SQL and declarative style of map-reduce to join the points or lines for the road network. This work has following features:

It finds the number of neighborhood objects in the specific region.

It detectes neighbors to make relations between nodes and ways.

It finds the patterns that formed by nearby objects based on distance.

This technique deals with OSM data and requires a crowd-sourcing data set to mine specific patterns. Processing of large data set is very time consuming process but using hadoop and pigeon it has performed fastly. Pig Latin language is used for Spatial Data Mining with kNN joins. This method is used for street mapping to mine patterns along with the dimensions and directions of particular area to select the end points and give them shape of lines or points to meet the requirements of road network as described in Figure 7. The following steps are proposed to achieve the given objectives:

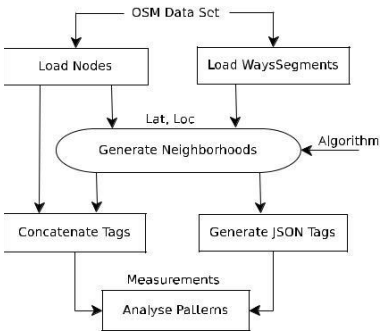


Fig. 7. Experimental Setup and Design

VI. Methodology

The identification of spatial patterns is carried in this re-search work using co-location pattern mining technique shown in 9. Firstly data is fetched from OSM into the system. This data is filtered by nodes and ways and mapped to mine

patterns. The following steps are used to mine spatial patterns from OSM data.

This work has been carried out to accomplish these ob-jectives as to deploy SpatialHadoop toolkit. Development of spatial data mining technique for Crowdsourced data and compare with other contemporary techniques.

A. Load data from OSM into system

Firstly we have taken a dataset from OpenStreetMap [10] in .osm or .xml format and placed into Hadoop Distributed File System. It is used by mapreduce paradigam to get input. This data is accessed by grunt shell in query processing system. Connecting nodes and ways with their respective IDs produce shapes of the ways. The data set of Ludhiana is loaded Figure 8 taken from [20].

Fig. 8. OSM data on Map

B. Filter the Spatial Data

After this data is filtered by their respective IDs based on latitude and longitude. We have joined the filtered nodes and ways to create shapes.

C. Grouping nodes and ways

After filtration of spatial data, it is processed per tuples by grouping of joins. Way IDs are grouped with joins to produce shapes of patterns which are created to define specific patterns. The detailed procedure is shown in Flowchart of SDM technique Figure 9. This is based on particular points and lines to form the shapes.

D. Colocate Segments and tags

Pattern shapes defined the location. It flattens to produce IDs per line and location of tags, cogroup them into the specific category. Result has been called again by relation ID and function connect.

E. Locate Patterns

Filtered segments are used to define tags by concatenation using following line of code based on latitude and longitude of objects. Then result is stored in HDFS. The output file is obtained from HDFS and compared to the existing techniques with specific patterns that have been found.

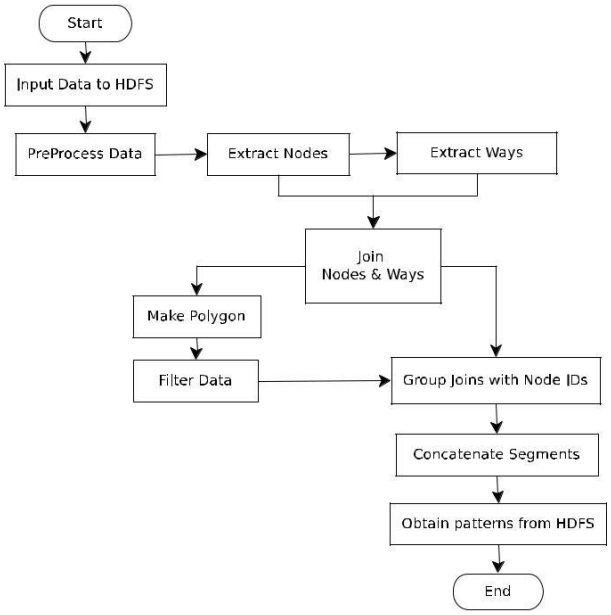


Fig. 9. Flowchart of SDM method

Algorithm 1: Algorithm for Co-location Pattern Mining

Input: Values
ts = threshold distance of neighbor relationship; DC = a set of tuples having two points;
TC = transactions of near by objects ;

Output:
DT: a set of near by objects;
R: is the complete set of patterns;

Steps
DT Empty
for each original node n of N do
DC Empty
DC NodeExpansion(n;ts)
for tuple < n;n^0;(n;n^0) > inDC do
TC Empty
TC NeighborOb jects(< n;n^0;(n;n^0) >;ts)
Insert elements of TC into DT
end for
end for
return DT

VII. Proposed Algorithm

We have propsed a algorithm to generate neighborhoods. This is used to measure the distance between objects.

