A Proficient Data Repertory Acumen layout to Supervise Data Provenance in Geo science Application

K.S. Kannan, M. Hemalatha

Abstract— **Geospatial Acumen layout elucidates the storage of geosciences tracing data as well as administering data chronicle. Inception of the data is provided by data provenance. Data provenance untangles the contingency among events and source and set up the data product. Geospatial Acumen layout interpolates workflow and fine grained data chronicle, while the former unfolds information about the reproducibility of the data product and the latter about the creation of output from its input value. Classification and clustering techniques are applied to deduce fine-grained data. Data preprocessing appends data LIRT (leaning, integration, reduction and transformation). The above said process is exercised to geosciences data. The denouement of Geospatial Acumen layout is fine-grained data aggregation based on workflow with minimum warehousing. The supremacy of Geospatial Acumen layout is, it is self –adaptable to any technological prototype. The user interaction of the Geospatial Acumen layout is done by a specific compiler.**

Index Terms— **Acumen layout, Geographical Information system(GIS), Fine grained provenance.**

I. INTRODUCTION

The Geospatial Platform is aimed to provide the basic service and capability. The platform used to be an Internet-based capability offering shared and convictional geospatial data, examines, and applications that have to be used by the civic and by government agencies and associates to meet their mission needs. The Geospatial Platform is a managed collection of frequent geospatial data, services, and applications donated and managed by convictional sources and crowded on a common infrastructure. It provides tools that help consumers to find and access the data and services irrespective of their sources.

Geospatial provides information that states the spot and names of features, either below or above the earth's surface. Lucidly this act as the crucial topographical particulars establish on a map. In addition to that it also includes diverse location- correlated datasets such as usage of land and Population thickness. A Geographic information system amalgamates hardware as well as software, and it imprisons, investigates, supervises and exhibits all forms of geographically related information. GIS permits the users to appraise trends, collaborate between regulations, understand the panorama and make better more informed conclusions about compound scenarios and policy. It helps the user for answering questions and also for solving problems by looking at data that used to be quickly implicit and easily shared. GIS technology can be incorporated into any endeavor information system framework.

The term dataset or data product in the context refers to the data available in many different forms like file or tables and also in elementary collections. It also means that data in a collection of closely associated tables, corresponding to exacting experiment or event. The provenance of a data product has two significant features they are the ancestral data products i.e. the place from which the data product is evolved, and the conversion process of these ancestral data products, probably through workflows, which helped to derive this data product.

Data provenance or origin means that process of constructing the particular piece of data for data product. Data provenance refers to lineage which explains the association among events and source data in constructing the data product. It is produced once the data is routed. A user can acquire this data product by questioning the place where it is traced. In argumentative analysis Data provenance or origin plays an important role for facilitating the set of digital proof performed through a post-incident search. It is broadly used for argumentative analysis as well as for scientific associations and also in legal happenings. According to database systems, data provenance gives the explanation of how the transformation of data product obtained from its input value. As per geosciences domain data provenance describes the origin history of data product from its initial value.

Workflow provenance is a notable value for scientists that are achieved from the basis of data products produced by complex reformations. With the help of workflow one can make certain of the quality of the data, fixed on its native data and supports track back the sources of errors, further it allow automated transformation of derivations to update a data, and provide attribution of data basis. Workflow provenance states the dependencies among activities. With the help of workflow provenance there will be significant decrease in storage overhead of provenance data by concluding fine-grained provenance. Fine-grained provenance states that the formation of data product and its progression from input values. Fine-grained provenance helps for tracing the

Manuscript received May 20, 2014.

K.S.KANNAN, Assistant Professor, Computer science and Engineering, NPR College of Engineering and Technology Tamilnadu, India

M.Hemalatha, pursuing in Master Of Engineering (ME) Degree from ANNA University (PU) Chennai ,TN ,India.

value of resultant data product. The facilitation of the redirect able results is obtained with the help of fine-grained provenance. Reproducibility is obtained with the help of fine-grained provenance and storage transparency can be removed by using the workflow provenance.

