JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 15, 2020

GEO-SPATIAL DATA ANALYSIS OVER OPEN STREET MAP USING LATENT SEMANTIC ANALYSIS AND GENETIC ALGORITHM

Sonam[1]
MTech. Student, Dept. of Computer Science & Engineering R.D. Engineering College at Duhai, Ghaziabad, India

Prof. Jaideep Kumar[2]
Professor & Head of Department, Dept. of Computer Science & Engineering R.D. Engineering College at Duhai, Ghaziabad, India

Received: 14 March 2020 Revised and Accepted: 8 July 2020

Abstract:
The expansion or improvements of metropolitan areas and rural area is moderately speedily going on, particularly the metropolis of India like (Delhi, Mumbai and more), in addition, amplify the requirement of information retrieval about the said location like transportation infrastructure, commodities, hospitality, business opportunities, and agricultural aspects. Escalating demand for such resources is very essential for the sake of progression and development for masses therefore, to omit the obstacles or to remit decrease the level of service. The proposed scheme uses Machine Learning Techniques i.e. Latent Semantic Analysis and Genetic Algorithm which inculcates for swift and prompt information espionage or information retrieval system from huge corpus or map repositories. However, the supervision model is prepared based on the investigation of service derived from the ratio of data volume and capacity in context to locations and references. The method used to obtain data from geometric data and land use are done with image interpretation and measurement vector data as XML from OpenStreetMap. The result of the study presents a level of accuracy of approx 79% the scheme mainly uses geometrical data available on OSM to extract land use data for information retrieval effectively.

Keywords: OpenSteetMap, Machine Learning, Latent Semantic Analysis, Genetic Algorithm, Singular Value Decomposition, Xtensible Markup Language.

1. INTRODCUTION

Geo-Spatial analysis or spatial information comprises one prescribed technique that studies entities using their topological, geometric, or geographical assets. "Spatial investigation comprises various procedure, many of which are still in their initial expansion, with dissimilar methodical procedures applied in fields such as Botany, Forestry, Horticulture and Agriculture with the study of the placement of Botany, Forestry, Horticulture, and Agriculture with the use of "places and routes" algorithms to build structures complex cable. In a more limited sense, spatial analysis is a technique applied to structures at the human scale, especially in the analysis of geographical data. Complex issues arise in spatial analysis, many of which are not clearly defined or not fully resolved, but form the basis for current research. The most basic of these is the problem of determining the spatial location of the entity being studied." The classification of spatial analysis techniques is difficult because a large number of different fields of research are involved, various essential procedures can be chosen, and various forms of data can be taken. [1-2]

Geo-Spatial analysis might be considered [3] has emerged with initial efforts on cartography and surveys but many fields have contributed to its improvement in modern forms. Biology contributes through botanical studies on global plant distribution and location of local plants, studies of animal movement ethnology, ecological landscape studies of vegetation blocks, spatial population dynamics ecological studies, and biogeography studies. Epidemiology contributes with early work on disease mapping, especially John Snow's work on mapping cholera outbreaks, with research on mapping the spread of disease and with site studies for health care delivery. Statistics has made a major contribution through work in spatial statistics. Economics has contributed mainly through spatial econometrics. Geographical information systems are currently the main contributors because of the importance of geographic software in modern analytic tool boxes. Remote sensing has contributed extensively to morph metric analysis and grouping. Computer science has contributed extensively through the study of algorithms, especially in

computational geometry. Arithmetic with persistent approaches can present basic tools for investigation and to uncover the complication of spatial space, for example, with modern work on fractals and invariant scales. Scientific modeling provides a useful structure for new mechanisms [1-3].

Geo-Spatial analysis faces numerous essential problems in the classification of the object of research, in the development of analytical procedure to be utilized, however in the utilization of central processing unit (computer) for analysis, in the restrictions and peculiarities of known analysis, and in the arrangement of methodical results. Numerous of these problems are active matter of contemporary investigations. Ordinary mistakes frequently happen in spatial analysis, some because of space mathematics, some because of the way data is spatially presented, some because of the available tools. Census data, because it protects individual privacy by combining data into local units, raises a number of statistical problems. The fractal nature of the coastline makes measuring exact length difficult if not impossible. Computer software that attaches a straight line to a coastline curve can easily calculate the length of the line it determines. However, this straight line may not have meaning inherent in the real world, as shown for the maps.

These problems are challenges in spatial analysis because of the power of maps as a medium of presentation. When the consequences are obtainable as a map, the appearances coalesce spatial data that is generally accurate with analytic results that may be inaccurate, which leads to the notion that analytic results are more accurate than the data to be shown.

The definition of an entity's spatial existence limits the probable analysis that can be functional to the entity and persuade the final conclusions that can be reached. While this possessions is essentially true of all analyzes, it is very important in spatial analysis because the tools for defining and studying entities support the specific categorization of the entity being studied. Statistical techniques prefer spatial definitions of objects as points because there are very few statistical techniques that operate directly on line, area, or volume elements. PC tools sustain the spatial characterization of objects as harmonized and separate elements due to the limited number of database elements and computational structures available, and the ease with which these primitive structures can be created [3-5].

Open Street Maps: OpenStreetMap (OSM) [12-16] is a GIS Web product that can also be operated using web browsers and smartphone so that it can be categorized smart GIS. The use of smart GIS really helps the mapping process because it is more efficient in terms of time, device, and can be taken to any area. Through the Open Data Commons Open Database License 1.0, OSM contributors can own, modify, and share map data widely. There are various types of digital maps available on the internet, but most have legal and technical limitations. This makes the community, government, researchers and academics, innovators, and many other parties unable to freely use the data available on the map. On the other hand, both the OSM base map and the data available in it can be downloaded for free and open, for later use and redistribution. Thus it is hoped that utilizing OSM will become an alternative source of data, especially for mapping urban areas and rural areas, one of which is related to locations and directions. (OpenStreetMap, 2019).

