International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-11, September 2020

Geotagging: Systematic Anatomization and Conceptual Model for POI Verification

Monika Sharma, Vinod Bothale, Meenakshi Nawal, Mahesh Bundele

Abstract*: People have been contributing large amount of data every day in Geographical Information Portal however to harness the real power of this tremendous amount of data, it must be managed efficiently. There are some challenges with this data, such as authentication and skewedness. Maximum public places even in a small village must be geotagged to provide better citizen centric services specifically in developing countries. Hence, it's a need of hour to enrich the unique Indian GIS portal "Bhuvan" with Point of Interest (POI) data where one can find all necessary information. Although POI on Bhuvan is improving day by day however to overcome the challenges mentioned above is an important task. Therefore, a framework is required to expedite the tagging and authentication process of the tagged data in an efficient manner to exploit the power of POI data. In this paper, we have explored the available techniques to improve and verify the crowd sourced data and propose a conceptual model to accomplish the objective of verifying and managing the geotagged data to enrich the portal. A trust rank parameter is introduced to ensure the quality of the POI data. This will be calculated using multisource verification model using state of the art open source technologies available. This verified POI data can be used further in knowledge Graph creation to get better search facility.

***Keywords*: Authentication, Crowdsourcing, Point of Interest, Tagging.**

I. INTRODUCTION

Various proprietary GIS portals are available where people are generating a tremendous amount of data. Mostly this data is not covering all the regions uniformly. Various studies show that people use to tag famous places a lot, and same is with the urban data while there is a lack of information available for rural areas. However, each place has its own significance and to enrich the geotagged data related to rural areas, will also help in producing citizen centric services in these areas. This can be visualized with the help of a scenario where a road accident has happened nearby any village, then the emergency hospital and nearest police station can be located using geotagged data of that particular area quickly for people who are not familiar with the place. Hence, it is vital to get maximum public places to be geotagged with reliable information. It is an important to find a way to collect large amount of data quickly. Crowdsourcing is the easiest method to enrich the Point of Interest data on any GIS portal. Though there are some challenges with crowdsourced data like aggregation, verification and management of this data. Many people can tag the same place with different names like a worship place, some people may tag it with a temple, others can tag with the regional name like mandir but both are pointing to the same place. Hence to aggregate the data and removing the duplicates without losing the additional keyword is also necessary step.

Along with this, verification of the crowd sourced data is an important task because further this data will be used to provide better public services. To ensure the authenticity of the crowdsourced data is an important and challenging task. As of now, validating the crowd sourced data is a manual task in Bhuvan portal which we have chosen as the case study. Since manually verifying the large amount of data is time-consuming process and not feasible hence it is important to focus on this issue to expedite the task of POI data collection with quality assurance. Geotagged data in context would be huge data e.g. Plain xml data size in an open street map is 1256.3 GB on 1st July, 2020 as per Wikipedia information [47]. Hence, to manage this data is an intriguing job. In this study, we have explored various research work done in all these challenging areas and studied available methods and techniques at each stage to find the way to overcome the stimulating situations.

We explored "Bhuvan" GIS portal to understand the significance of our work with respect to various Indian scenarios. We have observed that people have contributed large amount of data on Bhuvan portal but the verification task is still needed and geotagged data is not synchronized therefore, searching a place may not give correct results. Hence, we decided to develop a framework which can help to overcome above mentioned challenges i.e. data verification, aggregation and management to harness its true power. Our proposed system provides a framework to collect, verify, aggregate and managing the POI data using crowd sourcing with the help of progressive technologies available. This data will be further used to generate the knowledge graph to apply data discovery methods in order to empower the better use in public utility applications.

Revised Manuscript Received on September 30, 2020.
***** Correspondence Author
Monika Sharma*, Computer Engineering, Poornima University, Jaipur, India. Email: smonika15@gmail.com
Vinod Bothale, National Remote Sensing Center, ISRO, Hyderabad, India. Email: vinod_bothale@nrsc.gov.in
Meenakshi Nawal, Computer Engineering, Poornima University, Jaipur, India. Email: meenakshi.nawal@poornima.edu.in
Mahesh Bundele, Computer Engineering, Poornima University, Jaipur, India. Email: maheshbundele@poornima.edu.in

Our proposed solution for POI verification is a novel idea as per our best knowledge because we are using Lay out map's meta data as the ground truth data since it is an authentic document provided by Indian government. However, we cannot rely on this data solely as it may change over the period of time.

Hence, we are applying multiple factors while assessing the quality of POI data and each factor has its own weightage which cannot be ignored. Although it will increase the computation cost but we are using Hadoop map reduce framework and dividing the verification task into multiple sub task which can be executed simultaneously. So, the performance will be improved. Outcome of the verification task is the value of trust rank parameter which will be associated with each POI tag data. This research paper is organized in four sections. Related work done is represented in section two. In section three, the proposed model is explained and the last section proposes the conclusion and future plans to achieve the objectives laid in the proposed model.

II. RELATED WORK

As mentioned above, the proposed system must be holistic which can overcome the challenges cropped up at each stage. Hence, the existing methods, techniques and approaches related to each stage was explored thoroughly and presented in following subsections with headings "crowd sourcing platform", "aggregation", "evaluation of authenticity" and "knowledge base Creation". More than seventy research papers were explored and categorized as per the stages involved like collecting the data which will be referred in the paper as crowd sourcing platform, aggregating collected data, verification of the data and managing the data. Summary of the research work explored is represented in the form of a table which will be helpful to get a quick glance of methods, approaches or techniques used along with the used data set and important observation of the research work. We have eliminated the review papers, or the incomplete information papers from the respective tables. However important takeaway messages from such papers were included in the discussion at respective stages. During this study we observed that mostly verification algorithms were designed for quantitative data rather than qualitative data. Generic benchmark frameworks were not tested on real and large data set. Validation techniques were explored in detail and given in the subsection under subheading "Evaluation of Authenticity" of section II. There is no open source framework available which can be directly applied for mapping, verification and knowledge base creation for open source GIS portal. For managing end to end processes including tagging, verification, aggregation and Knowledge graph creation, customized solution needs to be developed as per Indian scenarios. Proposed system is providing end to end solution for geotagged data collection to knowledge graph generation for Bhuvan GIS Portal with ensured quality of data even for less popular places. Study of existing methods and technologies used at each stage is given in subsequent four subsections.

