A Survey on Location-Based Geospatial Data Mining using Digital Image Processing Techniques

Anita Harsoor, Bhavani Soma

Poojya Doddappa Appa College of Engineering

Corresponding author:Bhavani Soma, Email: bsssoma3@gmail.com

In our daily lives, we face many potential problems, such as potholes, garbage filling, uncomfortable parking, broken/damaged roads, etc. To report a complaint about this kind of problem becomes a great hassle. To solve such problems we will develop a Geospatial Data Mining System called GDMS which will be an android application for the retrieving and analysing of textual data with geographical location and image as proof of the problem which can be used anywhere and at any time. The system contains four components: data collection, data pre-processing, data analysis, and data visualization. GDMS system will be used to collect data that will be used by common people/civilians to register the complaint about the problem that was found. This paper presents research done on relevant topics/techniques of image processing.

Keywords:Data Pre-processing, K-Means, SVM, Data Mining

2023. In Saroj Hiranwal & Garima Mathur (eds.), *Artificial Intelligence and Communication Technologies*, 569–575. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-955020-5-9-55

1 Introduction

In our daily life, there exist many possible problem, such as street lamp damage, manhole cover damage, road damage/broken, water leakage, uncomfortable parking, etc. Manually detecting such problems may require a lot of resources like spending a lot of manpower, a waste of financial resources and it also requires a lot of time. To collect such possible safety problems/issues in time and to find out the most urgent problems to be solved without much of the resources we will build a GDMS system.

To solve this kind of problem, GDMS system developed for collecting unusual events by giving in the textual descriptions with geographical locations and image as a proof. It will provide convenience for the public to register a complaint about the problem found around them at anytime and anywhere. Once a civilian finds comes across any such problem, he/she can use the GDMS system to report about it by posting the required information. Other community residents may also upload similar information of the same clue, which may result in the system collecting much redundant information on similar events, represented as geospatial data containing geographical location, text description, and photos.

Then the system needs to process, index, and analyze such events for efficient retrieval and access. By collecting huge amount of data the need arises to develop an effective data storage and analysis tool. To address the problem, we propose a novel geospatial data mining system called GDMS. In our system, the collected data are preprocessed and analyzed by combining clustering techniques with classification models, based on group of similarities which will be done by using k-means clustering technique, these groups will further be classified using svm classification technique where line or hyperplane will divide the groups and then the groups will be labeled and moved to the class to which group belongs to. We will than visualize the clustering and classification results with a degree of urgency on the google map to the government staff. The block diagram of the GDMS system is shown in the Figure 1.

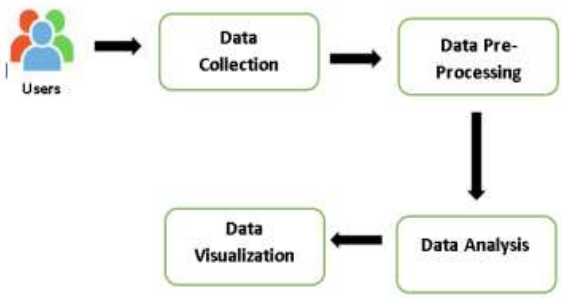

Figure 1. Block Diagram of Geospatial Data Mining System.

The block diagram of the GDMS system consists of four main components. The following are the components of the system:

i. Data collection: a large number of geospatial data is collected i.e., textual description and geo-location of the problem with images as proof from the Geo-spatial Data Mining (GDMS) system that is used by common people, to report social problems.

ii. Data Pre-Processing: Clustering which is an unsupervised approach will be used in preprocessing phase. K-means algorithm is the clustering technique focuses on the problem of finding textual topics of clusters containing text descriptions with geographical locations with a cluster of images.

iii. Data analysis: the clustered group will further be classified using SVM/CNN Classification technology, which is a supervised approach, where the similar groups will be moved to one class and the other group will be moved to other class.

iv. Data visualization: we will demonstrate an effective visualization tool that show the location of the particular problem on the map.

2 Methods

In this section we are discussing about the different methodologies used in the relevant work. After reviewing the related works to outline the research problem we need to define problems/events, which have multiple interpretations which we face in our daily life such as potholes, garbage filling, uncomfortable parking, water leakage, etc. We define the event as a real-world occurrence that can be characterized by time, context, and possible location.

a. Data Collection: The data collected are textual descriptions with geographical location and photos as proof of the event that common people have come across [1]. In previous work, different data collection methods like supervised, unsupervised ([3], [4]), and in some research works [5] semi-supervised methods have also been used for helping researchers to understand which method for data collection will be helpful for their domain. In the case of the supervised method, data must be trained in advance, whereas for the unsupervised method data can be trained in the later process, and semi-supervised falls in between supervised and semi-supervised.