2. LITERATURE REVIEW

OpenStreetmap is created by a mutual map community contribute and maintain data on roads, trails, cafes, stations, and many other things throughout the world. OpenStreetmap emphasizes local knowledge. Donors use aerial photography, devices GPS, and field maps to verify OSM accuracy and always updated [12]. Openstreetmap users continue to grow every year. The development of urban areas that is quite rapid, especially in the cities, helped increase the need for transportation facilities and infrastructure. The increasing number of motorized vehicles has an impact on the decreasing level of road services. Road management is based on the results of analysis of the level of road services obtained from comparison of traffic volume and road capacity. The method used to obtain data that is geometric data of roads and land use is done by interpreting images and measuring vector data from OpenStreetMap [12], as well as field survey activities. The results of the study present the level of accuracy of vector data, especially geometric roads and the ability of imagery available in OSM to tap land use data. Each of the accuracy test results shows that vector and raster data in OSM are suitable as alternative data sources for mapping road services. In general, road conditions in city have a poor service level, which is below class C to F. Road management recommendations given in general are traffic lights, parking, road markings, and road geometry improvements.
The OpenStreetMap project (abbreviated OSM) was launched by Steve Coast in summer 2004 founded in London. The aim was to collect data for a free card. The back then available cards were either expensive or were not under a

free license available for any use. So there is a similar one behind OpenStreetMap Motivation like behind free software - the desire to be able to do what you want with it wants [13]. A large number of volunteers - on November 14, 2016 there were 3 199 742 user accounts [14] - compiled the data. Only some of these users ever have edited an OSM object. In December 2011 there were 505,000 users only 38 percent edited an object [15]. As of July 31, 2019, there were approximately 562,000 of the 2.2 Millions of users (26 percent) who had ever edited an object [16]. The OpenStreetMap Foundation has been behind the OpenStreetMap project since 2006, which operates the server and since switching from the Creative Commons license Attribution Share under the same conditions 2.0 to the Open Database License 1.0 (ODbL) in September 2012 also holds the rights to the data and it under the Conditions of the ODbL available to everyone [16, 17, 18]. Unlike the Wikipedia, OpenStreetMap has no relevance criteria, but only records things that are observable on site and not directly personal Data like doorbell signs. Exceptions to this rule only exist for things which are considered important for a card, such as B. Administrative limits OpenStreetMap started from scratch in many places, only in some areas was early Years of uniform import coverage of data from external sources. In the other areas (especially in Europe) there has been a powerful one Community formed, which records the data manually on site and their focus is located in German-speaking countries (from the start of the project to December 2011, 31 percent of all users had data mainly in Germany, Austria or Switzerland recorded [19]). Bing Maps aerial photos have been available since November 2018.

LSA (Latent Semantic Analysis) is a procedure of analyzing a document to find the meaning or concept of the document by comparing the semantic similarities. The meaning or concept of the words contained in the writing will be a reference comparison without looking at the linguistic characteristics of a writing. The LSA method maps the words or documents to a concept space and comparisons are made on this space. The concept space or more commonly referred to as latent semanticspace is the result of mapping from a high dimensional matrix to a smaller dimension. Although in smaller dimensions, the matrix is a matrix that represents the contents of the whole document. The hallmark of LSA is a technique called Singular Value Decomposition (SVD). SVD is used to perform matrix decomposition after weighting and then measure its similarity with the data to be tested [25]. In 1990 through a journal titled "Indexing by Latent Semantic Analysis" by Scott Deerwester, Susan Dumais, George Furnas, Richard Harshman, Thomas Landauer, Karen Lochbaum and Lynn Streeter, an algorithm was introduced to index words in documents and plot them into a vector base that called Latent Semantic Analysis (LSA)[26]. LSA algorithm is one of the development algorithms from the field of Information Retrieval, which is able to collect a large number of documents in a database and connect relationships between documents by matching the given query. More specifically, the LSA algorithm is a method of making vector-based terms (terms) that are considered capable of capturing the semantics of a document or sentence. The main function of this LSA is to calculate the similarity (similarity) of documents by comparing the vector representation of each document. In forming vector-based term representations, LSA will form a matrix that represents the relationship between terms and documents called semantic spaces, i.e. words and documents that are closely associated will be placed close to each other represented by vectors. LSA in its calculations using Singular Value Decomposition (SVD)[27]. SVD represents semantic space in the form of matrices that have smaller orders than the original matrix order, but matrix calculations still produce matrices that are almost the same value. SVD is a linear algebra theorem which is said to be able to break the block of a matrix A into three new matrices, namely an orthogonal matrix U, diagonal matrix S, and Transpose matrix orthogonal.

The LSA method maps the ways or nodes to a concept space and comparisons are made on this space. The concept space, or more commonly referred to as latent semantic space, is the result of mapping from a high dimensional matrix to a smaller dimension. Although in smaller dimensions, the matrix is a matrix that represents the contents of the whole document. The hallmark of LSA is a technique called Singular Value Decomposition (SVD). SVD is used to perform matrix decomposition after weighting and then measure its similarity with the data to be tested [28].
The “concept of Latent Semantic Analysis (LSA) is an IR method that constructs the structure of document collections in the form of vector spaces using linear algebraic techniques, namely singular value decomposition. According to [27], in general the LSA concept includes several points as follows: “

1. *Text Operations on Queries and Document Collection*: Queries from users and document collections are subject to text operations processes. The text operations process includes,
a. Parse each word from the document collection.
b. Discard words that are stop words.
c. To stem the words for the next process.

2. *Matrix Creation*: The results of text operations imposed on the document collection are subject to the matrix creation process. The matrix creation process includes,
a. Count the frequency of occurrence of words,
b. Construct a word-document matrix, where the matrix row shows the word and the matrix column shows the document. For example, the matrix element in row 1 and column 2 shows the frequency of occurrence of the 1st word in the 2nd document.