The layout that handles both fine-grained provenance as well as workflow provenance need to be nonspecific, storage proficient and should be flexible to any known logical schema. The layouts which satisfy these requirements need to perform close examinations of full problem domain.

II. Related Works

In [1] peng yue et al. provides overview of recent methods on data provenance in both the general information domain and geospatial domain. The results will be helpful for geoscientists for development of geospatial provenance-aware service-oriented applications in Cyberinfrastructure in understanding the validity of developing and using provenance-aware geospatial applications, assessing what operational system framework and approaches are available and applicable within their applications, and identifying the critical issues and directing future research agenda for provenance-aware applications in Geo-Cyberinfrastructure. According to cyberinfrastructure data provenance means that derivation the history of data from its data product. The main task of cyberinfrastructure is to build shared and composite geoprocessing workflows. This is then used for combining allocated data and services in geoscientific checks. This will be helpful for the scientist to obtain consistency of the data products, confirm and copy scientific results in Cyberinfrastructure based on its requirement. The factors that have to be deal with provenance based applications are provenance representation in different application, capturing which has to be performed physically, storage by using existing metadata, and query by using its interface and its language, visualization with more understanding and finally its application.

In [2] Paul Groth et al. make notes on dataset all the way through metadata which is essential for organizing and arranging data which are performed through reconstructing provenance. The concept states that instead of manual collection of dataset its much fewer to collect the models that utilize the dataset. This helps the scientists for uploading the dataset to have collected with informal explanation but there is no link for structured data with them. And other scientists can utilize it for downloading required dataset and use them for analyzing models executed in software. For reconstructing the provenance it is needed to gather the process of alteration performed on unique dataset to obtain the resultant dataset. After reconstructing the dataset it has to be promulgated into the starting dataset. Three kinds of approaches are used for reconstructing provenance they are mining approach, leveraging with network topology and perform leveraging with implementing environment. In mining approach documents are grouped with the help of cosine similarity. In the second technique the dataset are reconstructed through sharing of information that are available. And in the third approach rebuild of dataset depends on the knowledge of information.

In [3] wang chiew tan et al. articulates the usage of provenance in past, current and in future. The statement provenance is used identically with the word lineage in the record community. It also mean as source provenance or source category. Provenance means origin or source. Workflow provenance termed as the record that obtained from the last result of workflow. There will be varied amount of information recorded for workflow. Fine-grained provenance provides brief account of data that are obtained from the transformation process. According to past, data provenance has to notice the demand to execute for relation of tuples and also for the occurrence of data at various granularities. Past data provenance also deals with the why provenance and where to use the provenance. Current usage of data provenance has DB notes which handle extension, DB notes extension which use queries based on relational algebra and the express of language through the propagation of explanation. Future usage of data provenance is said to be in spider which coats the schema mapping which handles relationship between source schema and with the target schema. The main task of spider is for correcting the errors in programming language similar to that of debuggers. In future data provenance is based on SQL queries. SQL queries act as a construction blocks for database models. It helps for providing information about how the reasoning of provenance performed.

In [4] Tanu Malik et al. describe the methods of chasing and sketching allocated data provenance. Collection system simply performs collecting data provenance and delivered it to the centralized servers. But several data intensive application wants data to be maintained locally in a decentralized manner. At the time of fine-grained granularity this will provide huge collection of data. However while chasing sketching and querying of allocated data provenance results in complexity. In order to handle the complexity and to provide answer for the successive queries of allocated data provenance, provenance sketches are provided. The provenance sketches are extended as a part of SPADE system. In the chasing process three kind of categories are used in the first category the data and its metadata are in centralized server, the second category maintains the data locally but the control of metadata are still there in the centralized server, and third category provides the data and its metadata are in distributed server. In the allocated data provenance recording are performed either in intra host dependencies or in inter host dependencies. In intra host dependencies graph based on file vertices, process vertices and edges. Inter host dependencies provenance graph based on provenance vertex. The SPADE which acts as a querying tool consists of the following functions. It mentions its file vertex and obtains its lineage or it mentions its file vertex along with a threshold and obtain lineage or it mention two file vertexes and by using these two vertices provide lineage between them.