This work has been tested on [75.484, 30.618, 76.309, 31.141] defines latitude and longitude of a location, taken from [20]. Data set of Ludhiana city has been used which is given above. It gives number of objects to define patterns in this region. It is also checked on world's road data taken from SpatialHadoop (official website) data sets. Both data sets are different in size. And this process has taken very less time to compute

978-1-5090-4715-4/17/$31.00 ©2017 IEEE

this data.

VIII. Results

The results has obtained by extraction process, give details about number of objects in following Table II. It shows the number of objects in given data set using spatial data mining process.

TABLE II
Number of Objects with facility types

ID	Facilities	Total Objects
1	Banks	158
2	ATM	278
3	Hotels	185
4	Parking Lot	166
5	Super Market ans Mall	28
6	Schools	579
7	Library	15
8	Colleges	85
9	Hospitals	225
10	Pharmacy	1002
11	Ambulance	125
12	Gurudwara	368
13	Temples	503

In this work SpatialHadoop has support for crowdsourced data types and operations. This has adapted spatial structures such as R-tree, R-+tree and Grid for index creation. As compared to PostgreSQL rule based data mining technique, our proposed system has used SpatialHadoop which is a new technique for data analysis to mine spatial patterns, result has obtained in less time. This comaprison is based on iterations of knn algorithm in spatialhadoop and postGIS. For spatial join the two lables of a data, being joined to obtain the same result, but in different time intervals. The below given Figures 10a and 10b have shown the performance of spatialhadoop and postgis in different data sets.

By comparing these spatial data mining techniques it proved that spatialhadoop has taken less time to mine spatial patterns as compared to postgis by use of spatial joins such as kNN joins around distance of 200m.

The participation index (PI) is calculated using observed data. In case of co-location pattern Participation index is defined as

P = {p1, p2,...,pn},

and the participation index can also be described as PI(P) is defined as

pi P {PR(P,pi)}.

The comaprsion table of participation index is given which defines the neighborhood approach and efficiency of both methods to associate the co-located objects based on facilities in Table III:

Co-located neighbors of above facilities in neighborhood relation has been represented in Figure 10 which also shows the different type of patterns extracted by using spatialhadoop based on co-located mining algorithm in short span of time.

TABLE III
Co-location Patterns with neighbor distance of 200 m

ID	Co-location	PostgreSQL(PI)	SHadoop(PI)
1	(Banks,ATMs)	0.705	0.562
2	(ATMs,Hotels)	0.696	0.526
3	(ATMs, SuperMarket and Mall)	0.985	0.756
4	(Hotels,Parking Lot)	0.621	0.512
5	(ParkingLot, SuperMarket and Mall)	0.785	0.603
6	(Schools,Library)	0.674	0.523
7	(Library,Colleges)	0.752	0.514
8	(ATMs,Colleges)	0.435	0.306
9	(Hospitals,Pharmacy)	0.561	0.425
10	(Hospitals,Ambulance)	0.489	0.366
11	(ATMs,Hospitals)	0.389	0.245
12	(Gurudwara,Temples)	0.397	0.312

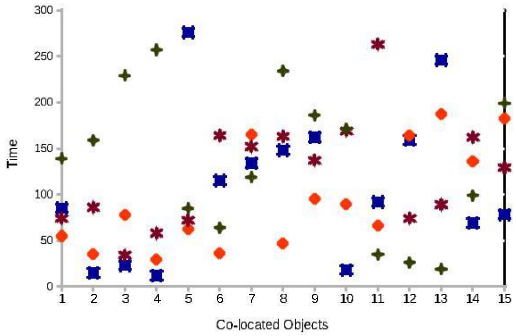
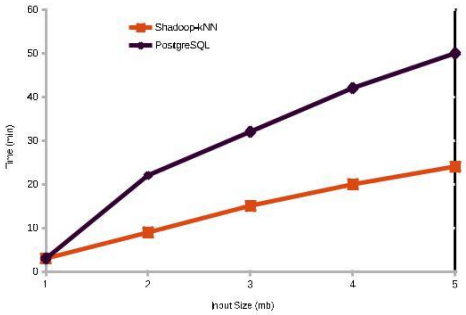


Fig. 10. Different Co-located Patterns

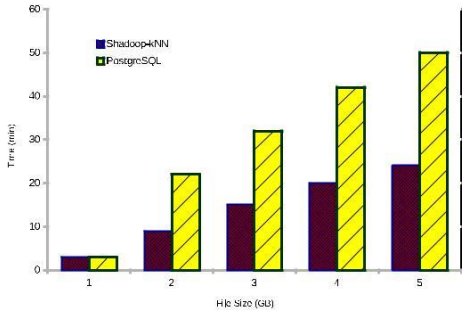
These patterns has described the number of facilitis found in particular area of mining process.