3. *SVD Decomposition:* The word-document matrix formed, A in size m × n, is then subject to SVD (singular value decomposition) decomposition. The SVD results are in the form of 3 (three) matrices, so that matrix A can be written as: A = U S V T. To simplify the explanation, say u1, u2, ..., uk are column vectors of the matrix U, σ1, σ2, ..., σk are entries in the main diagonal of the matrix S, and v1, v2, ..., vk are column vectors from the V. matrix. The rank of the matrix A, k is the number of nonzero entries located on the main diagonal of the matrix S, namely σ1, σ2, .. , σk. k is also the number of singular values of A. d. From k singular values of A, r is chosen the largest singular value, that is σ1 ≥ σ2 ≥ ..., σr> 0, with r <k. Linear algebra that has many internal functions document and text processing. SVD is known as a very powerful technique, with regard to solving problems of equations or matrices, both singular and numerically approaching singular. The advantage of SVD is the ability to be used on all real sized matrices (m, n). If A is real matrix with size m x n and decomposed with Singular Value Decompotition (SVD) becomes:

$$A = USV^T(eq.1)$$

Where:-
U = orthogonal matrix measuring mxm, which is the AA^T eigen vector
S = diagonal matrix containing a singular value of size mxn, which is the square root of the eigen value of AA^T and A^TA
VT = orthogonal matrix of size nxn, which is an eigen vector from A^TA
Suppose that A can be written with column vectors, which are also word vectors for each document, namely:

$$A = [\vec{a_1}\vec{a_1} \ldots . \vec{a_1}] with \vec{a_1} sized m * 1, i = 1,2, \ldots \ldots, n. (eq.2)$$

For the U matrix, it is written using column vectors, which are:

$$U = [u_1\vec{u_2} \ldots . \vec{u_r}] with \vec{u_1} sized m * 1, i = 1,2, \ldots \ldots, r. (eq.3)$$

The S matrix is sized r * r, namely:

$$S = [\sigma_1\ 0\ \ldots\ 0 \vdots \vdots \ddots \vdots\ 0\ \sigma_3\ \ldots\ 0\ 0\ 0\ \ldots\ \sigma_r\ ], where\ \sigma_1 is the singular value i = 1,.2 \ldots, r (eq.4)$$

and matrix V is written using row vectors, which are:

$$V = [V_1\ V_2\ \vdots\ \vec{v_n}], where V_1 sized 1 * r, i = 1,2, \ldots, n. (eq.5)$$

Genetic algorithm as a branch of evolution algorithm is an adaptive method commonly used to solve a search for value in an optimization problem [32]. This algorithm is based on the genetic processes that exist in living things, namely the development of generations in a natural population, gradually following the principle of natural selection. By imitating this theory of evolution, genetic algorithms can be used to find solutions to problems in the real world. Genetic algorithm was first introduced around 1975 by John Holland in his book entitled "Adaption in Natural and Artificial Systems" [33] and then developed with students and colleagues [33]. This algorithm works with a population consisting of individuals, each of which represents a possible solution to an existing problem. An individual is represented as a collection of genes called chromosomes. By using genetic algorithms, the resulting solution is not necessarily a the exact solution of the optimization problem that was solved. Some important definitions in genetic algorithms are as follows [34-40]:

• Genes: A value that states the basic unit that forms a certain meaning. In genetic algorithms, this gene can be a binary value, float, integer or character, or combinatorial.

• Chromosome: Is a combination of genes that make up a particular value.

• Individual: State a value or state that states one possible solution of the problem raised. In some problems that can be solved by genetic algorithms, this individual can also be a chromosome itself.

• Population: Is a group of individuals who will be processed together in one cycle of the evolutionary process?

• Generation: Express one cycle of the evolutionary process or one iteration in the genetic algorithm.

In an optimization problem, there is an objective function or an objective function which is an evaluator function or the function that you want to optimize. Whereas in the genetic algorithm, the function is called the fitness function; Each individual has a certain fitness value. Before genetic algorithms can be implemented, a representative code for the problem must be designed. For this, a solution in the problem space is encoded in the form of a chromosome consisting of the smallest genetic component, the gene. With the theory of evolution and genetic theory, the application of genetic algorithms will involve several operators, namely selection, crossover, and mutation. The stages of the genetic algorithm are first forming an initial population, then the population is evaluated by a fitness function that has been determined, then the population is processed (recombined) with using genetic operators such as selection, crossover, and mutation so as to produce a new population for the next generation. The process or stages are repeated until they reach certain stop criteria. The stop criteria can be a generation limit or a certain optimal value desired. Standard flow chart form of genetic algorithm according to Goldberg (DE);

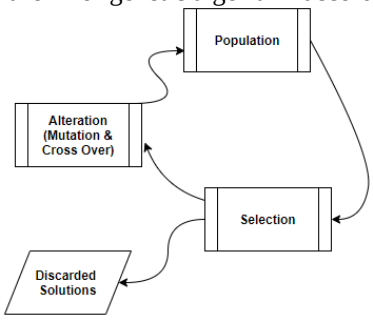


Figure 1: Genetic Algorithm

The standard form of genetic algorithm in Figure 1; can also be written in pseudo code as follows:
Genetic Algorithm Procedure; begin
t : = 0;
Initialization P (t); P (t) evaluation;
do while (Conditions not met)
t : = t + 1;
Select P (t) from P (t - 1); Recombination P (t);
End do;
end;

3. PROPOSED WORK

Under the proposed scheme we are going to extract the geo-spatial data or corpus from Open Street Map using map-extractor facilitation for the specific country, state or region. Thereafter, using the latent semantic analysis we will process the data using singular value decomposition which we remove the noise and will form the un-structured layout with respective entities and attributed for feature detection like relations and indexing. Subsequently, using a genetic algorithm with respect to elements and attributes the effective and evolutionary references will be espionage which will mitigate the retro respective data with its semantic models using crossover and mutation to provide appropriate and accurate information promptly. The below figure depicts the workflow of the proposed scheme for perusal and ready reference.