In [5] Ziheng Sun et al. describes a workflow structure for web geoprocessing. Structure is used for assembling, organizing, accumulating, and serving the information about provenance that are created during the function of workflow for Web geoprocessing. Workflow of web geoprocessing consists of three phases they are process modeling, process modeling instantiation and execution of workflow. The process modeling provides an abstract for composite process which provides control flow as well as data flow for process nodes. The process modeling instantiation states that abstract of the process directed either into a concrete workflow or into a service chain which are executable. In workflow execution value adjoined data product is generated with the help of workflow engine that executes the chaining result. The phase of process modeling is used for recording the process models of workflow provenance. The instantiation phase is merely used for selecting services. And the execution phase is used for running the instances that are obtained from services. Workflow provenance of web geoprocessing allows the users to have information about the knowledge of process model, selected service of service chain and executed data and its parameter. Geospatial provenance has three levels. The knowledge level of geospatial provenance consists of data used in geospatial process and its service type and also the modeling of complex process. The process level of geospatial has atomic and its composite process. The service level is used for linking service that are obtained from atomic and its service chains.

III. Existing System

Existing system uses database to store fine grained data storage. In all contexts, provenance can be defined at different levels of granularity. Fine grained data federation is defined at the value-level of a data product, which refers to the determination of how that data product has been created and processed starting from its input values. It helps scientists to trace the value of an output data product. It could be facilitated to have reproducible results as well. On the other hand, coarse-grained or workflow provenance is defined at the higher level of granularity. It captures association among different activities within the model at design time. Workflow provenance can achieve reproducibility in a few cases where data are collected beforehand, i.e., offline data. In cases of streaming data, workflow provenance itself cannot achieve reproducibility due to the creation of new data products and update of existing data products during the model execution.

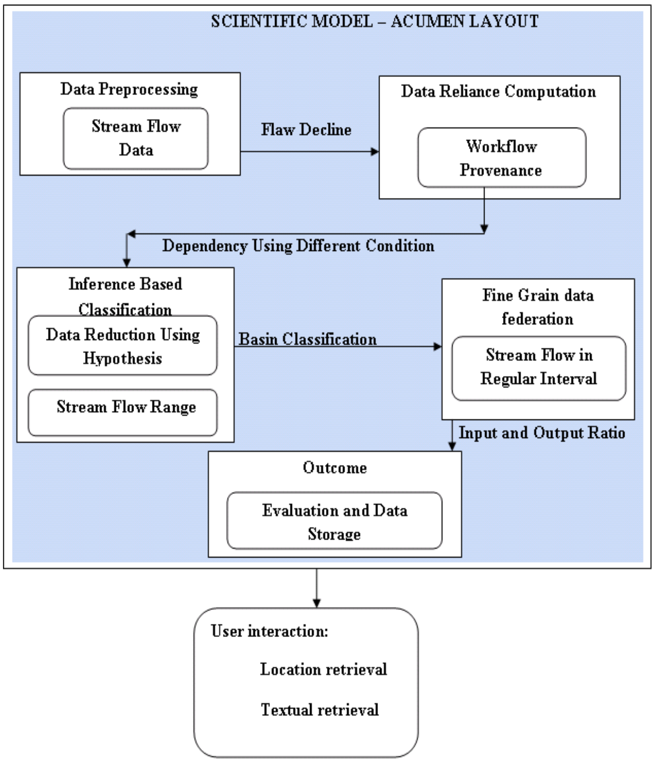
Streaming data might have different data arrival patterns. Data arriving at regular intervals are referred to as constant sampling data (e.g., temperature measurements sent at regular intervals). On the other hand, data might also arrive at an irregular interval, such as buying and selling quotes on an instrument in a stock market. These are referred to as variable sampling data.