A. Crowd Sourcing Platform

Crowd Sourcing is the practice where large number of people involved in collecting the data contributing in any project / task allocated to them [48]. Crowd sourcing would be a useful tool mainly in Urban Planning, Public Health, Area Mapping, Smart City and Disaster management services. Through crowdsourcing, it is easier to capture and analyze the vast amount of data for better planning and improving the existing systems such as Traffic management, Event Management, Area Planning and E-Governance. Crowdsourcing is a crucial game player in disaster management and recovery activities due to the direct involvement of citizens to get the updated information through various modes like social media and mobile phone apps. Here in this study, we focused on the crowdsourcing done in geotagging context. In a review paper, authors presented the detailed survey about the approaches, technologies used in crowdsourcing and compared various platform using different parameters like Definition language, job support, process control, data management, development, quality control and public availability [18]. They have tested the available prototypes and found the scope to improve motivation of the participants including to make the opportunity available for workers training. Selected research work done in crowdsourcing domain are summarized in table I. As we can see in the table, very few researchers have worked specifically in geotagging using crowdsourcing domain. During the study, some challenges were observed in crowdsourced geotagging which are tag extraction, aggregation and truth discovery in crowdsourced geo tagging which needs to be resolved in order to make the maps reliable.

B. Aggregation

As discussed earlier, there may be redundancy in crowdsourced data hence to manage it efficiently, it should be analyzed properly and redundant data needs to be removed. Research work done in this direction were explored and are summarized in table II.

As we can see various techniques of aggregating crowdsourced data were proposed like binary labels, Games with a Purpose, Mturk, voting strategy, EM Algorithm, confusion matrix-based approach for label estimation, ZenCrowd, GLAD, Mapco Bayes, SQAURE [26,27]. Threefold architecture for mobile crowdsourcing, mobile sensing, eHealth, smart grid, additive homomorphic encryption scheme, Sum and Min aggregate of time-series data, a privacy-preserving sum aggregation and a point to point privacy-preserving data aggregation methods were proposed [45]. Generic benchmark framework was proposed to compare various aggregation approaches like MD, HP, ELICE, EM, GLAD, SLME, and ITER [33]. This issue needs to be looked as per the POI data perspective.

C. Evaluation of Authenticity

In order to ensure the quality of crowdsourced data, there must be a fool proof method to verify it.

We studied various available methods and approaches to ensure the authenticity of collected data. Summary of the related work for truth discovery and verification is shown in table III. Various methods like repetitive methods, enhancement-based, Probability based graphical methods, unsupervised models, Truth finder for Spatial Events without user's location tracking using collapsed Gibbs sampling were used for truth discovery [31].

Machine learning approach using FLOCK machine learning platform, randomized Gaussian mixture model (RGMM), iCrowd, an adaptive crowdsourcing

Table- I: Summary of Related Work Done in Crowd Sourcing

Sr. No	Domain	Key Issue Addressed	Model/Methods	Data Set	Observation	Limitations
1	Crowdsourcing framework for Spatial data [37]	Proposed a generic crowdsourcing framework and dealt with maximum task assignment problem	Least Location Entropy Priority, Close Distance Priority, greedy approach for work assignment in spatial crowd sourcing framework	Synthetic and real data set named Yelp and Gowalla were used.	Maximized the assigned task	A better heuristic needs to be developed for spatial crowd sourcing domain
2	Crowdsourcing labels reliability [5]	Dealt with incorrect label problem to ensure quality of crowdsourced data	Machine learning algorithm named as expectation–maximization, saddle point along with leave-one-out-cross-validation approach was used	TREC 2011 data set was used	Classification accuracy was enhanced by identifying reliable labelers with less training sample	It was not tested for another domain
3	Crowdsourcing marketplace [36]	Non uniform distribution of power in workers of crowd sourcing platform	Open governance model	Data was generated on crowdsourced platform	Improved task quality and fairness with the open governance model	Feasibility in Indian scenario needs to be identified
4	A mobile Application for crowd sourcing [28]	Active Involvement and increasing productivity of the participants in crowdsourcing	Interactive Mobile app for task assignment	Anonymous and known Participants were involved directly	The indulgent and functional values had balance with moderate quality of Horn Heide Questionnaire and performance Qualification.	Final conclusion is pending
5	Disaster management using crowdsourcing [4]	To enhance preparedness in health domain during disaster	Software platform for Gateway for users and server layer to support backend processing	Test cases were bespoken with variety of symptom and cruciality level	Clusters were developed for all level of priority	Testing needs to be done in real time scenario
6	Crowd sourcing enabled machine learning framework [7]	Classification and prediction problem	Measures of Machine learning framework	YouTube, Wikipedia pages and Monet paintings were used	The consensus-weighted version significantly outperforms for classification	Not a generalized approach
7	Crowdsourcing enabled deep learning in biomedical imaging [1]	Very less ground truth data available for biomedical imaging	Aggnet, a deep learning enabled crowd sourcing layer on Convolutional neural network was proposed	CrowdFlower and MICCAI-AMIDA13 challenge data set	Robust for noisy labels	Large number of missing annotations
8	Crowdsourcing enabled disaster management system [25]	Coordination among citizens and volunteers during the disaster	Concepts of situated crowdsourcing as well as ubiquitous crowdsourcing based on a user-centered design approach	Author had planned to test during Kieler Woche sailing festival on real data	Seems useful but testing was not done to judge the performance and effectiveness	Needs to tested and verified
9	Classification of crowdsourced feed [23]	Filtering redundant information	Image similarity along with the analysis of geolocation and text	Limited testcase generated. Testing was planned with real data in future	Comprehensive approach which can be applied on many areas.	Only histogram approach is used for image similarity. Text analysis needs to be enhanced