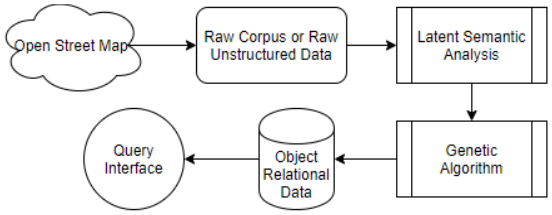


Figure 2: Workflow of Proposed Scheme using LSA and GA

***Latent Semantic Analysis*:** LSA is a method based on machine learning in excess of natural language processing, in exacting distributional semantics, of analyzing relationships between a set of credentials/data/corpus/documents and the terms they enclose by fabricating a set of conception related to the data/corpus/documents and conditions. LSA

presume that expressions that are close in connotation will transpire incomparable pieces of text (the distributional hypothesis). However, Latent semantic analysis aims to highlight hidden semantic relations among terms and enables protuberance of uncertainty and credentials in the same space defined with semantic dimensions. The preliminary set of documents is accessible in a form of matrix A where one column represents one document. Rows of the matrix are terms that occur in documents. Field Ai,j of the matrix is the occasion of the term i in the document j. If we imagine every row as one dimension of semantic space then every column is a vector that projects the corresponding document in such space. Matrix A, composed in the described way, has a very large number of rows/dimensions (n). To make this space easier to handle, the reduction of dimensions is necessary. Reduced space is called latent space because hidden (latent) knowledge of co-occurrence of terms is reviled. One of the techniques for dimension reduction is the Singular Value Decomposition (SVD). This decomposition reduces the space in such a way that the difference between projection in original and latent space is minimal. When SVD is applied to matrix A we get three new matrices: A = USVT where U and VT are orthogonal matrices and S is the diagonal matrix composed of singular values of A matrix. By restricting matrices U, V and S to their first k rows and columns, one can create a projection of matrix A in k-dimensional space (k << n). Such projection has minimum deviation because singular values in the S matrix are ordered in descending order and they are considered to be weights for the relevance of a particular dimension. If one takes the same approach for representing the queries as semantic vectors, then those queries can be projected in the same semantic space and compared to the vector of particular document. Angle between vectors is usually taken as quantify of relationship. The below diagram depicts the architecture of latent semantic analysis.

LSA Pseudo Code: In the analysis process involves the formation of vectors using XML documents thereafter vector training data documents the process of which is shown as under in pseudo code.

Query Vector Training Algorithm

1: For i = 1 to p do
2: a) Remove Noise from the Corpus (XML Documents)
3: b) Stem words/Nodes from the training answer Corpus
4: End For
5: Select for the index-term of Vector by matrix document
6: Vectors form of matrix using XML document, Amxp
7: Decompose the Amxp matrix using SVD, where the equation will be $A_{mxp=U_{mxr}*S_{rxr}*V_{rxp}.^T}$
8: Truncate / cut / reduce from U, S and VT and make the following equation: $A_{kxk=U_{mxk}*S_{kxk}*V_{kxp}}$
9: For j = 1 to p Do
10: Form key vector well formed XML Documents as $D_j = D_j^T * U_{mxk} * S_{kxk}.^{-1}$
11: End For

The "next step is the same as forming a query vector training, the reduced SVD matrix will used to form the query document answer vector students, In this process each XML document node will later formed query vectors that will be compared with queries vector training data answers so that values can be determined the similarity which is the basis for giving value automatically by the system. This process can be demonstrated by equation below along with pseudo code."

$$Q_j = Q_j^T * U_k * S_k{}^{-1} (eq. 6)$$

Information:
Qj: Vector query documents student answers
Q_j^T: Transpose vector query to students' answer documents
U_k: Orthogonal reduction matrix
$S_k{}^{-1}$: Singular reduction inverse matrix

Algorithm for the Formation of XML Document Vector Queries

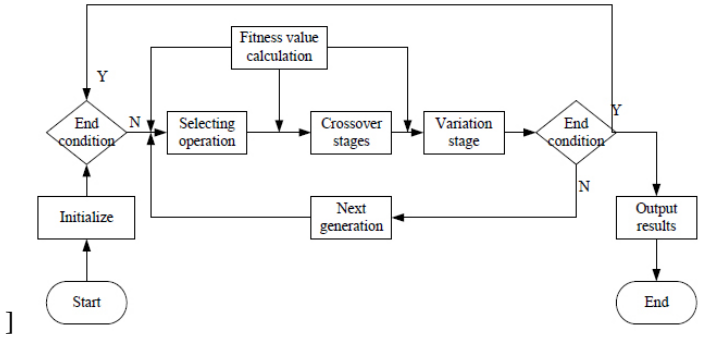
01: For i = 1 to p do
02: a) Remove Redudant Nodes from essay XML Document D'
03: b) Stem words from Nodes and Attributes
04: EndFor
05: The same matrix query dimensions are form rules using XQuery from matrix documents

06: For j = 1 to p Do

07: Form the answer vector as $Q'_j = Q'^T_j * U_{mxk} * S_{ksk}{}^{-1}$

EndFor

Genetic Algorithm: GA "was primarily initiated by John Holland in the 1970s (Holland 1975) as a result of examinations into the opportunity of computer programs that will undergo evolutionally in the Darwinian sense. GA is a component of a broader soft computing archetype acknowledged as evolutionary computation. They endeavor to disembark at the finest elucidation during a progression similar to biological evolution. This engages following the principles of survival of the fittest and crossbreeding and mutation to produce better solutions from a pool of obtainable solutions. Genetic algorithms have been seen as equipped for discovering answers for a wide assortment of issues for which no satisfactory algorithmic arrangements exist. The GA strategy is especially appropriate for enhancement and optimization, a critical thinking method wherein at least one generally excellent arrangement are looked for in an answer space comprising of an enormous number of potential arrangements or solutions. GA lessens or reduces the search space by consistently assessing the current generation of candidate solutions, disposing of the ones positioned as poor, and creating another age through crossbreeding and mutating those positioned as great. The positioning of up-and-comer arrangements is finished utilizing some pre-decided proportion of fitness or wellness. The GA evolutionary cycle begins with a randomly selected initial population. The progressions to the population happen through the procedures of choice dependent on fitness, and adjustment utilizing crossover and mutation. The use of choice and adjustment prompts a populace population a higher extent of better arrangements. The evolutionary cycle proceeds until a satisfactory arrangement is found in the present age of populace, or some control parameter, for example, the quantity of ages is surpassed or exceeded the below diagram depicts the genetic algorithm evolutionary cycle."