Existing work documents fine-grained data provenance unambiguously in a database. However documenting these data require a huge amount of storage to manage fine grained data federation especially in a data streaming scenario where a single incoming data product may contribute to produce multiple output data products. Sometimes, the size of provenance data becomes a multiple of the actual data. Because these provenance data are just metadata and they are not often utilized for end users, the explicit documentation of fine-grained provenance seems to be not practical and it's too exclusive. In existing system, the size of provenance data becomes a multiple of the actual data. Existing system does not have inference technique to reduce the amount of data and storage space

IV. Acumen Layout

The Acumen scheme is needed to be developed for geoscience application. Acumen scheme consists of both workflow provenance and fine-grained provenance. Stream flow data is taken for processing. It provides an inference-based layout, which has both workflow provenance and fine-grained provenance with minimum cost in terms of duration, training, and disk utilization. Proposed framework is applicable and can be used in any kind of scientific model, and it can handle different model vibrant, such as changes at the time of process as well as arrival pattern of input data product. Estimation of the layout in original use case along with geospatial data demonstrates that proposed layout is similar and suitable for users in geoscientific stream. Data preprocessing is applied to reduce and remove the redundant and noisy data respectively. Data preprocessing helps us to remove empty sinks from the stream flow data which increase accuracy of the process. Inference techniques are applied to infer the fine-grained provenance which reduces storage space and processing time. Reliancy based classification and similarity based clustering is applied to identify the different uses of sinks such as drought basin, flood basin and constant basin. Thus, acumen scheme results in accurate geosciences trace data for scientific model.

V. Steps Of Acumen Layout


SCIENTIFIC MODEL – ACUMEN LAYOUT
Data Preprocessing
Stream Flow Data
Flaw Decline
Data Reliance Computation
Workflow Provenance
Dependency Using Different Condition
Inference Based Classification
Data Reduction Using Hypothesis
Stream Flow Range
Basin Classification
Fine Grain data federation
Stream Flow in Regular Interval
Input and Output Ratio
Outcome
Evaluation and Data Storage
User interaction:
Location retrieval
Textual retrieval

A. Flaw Decline

The flow data of stream for various river sinks and its various durations are used for tracing the data provenance from input measurements of the data to its output product. By using flaw reduction, different sinks of various durations are segmented. The consideration of flaw reduction is based on the most and little difference of the forecasting points. Preprocessing of data is applied for flaw reduction. Preprocessor is actually program that is used for processing the input data and to produce the output data product which is used as input for another program. The result is said to be a preprocessed form obtained from the input data, which is often used by some subsequent programs like compilers. The character of the preprocessor is considered for amount and kind of processing occur.

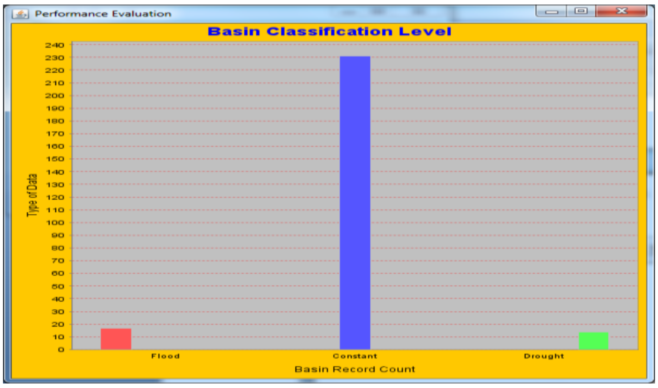
B. Data Reliance Computation

Data reliance in computer science is a state through which program statement (instruction) refers the data of a previous statement. In compiler theory, the procedure used to determine data relevancies among statements (or instructions) is called reliance analysis. Reliance occurs in a database when information stored in the same database table uniquely determines other information stored in the same table. In data Reliance computation module, basin Reliance is identified using attribute values. The fields used for data Reliance computation, KAF (Kanchan TM Arsenic Filter). KAF provide a filtration rate of 25 and 15 L/hour, respectively, sufficient enough to supply water. Minimum and maximum KAF values are considered for Reliance computation.