Table- II: Summary of Related Work Done in Aggregation of Crowdsourced Data

Sr. No	Domain	Key Issue Addressed	Model/Methods	Data Set	Observation	Limitations
1	Aggregation of crowdsourcing data [6]	Distinguish malicious annotator along with the quality check	Active learning strategy with Bayesian updating scheme, a Bradley-Terry model framework	Simulated data was generated	90% of the best accuracy was Achieved	Heuristics for narrowing the sample space needs to be found.
Sr. No	**Domain**	**Key Issue Addressed**	**Model/Methods**	**Data Set**	**Observation**	**Limitations**
2	Aggregation of crowdsourcing data [38]	Extracting accurate labels from crowdsourced datasets	Community-based Bayesian label aggregation model	CrowdFlower and Click Worker	Better performance and scalable implementation	Application areas needs to analyzed
3	Aggregation of crowdsourcing data [12]	Quality of label	Majority voting with unbiased random tie-breaker strategy	Datasets HCB, WB, WVSCM, RTE, Meval, MMSys were used.	Efficiency depends on labelling task and the environment where it is deployed	Unreliable worker's needs to be filtered out automatically
4	Aggregation of crowdsourcing data [45]	Investigation of voluminous crowd sourced data	Three party architecture with Privacy-preserving Verifiable Data Aggregation	SOCR open data was used	Verification cost remains constant irrespective of workers count	Generalization needs to be done to apply on various domain
5	Aggregation of crowdsourcing data [17]	Aggregation of collected label with social choice consideration	Axiomatic framework for collective annotation	New data set generated from Switchboard Corpus	Restrictions e.g. same category for each item and category-exclusivity used during study	Not explored for variant category
6	Aggregation of crowd sourced data [33]	Performance comparison of the available aggregation technique	Component based architecture with computing layer consisting of two modules, aggregation and simulation	MD, HP, ELICE, EM, SLME, GLAD, and ITER techniques for aggregating the response was compared	Results obtained from EM and SLME were most accurate	Improvements needs to be done with respect to large data size.

Table- III: Summary of Related Work Done in Trustworthiness of Crowd sourced Geotagged Data

Sr. No	Domain	Key Issue Addressed	Model/Methods	Data Set	Observation	Limitations
1	Geographical data base [11]	Detecting and Resolving inconsistencies in a Gazetteer	Minimum bounding Rectangle technique	OsmNames, GeoNames	20% places relocated to correct position	The area of children MBR must be greater than or equal to the known area of the place
2	Truth discovery in crowdsourcing application [31]	Verification of quantitative data in crowdsourcing scenarios	Additive and multiplicative model named, QTF-A and QTF-M	Google images posted on CrowdFlower	Human factors were not involved. It is suitable for verification of quantitative data but not for qualitative data	Not suitable for qualitative data
3	Big geodata and crowd sourced gazetteers [35]	Distributed environment for big geodata processing	Python based Web crawler, Distributed Hadoop cluster, ClouderaManagerWebUI	Data extracted from web using web crawler	Complexity of adding the elements on map was reduced by using geojson	Quality check was not performed

4	Truth discovery in crowd sourced data [42]	Identify true facts	Randomized Gaussian Mixture Model with Expectation-Maximization (EM) solutions based on census-taking and specimen.	Weather Forecast, Indoor Floorplan, and Stock Data	Hypothetically effective verification approach. However, worked only on continues attribute.	Suitable only for continues attribute
5	Data reliability in crowd sourced data [2]	Measuring accuracy of each contributor's data on crowd sourcing platform	Consensus grades Computation based on belief propagation, maximum likely hood and Cost of Disagreement, is based on iterated convex optimization	Crowd Grader	The Variance Propagation algorithm with DEBIAS, WAvg and ATT options performs best. Reduction in error is 20%	Suitable only for numerical data
Sr. No	**Domain**	**Key Issue Addressed**	**Model/Methods**	**Data Set**	**Observation**	**Limitations**
6	Crowdsourced data reliability [15]	Multiple noisy labels	PLAT algorithm was proposed based on Active learning approach for labeled and unlabeled data	12 datasets from the UCI database and amazon mechanical Turk repository were used	It handled the imbalanced multiple noisy labeling issue with high performance	Performance can be improved in case of imbalanced dataset
7	Crowdsourced data classification [20]	Classification of reliable data	BIMO, Neural Network based classification model	Vote, credit, biodeg, sick and Income data set were used	Proposed algorithm works on the directly collected crowdsourced data rather than integrated data.	It is suitable only in binary classification.
8.	Noise correction in crowdsourced data [29]	Noisy label correction	STC (self-training correction) and Cluster-based correction algorithms were proposed.	Binary and multiclass data set were used	Cluster based correction method performed better on both the data set (multiclass and binary class)	Improvement in label quality is required in case of very low noise.
9	Aggregation and truth discovery in crowd sourced data [27]	Choosing most reliable information among various sources	Probabilistic Model to learn most reliable source	Game, SFV, KBP 2013 dataset	System learned the question and answer simultaneously and detected true answer	Performance and accuracy rate can be improved
10	Truth discovery in crowdsourced continues data [21]	Verifying truth under dynamic scenario	Incremental Truth discovery framework	HAM weather, Wunderground (Wund)2, World weather Online (WWO)3 data sets	Guaranteed convergence	Temporal relations can be used to improve the integration performance
11	Truth discovery and task assignment on crowd sourcing platform [9]	Knowing worker's accuracy and assigning the task instantly on crowdsourced platform	Adaptive graph-based task Assignment framework	YahooQA and Item Compare data sets	Proposed solution performed better for instant task assignment	Focused on task assignment with answer either yes or no
12	Truth discovery in crowdsourced data [3]	Investigating for reliability and correctness of crowd sourced data	Partial ordering and belief function framework along with Sums and adapted Sums models for fact confidence	DBpedia and Author dataset	Non conflicting assertions were recognized	Other features need to be analyzed to improve the performance
13	Trustworthiness of crowdsourced data [41]	Detecting false information while considering multi source information	Multi-source information trustworthiness analysis machine learning framework	Orbitz, Priceline, and TripAdvisor (2013) datasets	Inconsistent information was found correctly.	Efficient factorization methods
14	Truth discovery in crowdsourced data [42]	Convergence of the truth in crowd sourced data	Randomized Gaussian Mixture Model	Weather Forecast, Indoor Floorplan, and Stock Data	Effective approach with theoretical guarantee for continues attributes	Suitable only on continues attribute