b. Data Pre-Processing: The data collected will be used for pre-processing the data. Clustering is the pre-processing technique used in most research studies. Clustering is the task of dividing the data sets into a certain number of clusters in such a manner that the data points belonging to a cluster have similar characteristics. Clusters are nothing but the grouping of data points such that the distance between the data points within the clusters is minimal. K-Means clustering is one of the most widely used algorithms ([1], [12], [15]). It partitions the data points into k clusters based upon the distance metric used for the clustering. The value of 'k' is to be defined by the user. The distance is calculated between the data points and the centroids of the clusters. ProFreq and LocFreq [13] are the clustering technique where locations with text are clustered in a group based on their similarities, location will be used from users' profiles or from the location of the event posted.

c. Data Analysis: The data analysis is done, by the combination of clustering and classification, where we can categorize documents into various zones and generate various data sets according to their geospatial features. Support Vector Machine (SVM) ([1], [2], [12])or Convolutional Neural Network (CNN) [1] are the classification techniques used for classifying the features based on their similarities and in some of the research works knowledge graph-based similarity method ([1], [4], [10], [14]) have been used.

d. Data Visualization: The problem that is reported by the common people using a mobile application [1] or web browser can be viewed which will give the knowledge of the most common and urgent problem. Google Maps geotagging API [15] with a textual description of the problem will be displayed on the screen [1].

3 Results

The methodologies used by the authors in their work and percentage of accuracy is as shown in the Table 1.

Table 1. Methodology and Accuracy Results.

Sl. No.	Methodology Used	Description	Percentage of accuracy	Reference
1.	Support vector machine (SVM)	• Support Vector Machine (SVM) is a supervised machine learning algorithm. It's a supervised learning algorithm that is mainly used to classify data into different classes. SVM trains on a set of label data. • All of the data in which the feature vector lies are on one side of the hyperplane belong to one class, and the others belong to another class.	96.59 %	1, 2, 3, 4, 12
2.	The lazy Learning algorithm, K Nearest Neighbor (K-NN) algorithm	• K-NN algorithm compares the new data/case with the already available data and based on the similarity the new data will be moved to that particular case. • It is called a lazy learning algorithm because it will not train the data set immediately but instead stores the data set and during classification performs the action on the data set.	84.21 %	3, 4, 10, 14
3.	Convolutional Neural Networks (CNN)	• Convolutional neural networks have become the state of the art for many visual applications such as image classification, and have also found success in natural language processing for text classification. • Convolutional neural networks are very good at picking up on patterns in the input image, such as lines, gradients, circles, or even eyes and faces. They can operate directly on a raw image and do not need any preprocessing. • This text classification approach cannot process sentences longer than the width of the input matrix. It requires a lot of time for processing the sentence.	97.2 %	1
4.	K-means	• K-means algorithm is the most commonly used unsupervised	93.1%	1, 12, 15

		clustering algorithm that divides the object/data into similar and dissimilar data and creates a cluster of similar data. • The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.		
5.	Adaptive Affinity Propagation (adaptive AP)	• Adaptive AP uses adaptive preference scanning to search the space of the number of clusters and finds the optimal clustering solution suitable to a data set by the cluster validation technique. • In adaptive AP the adaptive damping is designed to eliminate oscillations automatically instead of manually, and the adaptive escaping is developed to eliminate oscillations when the damping technique fails.	78.32 %	5

4 Discussion

The comparison of the methodologies based on the percentage of accuracy are shown in the Figure 2.

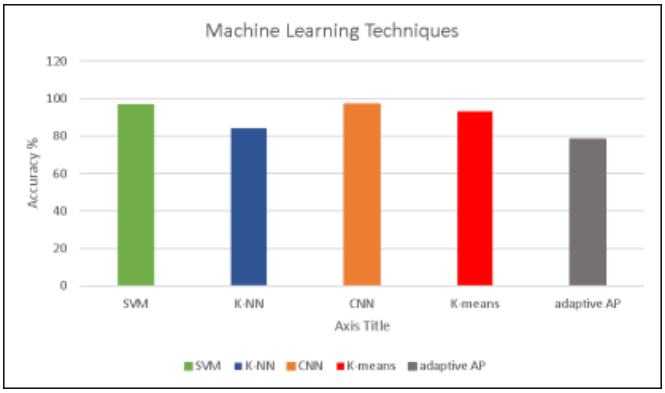

Figure 2. Comparison of Techniques Based on Percentage of Accuracy.

The study of the relevant work has concluded that the combination of both clustering and classification methods will give appropriate results. CNN has the limitation of taking a longer processing time in the case of textual classification. Using SVM with K-means will give more efficient results for processing both the images and text in future work.