C. Inference Based Classifications

In inference based classification module, the stream flow data is classified into a range of sinks based on the KAF values. KAF values and stream flow percentage of average and 30 year average stream flow values are used for classification. By combining the 30 year avg. values and percentage of avg. stream flow data, a hypothesis is defined and the basin which passes the hypothesis is classified separately. Bayesian classification technique is used for classification. A Bayesian network, Bayes network, belief network, Bayesian model or probabilistic directed acyclic graphical model is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies.

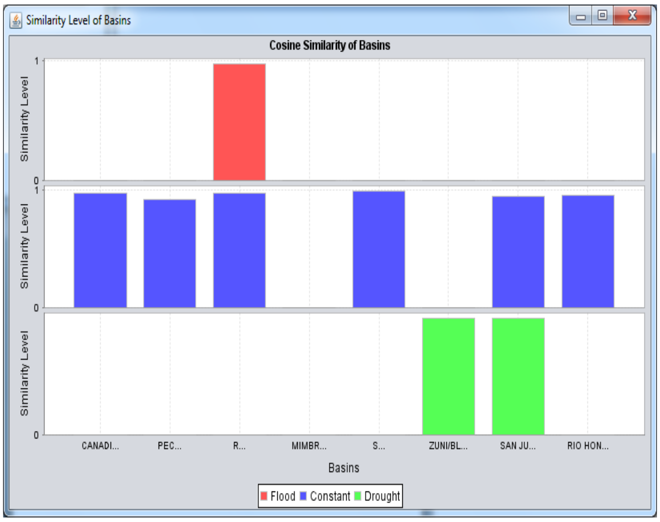
HYPOTHESIS=MAX VALUE-MIN VALUE



D. Fine-Grained Data Federation

In Fine-grained data Federation module, two types of stream flow data is considered such as Input-output ratio and Perseverance of output data product. Depends on data Reliance activities (workflow provenance), fine-grained data is mined and federated with respect to the above two stream flow data. The persistence and input-output ratio is computed for different durations and compared for Federation. Cosine similarity is used for data Federation. Cosine similarity is actually a measure of similarity stuck between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a Cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0, 1].

$$C.D \div |C|\,|D|$$



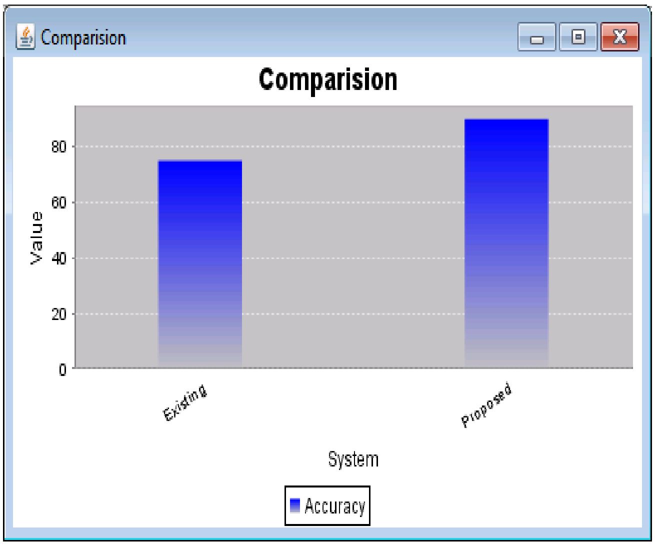
E. Data Storage and Evaluation

In data storage evaluation module, the fine-grained data is stored in database. Fine-grained data refers to the data tracing which traces the changes in input data till reaches the output data product. Thus with minimized space and time consumption, the Provenance data is stored efficiently in the database for the proof of scientific model. Evaluation includes optimized space and time consumption with accurate results shown in graph. Data storage and Evaluation detects the Flood and Drought sinks. It also provides information about the constant flow sinks.