]

Figure 3 Work Flow Model of Genetic Algorithm

The "negligible component of a genetic algorithm is known as a *gene*, which symbolize a element of information in the predicament sphere. A sequence of genes, recognized as a *chromosome*, signify one potential solution to the problem. every gene in the chromosome correspond to one element of the solution prototype."

1. *Selection***: "**In genetic progression, only the fittest survive and their gene pool contributes to the formation of the next generation. Selection in GA is also pedestal on a comparable procedure. In a familiar form of selection, known as *fitness proportional selection*, each chromosome's likelihood of being preferred as a excellent one is comparative to its fitness value."
2. *Alteration to improve good solutions***: "**The amendment steps in the genetic algorithm filter the good solution from the current generation to produce the next generation of candidate solutions. It is carried out by performing crossover and mutation."
3. *Crossover***:** "might be considering as artificial mating in which chromosomes from two individuals are combined to create the chromosome for the subsequent production. This is done by splicing two chromosomes from two different solutions at a crossover point and swapping the spliced parts. The suggestion is that some genes with superior distinctiveness from one chromosome may as a result combine with some good genes in the other chromosome to create a better solution represented by the new chromosome."
4. *Mutation* "is a unsystematic modification in the genetic constitution. It is useful for establish new distinctiveness in a population – impressively not accomplish all the way through crossover unaccompanied. Crossover only rearranges accessible characteristics to give new combinations. For

example, if the first bit in every chromosome of a generation happens to be a 1, any new chromosome created through crossover will also have 1 as the first bit. The mutation operators revolutionize the present value of a gene to a different one. For bit string chromosome this change amounts to flipping a 0 bit to a 1 or vice versa. Although useful for introducing new traits in the solution pool, mutations can be counterproductive and practical only occasionally and randomly."

Pseudo Code is depicted for ready reference:-

Step1*: Generate an initial Random Population*
While iteration <= maxiteration
Iteration = iteration + 1
Step2*: Calculate the Fitness of each Individual*
Select the Individual according to its Fitness
Step3*: Perform Crossover with probability pc*
Step4*: Perform mutation with probability pm*
Step5*: Population = selected individual after crossover and mutation*
End while

Algorithmic code as under:-

1. *Set t := 0;*
2. *Initialize P(t) :=* $\{S_{1\ \ldots\ldots\ S_N}\}, S_i \in \{0,1\}^n$;
3. $evaluateP(t): \{f(S_1)\}, u(S_1)\}, \ldots., \{f(S_N), u(S_N)\}$;
4. $findmin_{S\in P(t),u(S)=0}\{f(S)\} \vee min_{S\in P(t),u(S)>0}\{u(S)\}, setS \leftarrow S$;
5. $while(t < t_{max}) \neq truedo$
6. *Select* $\{P_1,P_2\} := \Phi\big(P(t)\big) \dot{-} \Phi$ *=matching selection method */*
7. *Crossover C: =* $\Omega_f(P_1, P_2) \dot{-} \Omega$ *=uniform crossover operator */*
8. *Mutate* $C \leftarrow \Omega_m(C, m_s, m_\alpha \epsilon) \dot{-} \Omega_m = static \wedge adaptivemutation *$
9. $C \leftarrow \Omega_{improve}(C) \dot{-} \Omega_{improve}$ *=heuristic improvement Operator */*
10. ***If*** $C \equiv anyS_i \in P(t) \frac{then}{} Cisredudant *$
11. *Discard C and go to 6;*
12. ***End if***
13. *Evaluate f(C), U(C);*
14. ***Find*** $aS' \in P(t) basedontheranking$
$replacementmethod \wedge replaceS' \leftarrow C$;
15. ***If*** *(u(C) =u(S*) = 0 and*
f(C) < f(S)) or (u(C) > 0, u(S*) > 0 and u(C) < u(S*)) then*
16. $S^* \leftarrow C$;
17. ***End if***
18. $t \leftarrow t + 1$
19. ***end while***
20. ***return S*, f(S*) and (S*)***

4. RESULT AND SIMULATION

As per the discussion in Section 1 and Section 3 this chapter will emphasis on implementation scenarios along with results and values respectively the modules will be data gathering from openstreetmap.org in form OSM format i.e. XML Compressed format therefore the data had to be uncompressed first, thereafter data cleansing or preprocessing forming the data in well formed manner, subsequently imposing Latent Semantic Analysis for vector and relation formation along with Singular Value Decomposition, thereinafter data will be evaluated using Genetic Algorithm for Optimization of relations and nodes. Consequently, data will be bundled in Object Relational Database System for accessing data or information with accuracy and precision.

Deriving Generic Information Using Latent Semantic Analysis: below code block depicts the call of libraries used using python like xml.etree for XML iteration, print for standard output, re for regular expression, nltk for natural language processing, math for mathematical formulations, numpy for scientific computing and scipy for machine

learning models therefore using the below code snippet the scheme will iterate from OSM file using LSA and extract the information like tags, ways and other important information."

```
import xml.etree.cElementTree as ET
import pprint
import re
import nltk
import math
import numpy
from nltk.corpus import stopwords
from nltk.stem.wordnet import WordNetLemmatizer
import scipy.io
from scipy import linalg
OSMFILE = "new_delhi.osm"
def matrix_reduce_sigma(matrix, dimensions=1):
    """This calculates the SVD of the matrix, reduces it and
        creates a reduced matrix.

        @params matrix the matrix to reduce
        @params dimensions dimensions to reduce.

        @return matrix The reduced matrix
    """
    uu, sigma, vt = linalg.svd(matrix)
    rows = sigma.shape[0]
    cols = sigma.shape[1]

    #delete n-k smallest singular values
    #delete ie settings to zero
    smallerBound = min(rows, cols)
    for index in xrange(smallerBound - dimensions, rows):
        sigma[index] = 0
```

```
root@localhost:/OpenStreetMap
File  Edit  View  Search  Terminal  Help
[root@localhost OpenStreetMap]# python3 1_LSA.py
{'bounds': 1,
 'member': 437,
 'nd': 87971,
 'node': 66333,
 'osm': 1,
 'relation': 55,
 'tag': 23035,
 'way': 15704}
```

Figure 4 Result Derived using LSA depicts the Information of Nodes, Relation, Tag and Ways in OSM file.