15	Truth discovery in crowdsourced data [39]	Uncertainty of truth value if in case of multiple reliable facts	Kernel Density Estimation approach	TripAdvisor Population(outlier) and unimodal synthetic data sets	With the confirmed single truth value, it performs well for utmost claimed in various sources	It works only for single truth existence.
16	Classification of noisy label in crowdsourced data [24]	Noisy label classification	Robust personal classifier model	Breast cancer dataset -UCI	High Classification accuracy and highly capable to detect the spammers.	Empirical comparison of models is required.
17	POI verification [8]	Identification of out dated POI	Semi-Supervised Verification, classifier (SVM, c4.5, bagging), learning algorithms like RBF, network, Adaboost, SVM and bagging were used	Yellow page website and open data	Improvement was observed in case of SVM in case of first data set. Accuracy was low for open data set	More sources should be used to get the data and accuracy needs to be improve in case of open data
Sr. No	**Domain**	**Key Issue Addressed**	**Model/Methods**	**Data Set**	**Observation**	**Limitations**
18	POI Management [43]	POI collection and management	Service oriented architecture using Baidu Map API, geo server	Baidu Map data	Collection and management of the POI data	No way to verify the collected poi data
19	POI Tag refinement [46]	POI tag refinement	Collaborative learning framework with multilabel classification approach	Beijing and Chengdu data set	Various data sources were used to construct feature for tag refinement	Framework can be used only for tag refinement

Table- IV: Summary of Related Work Done in Knowledgebase Creation

Sr. No	Domain	Key Issue Addressed	Model/Methods	Data Set	Observation	Limitations
1	Knowledgebase creation using crowd sourcing platform [19]	Knowledgebase construction	Crowdsourcing based blackboard model with linked data representation	Multilingual frequently asked questions in a domain of rental apartments, e learning content	A crowdsourcing platform was proposed to create knowledge base.	Evaluation is required
2	Knowledgebase creation using crowd sourcing platform [44]	Knowledgebase construction in cloud computing	Rule engine and service-based architecture was used	Weather fact display, schedule fact displays and schedule notice rules were used	A framework to collect, infer and manage the facts along with Knowledgebase.	Knowledge base and rule engine were designed in a single node.
3	Knowledgebase creation for health care data [34]	Knowledgebase construction in Chinese	Data collected from multiple origins, mixed automatically to construct Knowledge base.	Healthcare data extracted from conventional website and Chinese references	Knowledge base was created with 135,484 entities with 26821 symptom entities	Specific to Chinese healthcare domain.
4	Knowledgebase creation using crowd sourcing [40]	Automatic knowledge base creation	Three interrelated layer system for extraction, reasoning, and human feedback. CRF(Viterbi), Markov chain Monte Carlo (MCMC) inference, and approximate string matching for text analysis	Text corpus data set	Proposed system is consisting of three components such as extraction, reasoning and human feedback. As per the author, it can be used in various application domain	Testing required for scalability and efficiency.

5	Knowledgebase creation [22]	Entity classification and linking	A linked graph and coordinated inference method were used	YAGO, Wikipedia and TAC-KBP2013	Reasonably high accuracy	Large number of iterations are required to improve accuracy.
6	Knowledgebase creation using crowd sourcing [16]	Data collection using Crowd sourcing method	Knowledgebase creation by extracting data from crowdsourcing platform	2011 DBLP dataset	Responses received through Mechanical Turk were more accurate to emphasize the crowdsourcing	Tool should be designed to measure the important properties

framework, QTF-A and QTF-M, and graph-based estimation model were proposed which ensures the truth discovery in crowdsourcing environment [17,27,31]. Multi-source information trustworthiness analysis framework and fine-grained truth discovery model were proposed to find the truth in crowdsourced generated data [26,27]. RPC model, PLAT algorithm, Neural network-based approach, STC and Clustering based algorithm were proposed to correct Noisy and unlabeled data which is an important part of the truth discovery [24,29,30].

As we can see, very few have proposed methods to verify geotagged data. There are many challenges like most of the truth discovery algorithm works for quantitative data not for qualitative data. There are major issues for noise label detection and removal before aggregation.

Generic benchmark frameworks were not tested on real and large data set. Researchers worked for POI verification in recent years like identifying the out dated POI, tag refinement using machine learning approach etc. [8]. Most of the work done for geotagging, quality assurance was achieved either by using social media data or contributor's trust factor, which may not be correct in some cases. In proposed system, multiple sources are used to ensure the quality of geotagged data.

D. Knowledgebase Creation

Knowledge base is the vital component in the GIS framework. With the power of crowd sourcing, huge amount of data can be collected but to make this data a knowledgebase, it's important to represent the facts in an efficient manner. Various previous work done related to knowledge base creation were explored but we have picked only those who have worked on crowdsourced data. Summary of the selected work is given in table IV. Blackboard model-based method to link the relevant data [19] rule engine and SOA based knowledge base creation [44], automatic data fusion to create the Knowledge base were adopted and implemented [33]. MADDEN, MPP, PROB, CIIGA were used for automatic knowledge base creation [40]. LMKB, Ontological data representation and APIs for knowledge base creation were implemented [19].