F. User Dealing Interface

Initial step of this interface is performed based on applying forecasting points for database that obtained from previous steps. The collected information not fully adopt for our geo science application. Improvement on data cluster is performed by k-means huddling algorithm. The method randomly chooses the objects (attribute) as cluster centroid and calculates the relation between them to forms the set of clusters. And the set of clusters to calculates the relation between the each cluster individually and thus it forms the final cluster used for query processing. The algorithm states

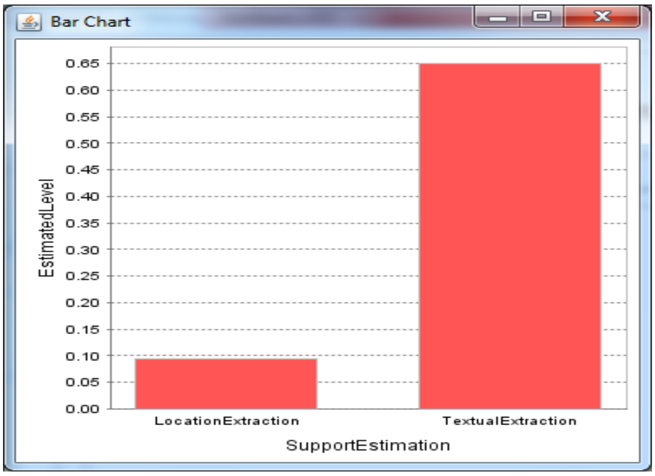
- Randomly select k data attributes from data set as initial centers. Repeat;
- Calculate the distance between each attribute di (1 <= i<=n) and all *k* huddles Cj(1 <= j<=k) and assign attribute di to the nearest huddle.
- For each huddle j (1 <= j<=k), recalculate the huddle center. Until no change in the center of huddles.



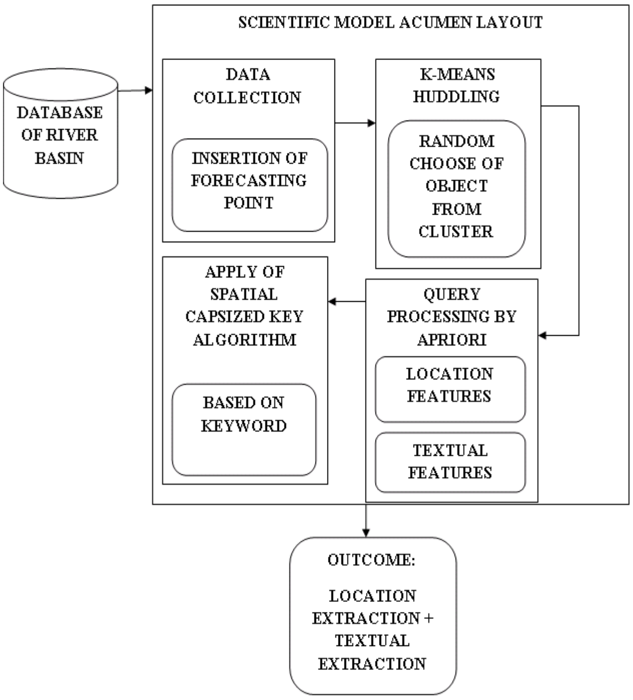
The above graph shows the comparision between bayesian and k-means huddling algorithm.

Query processing is performed by using Apriori algorithm. Apriori is used to obtain frequent item set from the database. After collecting the needs from user forms the query for data base processing. Support and Confidence of the data is used for calculating frequent item set for location as well as textual extraction.

- Support(A) = no. of transactions which contain the item set A / total no. of transactions contain the item set
- Conf(A→ B)=sup(A U B)/Sup(A)
- Freq(A→B)=sup(A,B)/Conf(A,B)



The spatial capsized key algorithm is used for providing result for the query provided by the users**.** The spatial capsized key provides effective search for user needs in textual oriented and the location oriented data. The capsized key list is the collections of points and points have the set of points and set of points have the set of key words and keywords relates the set of documents. The result for the keyword search is provided for both textual oriented as well as location oriented queries.