Optimization Of Corpus (OSM) using Genetic Algorithm: Under the scheme, genetic algorithms begin by initializing a set of randomly generated solutions. This set of solutions is called population. Each individual in the population is called a chromosome which describes a solution of the problem to be solved. A chromosome can be expressed in a symbol string, for example a collection of nodes. Chromosomes can change continuously called regeneration. For each production, chromosomes are calculated using a quantified instrument called the fitness function (level of fitness). To make the next generation, new chromosomes are called node formed by merged with two chromosomes from the current generation using operator's crossover / crossing or changing a chromosome by using a mutation operator. A new generation is formed by means of a selection made against parents and node based on the fitness value and eliminating the others. The more suitable chromosomes have a probability of being chosen. After several generations, this algorithm will converge towards the shape of the chromosome best, hoping to get it states the optimal solution of the problem being solved.

- *General Structure of Genetic Algorithms*

If P (t) and C (t) are the parent and node of t generation, the general structure of the genetic algorithm is as follows:
Genetic algorithm procedure:
begin

t ? 0;
initialization P (t); P (t)
evaluation;
while (termination conditions not met) do recombination P (t) to produce
children
C (t); C (t)
evaluation;
selection P (t + 1) from P (t) and C (t); t ? t + 1;
end

- *Operator and Evaluation Function*

Usually, initialization is assumed randomly. Recombination involves crossover and mutations to produce node. In fact, there are only two types of operations in genetic algorithms, namely genetic operations (crossover / crosses and mutations) and evolutionary operations (selection). In the theory of evolution, this mutation is a chromosome operator that allows living things to adapt to their environment even though the new environment is incompatible with the original parent environment. The biggest factor in the theory of evolution that causes a chromosome to survive, extinct, make a cross or mutation is the environment. In genetic algorithms, environmental factors are played by the evaluation function. The evaluation function uses chromosomes as input and produces certain numbers that indicate the performance of the problem being solved. In the optimization problem, the evaluation function is the objective function (objective function). The value of the evaluation function is called the suitability value (fitness value). This value will determine whether a string will appear in the next generation or die. Below is the simulation depicts the above scenario using python code blocks with results.

- *Selection*

The selection will determine which individuals will be selected for recombination and how node formed from selected individuals. The first step done in this selection is the search for fitness values. There are several selection methods, including: Roulette Wheel Selection : The roulette wheel selection method is the simplest method, and is often also known by name stochastic sampling with replacement. As the name implies, this method mimics the roulette-wheel game in which each chromosome occupies a circle piece on the roulette wheel proportionally according to its fitness value. Chromosomes that have a greater fitness value occupy a larger circle than a chromosome with a low fitness value. Table 1 illustrates an example of using the roulette wheel method.

Fitness Value	Chromosomes	Probability
K1	1	0.25
K2	2	0.5
K3	0.5	0.125
K4	0.5	0.125
Amount	4	

Table 1 Example of using the Roulette Wheel Selection Method.

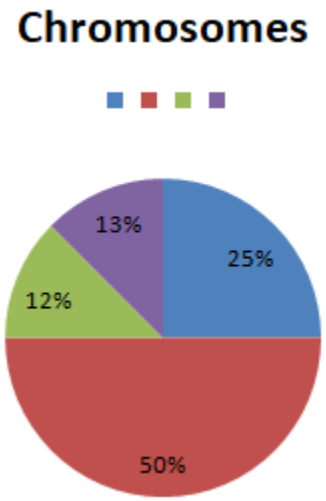


Figure5 Graph Representation of above Table Using the Roulette Wheel Selection Method;

- *Crossover*

One of the most important components in genetic algorithms is crossing or crossover. Cross or crossover function combines two different parent strings into two different descendant strings its parent; A chromosome that leads to a good solution can be obtained from the crossing of two chromosomes. Crosses can also be bad if the population size is very small. In a very small population, a chromosome with genes that lead to a solution will spread very quickly to other chromosomes. To overcome this problem with a certain probability That is, the crossing can be done only if a random number [0,1) is raised less than contextual range which is determined. From the results of studies that have been done by genetic algorithm practitioners, it is suggested that the probability of crosses is high enough, which is 80% to 95% to give good results. For some specific problems the 60% crossing probability gives better results.

```
id,user,uid,version,changeset,timestamp
28700996,PlaneMad,1306,7,31340004,2015-05-21T11:11:03Z
43242766,sowjanyaaa,2901480,6,39548603,2016-05-25T05:05:18Z
50283495,Oberaffe,56597,2,26277281,2014-10-23T12:56:56Z
92881405,PlaneMad,1306,8,31482143,2015-05-26T18:09:52Z
158316138,PlaneMad,1306,1,11210173,2012-04-07T07:25:13Z
175663885,NoelB,236361,2,12710718,2012-08-13T08:37:28Z
234852488,bdiscoe,402624,1,17479328,2013-08-24T04:57:53Z
242854404,PlaneMad,1306,3,31338475,2015-05-21T09:48:31Z
268635858,keepright! ler,1731253,1,21287952,2014-03-24T16:09:25Z
300602926,Petr Dlouhý,17615,1,25067584,2014-08-28T07:29:56Z
300603028,Petr Dlouhý,17615,1,25067584,2014-08-28T07:30:03Z
303424013,Rajesh Diwakar,2329016,3,25471472,2014-09-16T05:59:06Z
303576776,pratikyadav,2905914,5,32004690,2015-06-16T11:50:55Z
303581159,pratikyadav,2905914,8,31999608,2015-06-16T07:27:21Z
303656308,Oberaffe,56597,4,25762701,2014-09-30T08:42:54Z
304486525,sowjanyaaa,2901480,4,39468678,2016-05-21T12:07:00Z
304922251,stjohnscollege,3308183,4,34646322,2015-10-15T05:25:50Z
305104828,sowjanyaaa,2901480,5,39618171,2016-05-28T05:56:10Z
305709106,veekesh yadav,2354635,1,25767537,2014-09-30T13:21:21Z
310304726,Warin61,1830192,1,26425921,2014-10-29T23:43:59Z
311885110,Warin61,1830192,2,26680855,2014-11-10T06:09:44Z
312538019,Warin61,1830192,1,26748941,2014-11-13T05:22:56Z
313016464,Warin61,1830192,1,26812105,2014-11-15T23:51:56Z
345944555,sowjanyaaa,2901480,2,39683784,2016-05-31T09:03:22Z
345946915,sowjanyaaa,2901480,2,39683784,2016-05-31T09:03:22Z
345954171,Chetan_Gowda,2644101,1,31252512,2015-05-18T11:41:43Z
345955618,PlaneMad,1306,1,31252580,2015-05-18T11:44:38Z
```