III. PROPOSED MODEL

During this review process, we observed that there is no open source GIS portal framework available which can directly cater the need of enriching the Indian GIS portal with verification process of the crowdsourced data. As the POI data will grow, manual authentication process will not be feasible. Proposed system will allow users to geotag the locations which will be further verified automatically using social media data along with the layout plan provided by Indian government using GIS techniques involved. All the POI data will be arranged in hierarchy with the help of data mining algorithms to generate a knowledge graph. Component diagram is shown in figure1. There are mainly four components with label A, B, C and D. Component A is the process to get a georeferenced layout plan along with the asset attribute table. To get the authorized layout plan as georeferenced, we are using QGIS software. As an outcome of this processing, we get the attribute list along with the shapefile. Important features of the attribute table are shown in table V. We have created a broad category of POI like commercial, residential, health care etc. Category of each asset is also obtained while preparing the attribute list using QGIS and will be used during verification of POI as one of the criteria. Component B is the Web Portal which facilitates the registered users to geotag or search any POI. It also provides the gateway to super users to manage the entire system including all necessary master data management. Domain expert users are already part of this system and they also contribute their field information using this portal. This portal also connects various government departments so linking various heterogeneous data will give a strong fact for the knowledge graph which will be used to derive useful information.

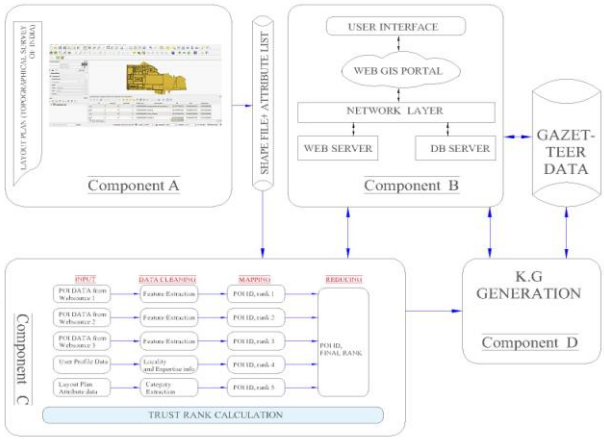

Fig. 1.Component Diagram of Proposed System

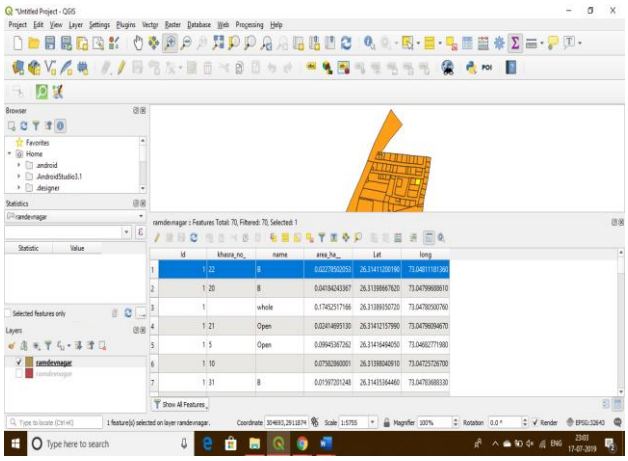

Fig. 2.Digitization of LayOut Plan

Table-V: Sample data obtained from QGIS

Asset Id	Khasra/plot no.	Name description	Latitude	Longitude	Category
18	0	E.S.I. Dispensary	28.378152 1388	76.9091075 323	HealthCare

Table- VI: API usage policy

Sr. no.	Source	API	Price
1	Google	Geocoding	INR356.774999997
2	Mapmyindia	Rev Geocode	No charges for 200 transaction/day
3	Open Street Map	Reverse	Not Applicable

Component C is a verification process for POI data based on Hadoop map reduce framework. POI data which needs to be verified, will be searched at three different web sources using supported APIs of the corresponding service provider. Return values of these API are different in format, hence data preprocessing is required. This cleaned data is sent to different map functions to calculate the rank value. Reducer will collect all the rank values calculated for the POI and emit the final rank values. Trust rank calculation is explained in detail in below subsection A**.**

Component D is the Knowledge graph generation process which uses the verified data.

This data is arranged in a hierarchical manner to get the relationship between these facts. Inference engine will help to make the search efficient.

In order to ensure the trustworthiness of the collected POI, quality must be verified. Therefore, a trust rank is calculated and associated with each POI. The higher the value of trust rank means the greater the reliability of data. Method of trust rank calculation is described in below subsection.

A. Trust Rank Calculation

Trust rank is calculated based on multiple parameters like user's domain expertise, locality, availability in layout map data and in various web sources. During the search on various web sources, it starts with the exact match of the latitude, longitude value. However, in case of unavailability, precision is reduced by ten meters. Based on the precision, rank value is assigned. According to each parameter, rank value is calculated and final value is the sum of all the individual rank values. This is done using the Hadoop map reduce framework which is explained in the section below.

B. Map Reduce Functions

As shown in figure 1, component C "Map reduce functions" are used for each parameter checking. Each function emits the rank value based on the single criteria and reducer will combine the rank values and emit a single final value.

Open Street Map, map my India and google map are used as other web sources to get the info on given POI. API name and access cost are given in table VI. Each source gives different feature names in JSON records. Hence, preprocessing of the json record needs to be done for getting the uniform features. After cleaning of JSON data, the category is predicted and matched with the ground truth data. As per the match result, rank value is assigned.

Similarly, the same POI data is checked for domain expertise and user's locality and the rank value is emitted accordingly.

Once all the map functions are executed, reducer will finally calculate a single value using all these individual values emitted by each map function. This value is sent back to GIS portal database and attached with the respective POI. All the verified POI data is arranged in a hierarchical manner and a knowledge graph is generated which can be used further for various applications.