SCIENTIFIC MODEL ACUMEN LAYOUT
DATABASE OF RIVER BASIN
DATA COLLECTION
INSERTION OF FORECASTING POINT
K-MEANS HUDDLING
RANDOM CHOOSE OF OBJECT FROM CLUSTER
APPLY OF SPATIAL CAPSIZED KEY ALGORITHM
BASED ON KEYWORD
QUERY PROCESSING BY APRIORI
LOCATION FEATURES
TEXTUAL FEATURES
OUTCOME:
LOCATION EXTRACTION + TEXTUAL EXTRACTION

G. Simulation Result

The given input will be Stream flow data of different river sinks for different durations. By using this input value the output obtained which will be

- Constant flow
- Variable flow

Variable flow helps to identify flood detection and drought in the river basin in the respective region. This result is used as input for user interface which provides result based on the keyword inserted by the user. The user interface acts as search engine for retrieval of information about the Geospatial data.

User interface result is based on location retrieval and textual retrieval.

VI. Conclusion

Scientists understand the importance of provenance data. Acumen layout is to manage provenance data especially for geo scientific applications. Therefore the approach provided can build workflow provenance graph automatically. Since every scientific model has different characteristics, thus incorporates the self-adaptability mechanism to the framework, which can select the appropriate method to infer fine-grained data provenance based on the model parameters. Further it is to improve user interface of the layout. This user interface for user is to get the information that is stored on data base. It forms the huddling of data more efficiently. User interface helps the user to retrieve their required details more quickly and accurately.

REFERENCES

[1] Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro and Huma Javed " Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity" Middle-East Journal of Scientific Research 12 (7): 959-963, 2012, ISSN 1990-9223

[2] Jun Wang, Marlon Pierce, Yu Ma, and Geoffrey Fox, Andrea Donnellan, Jay Parker, and Margaret Glasscoe "Using Service-Based GIS to Support Earthquake Research and Disaster Response" IEEE CS and the AIP

[3] Mohammad Rezwanul Huq, Peter M. G. Apers, and Andreas Wombacher 2013 "An Inference-Based Framework to Manage Data Provenance in Geoscience Applications" IEEE Transaction on issue:99

[4] Salmin Sultana, Mohamed Shehab, Elisa Bertino 2013"Secure Provenance Transmission for Streaming Data" IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 8, August 2013

[5] Tanu Malik, Ligia Nistor, Ashish Gehani 2012"Tracking and Sketching Distributed Data Provenance" IEEE sixth international conference

[6] Wang-Chiew Tan 2009 "Provenance in Databases: Past, Current, and Future" Foundations and Trends in Databases Volume 1 Issue 4, April 2009

[7] P. Yue, Z. Sun, J. Gong, L. Di, and X. Lu in Jul. 2011 "A Provenance Framework For Web Geoprocessing Workflows" Proc. IEEE Int. Geosci. Remote Sens. Symp., Jul. 2011, pp. 3811–3814

[8] P.Yue and L.He 2009 "Geospatial Data Provenance in Cyber infrastructure" Proc. IEEE 17th Int. Conf. Geoinformat., Aug. 2009, pp. 1–4.

[9] Yufei Tao , cheng sheng "Fast Nearest neighbor search woth keyword" on IEEE Transactions On Knowledge And Data Engineering

K.S.Kannan was born in Dindigul, Tamil Nadu (TN), India, in 1979. He received the Bachelor of Technology(B.Tech.)degree from the Pondicherry University (PU), Pondicherry, TN, India, in 2003 and the Master of Engineering (M.E.) degree from the Annamalai University, Annamalai Nagar, TN, India, in 2005,He is currently pursuing the Ph.D. degree with the Department of Information & Communication Engineering in Anna University,Chennai. His research interests include GeoSpatial Database,Distributed Systems.

M.Hemalatha was born in Madurai, TamilNadu (TN),India in 1990. She received the Bachelor Of Technology B.TECH from KALASALINGAM University (PU) Krishnan Koil ,TN , India in 2012. She is currently pursuing in Master Of Engineering (ME) Degree from ANNA University (PU) Chennai ,TN ,India. Her research interests include GeoSpatial and Data Mining.