These tools made OSM data more operative and provide functionality for this extraction and co-location of pattern. The different parameters has been used in this proposed work given below in Table IV.

Due to large spatial data sets SpatialHadoop provide high efficiency. Co-location pattern mining is important for service providers and for this work SpatialHadoop is becoming popu-lar tool in IT industry. Different sizes of data can be processed using this technique with spatial extensions. This is growing field for researchers to mine patterns in different forms based on latitude and longitude. This technique has been tested on latitude 75.484 and longitude 30.618 as current location and its nearby areas in bounding box.



(a) Spatial Join Performance (mb)



(b) Spatial Join Performance (GB)

TABLE IV
COMPARISON OF POSTGRESQL VS. SPATIALHADOOP

Tools	PostgreSQL	SpatialHadoop
Database	OSM Data	OSM Data
Size	Upto 10 GB	Up to 10 GB
Speed	Normal	High
Time for kNN Join	1 hour 20 min	32 min
Time to make Poly-gons	1 hour	20 min

IX. CONCLUSION AND FUTURE WORK

Crowdsourcing is the part of OSM in the field of spatial data analysis. It has been observed that mining of spatial patterns from spatial data is used for location services. This field has brought many hypothesis and exceptions to characteristics of OSM data. This work is based on mining of special patterns near around 200m by using k nearest neighbor distance. These spatial patterns are useful to find particular region facilities for the service providers. In this research, possible use of co-location pattern technique was demonstrated with tools such as SpatialHadoop and Pigeon language to process large data sets. This application has considered efficient for neighborhood relation between geographical objects.

Efforts have been made to implement this technique for mining of spatial patterns from OSM database but still there is a scope of improvement. Some improvements that can be taken up into consideration for future work are as Apart from co-location pattern mining other techniques can also be taken into consideration. Co-located patterns can be used by service providers. Some Issues like time factor and efficiency can also be taken into consideration for OSM data.

REFERENCES

[1] W. Yu, "Spatial co-location pattern mining for location-based services in road networks," Expert Systems with Applications, vol. 46, pp. 324–335, Mar. 2016.

[2] C. Sengstock, M. Gertz, and T. Van Canh, "Spatial interestingness measures for co-location pattern mining," in Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on. Brussels: IEEE, Dec. 2012, pp. 821–826.

[3] "SpatialHadoop," 2013, accessed 10-April-2016. [Online]. Available: http://spatialhadoop.cs.umn.edu/

[4] A. Eldawy, "SpatialHadoop: towards flexible and scalable spatial processing using mapreduce," in Proceedings of the 2014 SIGMOD PhD Symposium, Lake City, UT, June 2014, pp. 46–50.

[5] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins, "Pig latin: a not-so-foreign language for data processing," in Proceedings of the 2008 ACM SIGMOD international conference on Management of data. Van-couver, Canada: ACM, 2008, pp. 1099–1110.

[6] S. Zhang, J. Han, Z. Liu, K. Wang, and Z. Xu, "SJMR: Parallelizing spatial join with MapReduce on clusters," in IEEE International Conference on Cluster Computing and Workshops. New Orleans, LA: IEEE, Sept. 2009, pp. 1–8.

[7] A. Eldawy and M. F. Mokbel, "Pigeon: A spatial MapRe-duce language," in In Proceedings of the IEEE Interna-tional Conference on Data Engineering. Chicago, IL, USA: IEEE, March 2014, pp. 1242–1245.

[8] Y. Zhang, X. Li, A. Wang, T. Bao, and S. Tian, "Density and diversity of OpenStreetMap road networks in China," Journal of Urban Management, vol. 4, no. 2, pp. 135– 146, Dec. 2015.

[9] S. S. Sehra, J. Singh, and H. S. Rai, "A Systematic Study of OpenStreetMap Data Quality Assessment," in Information Technology: New Generations (ITNG), 2014 11th International Conference on. Las Vegas, NV: IEEE, Apr. 2014, pp. 377–381.

[10] "OpenStreetMap," 2004, accessed on 3-March-2016. [Online]. Available: https://www.openstreetmap.org/ #map=11/30.7949/75.8767

[11] J. Jokar Arsanjani, A. Zipf, P. Mooney, and M. Hel-bich, "An Introduction to OpenStreetMap in Geographic Information Science: Experiences, Research, and Ap-plications," in OpenStreetMap in GIScience. Cham, Switzerland: Springer International Publishing, 2015, pp. 1–15.