Figure 6: Ways Extracted from Corpus using Crossover (GA)

- *Mutation*

After the crossing process is complete, then a mutation process is imposed on the corpus. Mutation is the process of changing the value of one or several genes in 1 chromosome. Mutations function in making changes that are not caused by crossing. If the process of selecting chromosomes tends to continue on good chromosomes, it is very easy for early convergence to occur, which is to reach the optimum local solution. To avoid early convergence and to maintain differences in chromosomes in the population, in addition to taking a more efficient selective approach,

mutation operations can also be carried out. This mutation process is random, so it does not always guarantee that after the mutation process a better fitness chromosome will be obtained, but in the presence of this mutation it is hoped that the chromosomes obtained will have better fitness than before the mutation surgery. Figure 2.9 is an example of the mutation process.

Chromosomes before mutations	1 1 0 **0** 1 **1**
Chromosomes after mutations	1 1 0 **1** 0 **0**

Table 2: Example of the mutation process

However, "mutation has a controversy in its application in genetic algorithms because of its random nature so that it can interfere with chromosomes with the best fitness that has been obtained. Sometimes mutation is still used with a very small probability that is Pm < 1. So the possibility of chromosomes undergoing changes due to mutations is very small. However, the code block below depicts the scenarios as per the scheme."

```
Number of nodes:  664
Number of ways:  157
Number of unique users:  87
Top contributing users:  [(u'sowjanyaaa', 240), (u'PlaneMad', 66), (u'pratikyadav', 64), (u'Chetan_Gowda', 41),
(u'srividya_c', 40), (u'Rub21', 35), (u'shravan91', 30), (u'ruthmaben', 26), (u'Warin61', 25), (u'NoelB', 21)]
Number of users contributing once:  42
Common ammenities:  (u'atm', 1)
Biggest religion:  Hindu, Muslim
Popular cuisines:  choti wala, Haldiram, Bikaner Sweets, Hazaratganj Resturants
================================================================================
Accuracy  Achived : 78.86%
```

Figure 7: Results Achieved using LSA and GA with Accuracy of 78.86%

5 CONCLUSION AND FUTURE SCOPE

- *Conclusion*

From the results using Latent Semantic Analysis and Genetic Algorithm research on Open Street Map based Corpus, There are several conclusions that can be drawn, namely:

1. Latent Semantic Model is implemented to remove noise from huge corpus and bind the relation among the nodes to form the ways and direction for espionage the best data model.
2. Genetic algorithms designed and implemented generally provide solutions that are near optimal. After testing and obtaining the results, it can be said that the Genetic Algorithm works well to get the optimal solution (minimizing completion time) with accuracy of 78.86%.

2 After conducting the experiment, it was suspected to solve and to provide the accurate information with the probability of the right crossover and mutation using Object Relation Model (Schema Based Model) from the Corpus. This result is not absolute because the solution using Genetic Algorithm in principle uses the rules of random selection. Therefore, it may vary time to time as per the desired information to be retrieved from the corpus or Open Street map file.

* *Future Scope*

From the results of the study Of Latent Semantic Analysis and Genetic Algorithm on OSM based corpus, there are several suggestions that can be taken, namely:

1. The comparison or relation can be formed using cosine similarity index along with Singular Value Decomposition which can speed up the process.
2. For further research, chromosome representation can be used in other forms, such as Employment based illustration, Preference-list based illustration, Job-pair-relation-based illustration, Priority-rule-based illustration, Disjunctive graph based illustration, Completion time based illustration, Machine based illustration or Random key illustration for constant accuracy.
3. For further research, a cross operator / crossover other methods, for example job-based order crossover, partial mapped crossover, or other and mutation operations can also be done by other methods, for example inversion, insertion, or reciprocal exchange mutation.
4. Techniques like Adaptive Boost can be inculcated for swift and quick results.