IV. CONCLUSION AND FUTURE SCOPE

Crowdsourcing is the easiest way of POI data collection but at the same time quality may be a concern for this data. Therefore, in this paper, we studied the verification methods and approaches related to geotagging data collected using crowd sourcing. Paper described a multi-source authentication method for Point of Interest data collected via crowdsourcing techniques. These sources are meta data of layout maps, user locality information, domain expertise and social media data available on different web sources. Digitization of the layout map is done to get the attribute table using QGIS client. Government departments have already started the digitization process hence getting an attribute table will not be an issue during the production stage of the proposed methods. Hadoop map reduce framework is being used in order to make the process efficient. However, we need to evaluate the performance in case of separate jobs for each mapper versus single job for all the mapper functions. More study is required to generate the automatic knowledge graph from this verified POI data. As a future plan, we shall develop the automatic knowledge graph generation using POI data.

REFERENCES

1. Albarqouni, Shadi, Christoph Baur, Felix Achilles, Vasileios Belagiannis, Stefanie Demirci, and Nassir Navab. "AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images." IEEE Transactions on Medical Imaging, vol. 35, ED- 5, pp. 1313–1321, May 2016. doi:10.1109/tmi.2016.2528120
2. Alfaro L, Shavlovsky M. "Crowdsourcing quantitative evaluation: algorithms and empirical results." Technical Report UCSC-SOE-14- 03, School of Engineering, UC Santa Cruz. https://pdfs.semanticscholar.org/cdc8/5dda05559884af5b05f5015288cff5fbaaab.pdf. Accessed 27 July 2020.