REFERENCES

[1] Michael P. Bishop, Brennan W. Young, Da Huo, Zhaohui Chi, Spatial Analysis and Modeling in Geomorphology, Reference Module in Earth Systems and Environmental Sciences, Elsevier, 2020, ISBN 9780124095489, https://doi.org/10.1016/B978-0-12-409548-9.12429-7.
[2] Bankston Cotton, Environmental Psychology: Principles and Practices, Scientific e-Resources, Mar 4, 2019, ISBN 9781839474088 https://learn.arcgis.com/en/arcgis-book
[3] Kathryn Keranen,Instructional Guide for The ArcGIS Imagery Book Lyn Malone, Esri Press, 380 New York Street, Redlands, California 92373-8100 Copyright © 2017,Esri ,https://downloads.esri.com/LearnArcGIS/pdf/instructional-guide-for-the-arcgis-imagery-book.pdf
[4] Murayama, Yuji, Thapa, Rajesh Bahadur, Spatial Analysis and Modeling in Geographical Transformation Process, https://www.springer.com/gp/book/9789400706705
[5] Thapa, Rajesh & Murayama, Yuji. (2011). Spatial Analysis and Modeling in Geographical Transformation Process: GIS-based Applications. 10.1007/978-94-007-0671-2.
[6] Andreano, Maria & Benedetti, Roberto & Piersimoni, Federica. (2019). A Distance Correlation Index of Spatial Dependence for Compositional Data. Papers in Regional Science. 10.1111/pirs.12451.
[7] Congdon, Peter. (2019). Representing Spatial Dependence. 10.1201/9780429113352-6.
[8] Şen, Zekai. (2016). Spatial Dependence Measures. 10.1007/978-3-319-41758-5_5.
[9] Manley D. (2014) Scale, Aggregation, and the Modifiable Areal Unit Problem. In: Fischer M., Nijkamp P. (eds) Handbook of Regional Science. Springer, Berlin, Heidelberg
[10] Degbelo, A., Kuhn, W. Spatial and temporal resolution of geographic information: an observation-based theory. Open geospatial data, softw. stand. 3, 12 (2018)
[11] Ndehedehe, Christopher & A, Ekpa & O, Okwuashi & Simeon, Ogunlade. (2013). UNDERSTANDING ERRORS AND THEIR MEASUREMENT IN GEOINFORMATION. Journal of Environmental Sciences and Resources Managemen. Volume 5. Pp. 74 - 87.
[12] Jokar Arsanjani, Jamal & Zipf, Alexander & Mooney, Peter & Helbich, Marco. (2015). An Introduction to OpenStreetMap in Geographic Information Science: Experiences, Research, and Applications. 10.1007/978-3-319-14280-7_1.
[13] Mooney, Peter & Minghini, Marco. (2017). A review of OpenStreetMap data. 10.5334/bbf.c.
[14] Mocnik, F., Mobasheri, A. & Zipf, A. Open source data mining infrastructure for exploring and analysing OpenStreetMap. Open geospatial data, softw. stand. 3, 7 (2018). https://doi.org/10.1186/s40965-018-0047-6
[15] S. S. Sehra, J. Singh and H. S. Rai, "A Systematic Study of OpenStreetMap Data Quality Assessment," 2014 11th International Conference on Information Technology: New Generations, Las Vegas, NV, 2014, pp. 377-381, doi: 10.1109/ITNG.2014.115.
[16] https://en.wikipedia.org/wiki/OpenStreetMap
[17] Jokar Arsanjani, Jamal & Mooney, Peter & Zipf, Alexander & Helbich, Marco. (2015). An introduction to OpenStreetMap in GIScience: Experiences, Research, Applications.
[18] Zhang L, Pfoser D (2019) Using OpenStreetMap point-of-interest data to model urban change—A feasibility study. PLoS ONE 14(2): e0212606. https://doi.org/10.1371/journal.pone.0212606
[19] Mooney, P and Minghini, M. 2017. A Review of OpenStreetMap Data. In: Foody, G,See, L, Fritz, S, Mooney, P, Olteanu-Raimond, A-M, Fonte, C C and Antoniou, V.(eds.) Mapping and the Citizen Sensor. Pp. 37–59. London: Ubiquity Press. DOI: https://doi.org/10.5334/bbf.c. License: CC-BY 4.0
[20] Seto, T.; Kanasugi, H.; Nishimura, Y. Quality Verification of Volunteered Geographic Information Using OSM Notes Data in a Global Context. ISPRS Int. J. Geo-Inf. 2020, 9, 372.
[21] Jonathan Bright, Stefano De Sabbata, Sumin Lee, Bharath Ganesh, David K. Humphreys, OpenStreetMap data for alcohol research: Reliability assessment and quality indicators, Health & Place, Volume 50, 2018, Pages 130-136, ISSN 1353-8292, https://doi.org/10.1016/j.healthplace.2018.01.009.
[22] Nurin Swasti, Taufik Hery Purwanto, Application of OpenStreetMap (OSM) to Support the Mapping Village in Indonesia, IOP Conference Series: Earth and Environmental Sciencem, Published 1 November 2016
[23] https://www.openstreetmap.org/help
[24] https://wiki.openstreetmap.org/wiki/Beginners%27_guide
[25] P. Kherwa and P. Bansal, "Latent Semantic Analysis: An Approach to Understand Semantic of Text," 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), Mysore, 2017, pp. 870-874, doi: 10.1109/CTCEEC.2017.8455018.

[26] https://en.wikipedia.org/wiki/Latent_semantic_analysis
[27] Baker, Kirk. (2013). Singular Value Decomposition Tutorial. 2005.
[28] YongchangWang and L. Zhu, "Research and implementation of SVD in machine learning," 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), Wuhan, 2017, pp. 471-475, doi: 10.1109/ICIS.2017.7960038.
[29] Stewart, Sepideh & Thomas, Michael. (2006). Student thinking about eigenvalues and eigenvectors: Formal, symbolic and embodied notions. 487-495.
[30] Lee, Shyi-Long & Yeh, Yeong-nan. (1993). On Eigenvalues and Eigenvectors of Graphs. Journal of Mathematical Chemistry. 12. 121-135. 10.1007/BF01164630.
[31] Gene H. Golub, Henk A. van der Vorst, Eigenvalue computation in the 20th century, Journal of Computational and Applied Mathematics, Volume 123, Issues 1–2, 2000, Pages 35-65, ISSN 0377-0427,
[32] Vasuki, A. (2020). Genetic Algorithm. 10.1201/9780429289071-4.
[33] John H. Holland, Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence, https://ieeexplore.ieee.org/book/6267401
[34] Kumar, Kaushik & Zindani, Divya & Davim, J.. (2019). Genetic Algorithm. 10.1201/9781351049580-2.
[35] Chambers, Lance & Taylor, Michael. (2019). Genetic Algorithms. 10.4324/9780429437625-7.
[36] Judson, Richard. (2008). Genetic algorithms Genetic Algorithms. 10.1007/978-0-387-74759-0_218.
[37] Patil, Lalit. (2019). Lecture 10 : Genetic Algorithms.
[38] Alam, Tanweer & Dixit, Amit & Benaida, Mohamed. (2020). Genetic Algorithm: Reviews, Implementations, and Applications. 10.20944/preprints202006.0028.v1.
[39] Kumar, Sandeep & Sharma, Harish & Jain, Er. (2018). Genetic Algorithms. 10.1201/9780429445927-2.
[40] Awange, Joseph & Palancz, Bela & Lewis, Robert & Volgyesi, Lajos. (2018). Genetic Algorithms. 10.1007/978-3-319-67371-4_5.