3. Beretta, V., Harispe, S., Ranwez, S., & Mougenot, I. "How Can Ontologies Give You Clue for Truth-Discovery? An Exploratory Study." Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics - WIMS'16, 2016. doi:10.1145/2912845.2912848.
4. Besaleva, Liliya I., and Alfred C. Weaver. "Applications of Social Networks and Crowdsourcing for Disaster Management Improvement," 2013 International Conference on Social Computing, Alexandria, VA, 2013, pp. 213-219, doi: 10.1109/SocialCom.2013.38
5. Chen, Pin-Yu, Chia-Wei Lien, Fu-Jen Chu, Pai-Shun Ting, and Shin-Ming Cheng. "Supervised Collective Classification for Crowdsourcing." 2015 IEEE Globecom Workshops, December 2015. doi:10.1109/glocomw.2015.7414077.
6. Chen, Xi, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. "Pairwise Ranking Aggregation in a Crowdsourced Setting." Proceedings of the Sixth ACM International Conference on Web Search and Data Mining- WSDM '13, 2013. doi:10.1145/2433396.2433420.
7. Cheng, Justin, and Michael S. Bernstein. "Flock." Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15, 2015. doi:10.1145/2675133.2675214.
8. Chuang, Hsiu-Min, and Chia-Hui Chang. "Verification of POI and Location Pairs via Weakly Labeled Web Data." Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion (2015). doi:10.1145/2740908.2741715.
9. Fan, Ju, Guoliang Li, Beng Chin Ooi, Kian-lee Tan, and Jianhua Feng. "iCrowd." Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD '15, 2015. doi:10.1145/2723372.2750550
10. Farahat, Ahmed K., Ahmed Elgohary, Ali Ghodsi, and Mohamed S. Kamel. "Distributed Column Subset Selection on MapReduce." 2013 IEEE 13th International Conference on Data Mining. December 2013. doi:10.1109/icdm.2013.155.
11. Gao, Song, Linna Li, Wenwen Li, Krzysztof Janowicz, and Yue Zhang. "Constructing Gazetteers from Volunteered Big Geo-Data Based on Hadoop." Computers, Environment and Urban Systems 61 (January 2017): 172–186. doi: 10.1016/j.compenvurbsys.2014.02.004.
12. Georgescu, Mihai, and Xiaofei Zhu. "Aggregation of Crowdsourced Labels Based on Worker History." Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14) - WIMS '14, 2014. doi:10.1145/2611040.2611074.
13. Hong, Richang, Yang Yang, Meng Wang, and Xian-Sheng Hua. "Learning Visual Semantic Relationships for Efficient Visual Retrieval." IEEE Transactions on Big Data 1, no. 4. pp.152–161 December, 2015. doi:10.1109/tbdata.2016.2515640.
14. Huang, Jinxin, Lin Niu, Jie Zhan, Xiaosheng Peng, Junyang Bai, and Shijie Cheng. "Technical Aspects and Case Study of Big Data Based Condition Monitoring of Power Apparatuses." 2014 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC). December 2014. doi:10.1109/appeec.2014.7066164.
15. J. Zhang, X. Wu and V. S. Shengs, "Active Learning with Imbalanced Multiple Noisy Labeling," in IEEE Transactions on Cybernetics, vol. 45, no. 5, pp. 1095-1107, May 2015, doi: 10.1109/TCYB.2014.2344674
16. Kobren, Ari, Thomas Logan, Siddarth Sampangi, and Andrew McCallum. "Domain Specific Knowledge Base Construction via Crowdsourcing." In Neural Information Processing Systems Workshop on Automated Knowledge Base Construction AKBC, Montreal, Canada. 2014 Google Scholar
17. Kruger, Justin, Ulle Endriss, Raquel Fernández, and Ciyang Qing. "Axiomatic analysis of aggregation methods for collective annotation." In Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems, pp. 1185-1192. International Foundation for Autonomous Agents and Multiagent Systems, 2014 Google Scholar
18. Kucherbaev, Pavel, Florian Daniel, Stefano Tranquillini, and Maurizio Marchese. "Crowdsourcing Processes: A Survey of Approaches and Opportunities." IEEE Internet Computing 20, no. 2.pp. 50–56. March 2016. doi:10.1109/mic.2015.96.
19. Kuwabara, K., & Ohta, N. (2014). "Toward a Crowdsourcing Platform for Knowledge Base Construction". eKNOW 2014, The Sixth International Conference on Information, Process, and Knowledge Management, (c), 89–92
20. Li, Jingjing, Victor S. Sheng, Zhenyu Shu, Yanxia Cheng, Yuqin Jin, and Yuan-feng Yan. "Learning from the Crowd with Neural Network." 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA). December,2015. doi:10.1109/icmla.2015.14.
21. Li, Yaliang, Qi Li, Jing Gao, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. "On the Discovery of Evolving Truth." Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15. 2015. doi:10.1145/2783258.2783277.
22. Lin, Hailun, Yantao Jia, Yuanzhuo Wang, Xiaolong Jin, Xiaojing Li, and Xueqi Cheng. "Populating Knowledge Base with Collective Entity Mentions: A Graph-Based Approach." 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014) August,2014. doi:10.1109/asonam.2014.6921648.
23. Liu, Kaixu, Gianmario Motta, Linlin You, and Tianyi Ma. "A Threefold Similarity Analysis of Crowdsourcing Feeds." 2015 International Conference on Service Science (ICSS). May, 2015. doi:10.1109/icss.2015.14.
24. Liu, Zhiquan, Luo Luo, and Wu-Jun Li. "Robust Crowdsourced Learning." 2013 IEEE International Conference on Big Data. October, 2013. doi:10.1109/bigdata.2013.6691593.
25. Ludwig, Thomas, Christoph Kotthaus, and Volkmar Pipek. "Situated and Ubiquitous Crowdsourcing with Volunteers During Disasters." Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, September, 2016. doi:10.1145/2968219.2968585.
26. Ma, Fenglong, and Jing Gao. "Probabilistic Models for Fine-Grained Truth Discovery from Crowdsourced Data." 2015 IEEE International Conference on Data Mining Workshop (ICDMW). November, 2015. doi:10.1109/icdmw.2015.109.
27. Ma, Fenglong, Jiawei Han, Yaliang Li, Qi Li, Minghui Qiu, Jing Gao, Shi Zhi, Lu Su, Bo Zhao, and Heng Ji. "FaitCrowd." Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15, 2015. doi:10.1145/2783258.2783314.
28. Melenhorst, Mark, Jasminko Novak, Isabel Micheel, Martha Larson, and Martin Boeckle. "Bridging the Utilitarian-Hedonic Divide in Crowdsourcing Applications." Proceedings of the Fourth International Workshop on Crowdsourcing for Multimedia - Crowd MM '15. 2015. doi:10.1145/2810188.2810191.
29. Nicholson, Bryce, Jing Zhang, Victor S. Sheng, and Zhiheng Wang. "Label Noise Correction Methods." 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). October, 2015. doi:10.1109/dsaa.2015.7344791.
30. xof Image Labeling in Crowdsourcing." 2015 IEEE International Conference on Image Processing (ICIP). September 2015. doi:10.1109/icip.2015.7351042.
31. Ouyang, Robin Wentao, Lance M. Kaplan, Alice Toniolo, Mani Srivastava, and Timothy J. Norman. ""Aggregating Crowdsourced Quantitative Claims: Additive and Multiplicative Models," in IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 7, pp. 1621-1634, 1 July 2016, doi: 10.1109/TKDE.2016.2535383
32. Ouyang, Robin Wentao, Mani Srivastava, Alice Toniolo, and Timothy J. Norman. "Truth Discovery in Crowdsourced Detection of Spatial Events." Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM '14. 2014. doi:10.1145/2661829.2662003.
33. Quoc Viet Hung, Nguyen, Nguyen Thanh Tam, Lam Ngoc Tran, and Karl Aberer. "An Evaluation of Aggregation Techniques in Crowdsourcing." Web Information Systems Engineering – WISE 2013.pp 1–15. doi:10.1007/978-3-642-41154-0_1.
34. Ruan, Tong, Mengjie Wang, Jian Sun, Ting Wang, Lu Zeng, Yichao Yin, and Ju Gao. "An Automatic Approach for Constructing a Knowledge Base of Symptoms in Chinese." 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Shenzhen, 2016, pp. 1657-1662, doi: 10.1109/BIBM.2016.7822767.
35. S. Pradeepa and K. R. Manjula, "Construction of gazetteers from geo big data using machine learning technique on Hadoop," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 1619-1622.
36. Snehal Gaikwad, Durim Morina, Rohit Nistala, Megha Agarwal, Alison Cossette, Radhika Bhanu, Saiph Savage, Vishwajeet Narwal, Karan Rajpal, Jeff Regino "Daemo: A self-governed crowdsourcing marketplace." In Adjunct Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology, pp. 101-102. ACM, 2015. soi:10.1145/2815585.2815739.
37. To, Hien, Cyrus Shahabi, and Leyla Kazemi. "A Server-Assigned Spatial Crowdsourcing Framework." ACM Transactions on Spatial Algorithms and Systems 1, no. 1. pp 1–28 , August 13, 2015. doi:10.1145/2729713.

38. Venanzi, Matteo, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. "Community-Based Bayesian Aggregation Models for Crowdsourcing." Proceedings of the 23rd International Conference on World Wide Web - WWW '14, 2014. doi:10.1145/2566486.2567989.
39. Wan, Mengting, Xiangyu Chen, Lance Kaplan, Jiawei Han, Jing Gao, and Bo Zhao. "From Truth Discovery to Trustworthy Opinion Discovery." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 13, 2016. doi:10.1145/2939672.2939837.
40. Wang, Daisy Zhe, Yang Chen, Sean Goldberg, Christan Grant, and Kun Li. "Automatic knowledge base construction using probabilistic extraction, deductive reasoning, and human feedback." In Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction, pp. 106-110. Association for ComputationalLinguistics, 2012.https://www.aclweb.org/anthology/W12-3020/
41. Xiao, Houping, and Jing Gao. "Multi-Source Information Trustworthiness Analysis." 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, 2015, pp. 1600-1601, doi: 10.1109/ICDMW.2015.212.
42. Xiao, Houping, Jing Gao, Zhaoran Wang, Shiyu Wang, Lu Su, and Han Liu. "A Truth Discovery Approach with Theoretical Guarantee." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 13, 2016. doi:10.1145/2939672.2939816.
43. Yuanrong He, Yuanmao Zheng, Jian Deng, and Huoping Pan. "Design and Implementation of a POI Collection and Management System Based on Public Map Service." 2016 Fourth International Conference on Ubiquitous Positioning, Indoor Navigation and Location Based Services (UPINLBS) (November 2016). doi:10.1109/upinlbs.2016.7809971.
44. Zhou, Rui, Jing Li, Jinghan Wang, and Guowei Wang. "A Knowledge-Based Development Approach with Fact and Service for End-User in Cloud Computing." 2013 IEEE 37th Annual Computer Software and Applications Conference Workshops, Japan, 2013, pp. 277-282, doi: 10.1109/COMPSACW.2013.44.
45. Zhuo, Gaoqiang, Qi Jia, Linke Guo, Ming Li, and Pan Li. "Privacy-Preserving Verifiable Data Aggregation and Analysis for Cloud-Assisted Mobile Crowdsourcing." IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications, San Francisco, CA, 2016, pp. 1-9, doi: 10.1109/INFOCOM.2016.7524547.
46. Zhou, Jingbo, Shan Gou, Renjun Hu, Dongxiang Zhang, Jin Xu, Airong Jiang, Ying Li, and Hui Xiong. "A Collaborative Learning Framework to Tag Refinement for Points of Interest." Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (July 25, 2019). doi:10.1145/3292500.3330698.
47. https://wiki.openstreetmap.org/wiki/Planet.osm
48. https://crowdsourcingweek.com/what-is-crowdsourcing/

AUTHORS PROFILE

Ms. Monika Sharma is currently pursuing the Ph.D. degree in Computer Science & Engineering with Poornima University, Jaipur, India. She completed M. Tech in Computer Science and Engineering in 2005. She had worked at various positions in the software industry and Academics. She was awarded the National Merit Scholarship Award from the ministry of HRD, Government of India, for excellent performance in Secondary Board Exam. She has presented and published many papers in conferences and journals. She is a member of IEEE, IETE and ACM. Her area of research is GIS, Data Science and software engineering.

Mr. Vinod Bothale (**Deputy Director of Data Processing Area**) has obtained his Bachelor's degree in Electrical Engg from VNIT, Nagpur and is gold medalist from IIT Roorkee during his Masters in Engg. He joined ISRO Satellite Center in 1986 and worked on design of control systems for IRS and INSAT satellites. Subsequently at Regional Remote Sensing Centre, Jodhpur, he immensely contributed in development of decision support systems using geo-spatial technologies. He was a regional coordinator for Village Resource Centre (VRC) project and provided Web GIS solutions for natural resource management under VRC using open source software. On deputation to German Aerospace Centre-DLR, he worked on microwave SAR interferometry and contributed to improve the SAR processing chain at DLR, Germany. During 2009-2012, he was a Director, Maharashtra Remote Sensing Application Centre (MRSAC), Nagpur on deputation from ISRO. He operationalized MRSAC Geoportal for enabling e-governance in the state of Maharashtra. He also brought reforms in administrative and technical areas in MRSAC, Nagpur. He has rich experience in Image processing of Remote Sensing data, Geospatial applications to natural Resources management and Software application development. He has been the designer and architect of Spatial Data Infrastructures (SDI) and Web Portals. He has been founder designer of Bhuvan Geoportal and water resources portal, India-WRIS. Presently at NRSC/ISRO, he is a Deputy Director, Data Processing Area. His areas of research interest are GIS software, Image Processing, Data Processing and Remote sensing etc.

Dr. Meenakshi Nawal is currently working as Associate Professor with the Poornima University, Jaipur, India. She is Gold Medalist in M.Sc. (IT) from MDS University Ajmer. She has completed Ph. D (Computer Science) from Banasthali Vidhyapith, Jaipur. She is having 13 years' experience of Academics and Research. Her area of research is "Patient Authentication and Security measures in Remote Health Monitoring". She has attended many National and International Workshops, Conferences and Published papers in National conference. Her areas of research interest are Machine Learning, Deep Learning, Image Processing, Big Data Analytics and Software Defined Networks etc.

Professor (Dr.) Mahesh M. Bundele has completed his Bachelor's degree in Electronics and Power in 1986 from Nagpur University and immediately joined as Lecturer in Electronics at Babasaheb Naik College of Engineering, Pusad, Yavatmal district. He did his Master's in Electrical Power System and Doctoral in Computer Science & Engineering with a topic "Design and Implementation of Wearable Computing System for the Prevention of Road Accidents" from Amravati University in 1990 and 2013 respectively. He has worked as Lecturer, Assistant Professor, Professor and Head of CSE & IT during 25 Years at BNCOE and guided many research projects at UG and PG level on various applications in Electrical, Electronics & Computer Sciences including city and village Wi-Fi/Wi-Max design up to 195 villages. He has worked on various govt. and industry research projects. He was also appointed as Principal of Babasaheb Naik College of Engineering, Pusad from March 2011. He has worked in various capacities such as, member Board of Studies, Chief-Valuation officer etc. at University level. He has also worked for getting ISI to Krishak Motor pumps at Amravati. He has visited US, UK, China and Malaysia for research presentations. He has published more than 50 research papers in National and International conferences and journals. He is senior member of IEEE, Life member of ISTE and IEI and the member of ACM. Presently he is working as Member, STDCOM Technical & Professional Activities, and Member Execom, IEEE Delhi Section & Secretary, IEEE Rajasthan Subsection. He is having total 34 years of teaching including 7 years of research and internship. Presently he is working as Principal & Director, Poornima College Engineering, Jaipur, India. He has also worked Dean (R&D) at Poornima University and Heading Advanced Studies & Research Center dealing Master of Technology programs and Doctoral degree programs of the University. His areas of research interest are Wearable & Pervasive Computing, Software Defined Networks, Wireless Sensor Networks, and Smart Grid Issues etc.