5 Conclusion

After reviewing the techniques this paper concludes that the combination of classification and clustering will give more efficient results than using them separately. K-NN uses already trained data to classify the data, which is not appropriate in the case of our system. CNN has the limitation of taking a longer processing time in the case of textual classification. SVM trains on a set of label data. K-means divides the object/data into various parts and based on the similarity data is grouped and then it keeps on iterating until no best cluster is found. Text and geographical location are the data used in some of the recent works. SVM and K- means are the more appropriate classification and clustering methods that can be used in future works for fast, smooth and better results.

References

[1] Meihong Wang, Linling Qiu, Xiaoli Wang, (2019). “GDMS: A Geospatial Data Mining System for Abnormal Event Detection and Visualization”. IEEE International Conference on Mobile Data Management (MDM), 2019, DOI 10.1109/MDM.00-34

[2] C. H. Lee, H. C. Yang, and S. H. Wang, (2015). “A Location-Based Text Mining Approach for Geospatial Data Mining”. IEEE International Conference on Innovative Computing, pp. 1172–1175.

[3] X. Wang, Y. Wang, C. Gao, et al, (2018). “Automatic Diagnosis With Efficient Medical Case Searching Based on Evolving Graphs”. IEEE ACCESS, vol. 6, pp. 53307–53318

[4] S. Beliga, Ana M, Sanda M-I, (July 2015). “An Overview of Graph-Based Keyword Extraction Methods and Approaches”. JIOS Journal of Information and Organizational Sciences, VOL. 39, NO. 1.

[5] Wafa Zubair, Farzana Kabir Ahmad, and Siti Sakira Kamaruddin, (2020). “A Survey on Event Detection Models for Text Data Streams”. International Journal of Computer Science, 16 (7): 916.935 DOI: 10.3844/jcssp.2020.916.935

[6] Shengtian Sang, Zhihao Yang, Lei Wang, Xiaoxia Liu, Hongfei Lin, Jian Wang (2018). “SemaTyP: a knowledge graph based literature mining method for drug discovery“. National Library of Medicine National Center of Biotechnology Information DOI: 10.1186/s12859-018-2167-5, 19:193.

[7] Becker, H., Naaman, M., Gravano, (2010). “L.: Learning similarity metrics for event identification in social media.”. ACM, In: WSDM, pp.291–300.

[8] F. Xiao, S. Zhang, W. Liang, et al, (2018). “Efficient Location-Based Event Detection in Social Text Streams”. International Conference on Intelligent Science & Big Data Engineering, pp. 213–222.

[9] C. H. Lee, H. C. Yang, and S. H. Wang, (2009). "A Location-Based Text Mining Approach for Geospatial Data Mining". IEEE International Conference on Innovative Computing, pp. 1172-1175.

[10] Yongchun Xie, Yong Wang, and Linfeng Li, (06 Aug 2021). “Knowledge Graph-Based Image Recognition Transfer Learning Method for On-Orbit Service Manipulation”. Beijing Institute of Control Engineering, Beijing, China Science and Technology on Space Intelligent Control Laboratory, Beijing, China, Vol. 2021, Article ID 9807452, 9 pages, https://doi.org/10.34133/2021/9807452.

[11] T. Guo, K. Feng, G. Cong, et al, (2018). “Efficient Selection of Geospatial Data on Maps for Interactive and Visualized Exploration”. International Conference on Management of Data, pp. 567–582.

[12] C. H. Lee, H. C. Yang, T. F. Chien and W. S. Wen, (2011). " A Novel Approach for Event Detection by Mining Spatio-temporal Information on Microblogs”. IEEE International Conference on Advances in Social Networks Analysis and Mining.

[13] G. Acampora et al. (2020). “Automatic Event Geo-Location in Twitter”. IEEE ACCESS, July 2020. DOI: 10.1109/ACCESS.3008641

[14] X. Wang, Y. Wang, C. Gao, et al, (2018). "Automatic Diagnosis With Efficient Medical Case Searching Based on Evolving Graphs". IEEE ACCESS, vol. 6, pp. 53307–53318.

[15] P. Giridhar, T. Abdelzaher, J. George, and L. Kaplan, (Apr. 2015). "Event localization and visualization in social networks,". in Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS), pp. 35–36.

[16] B.J. Frey and B. Dueck, (2007). "Clustering by Passing Messages Between Data Points". March 2008 Scienc, DOI:10.1126/science.1150938, Vol. 315, No. 5814, p.p.972-976.

[17] C. Sallaberry, P. Etcheverry, and C. Marquesuzaa, (2006). "Information Retrieval and Visualization Based on Documents' Geospatial Semantics". In Poceedings of International Conference on Information Technology: Research and Education, Tel Aviv, Israel, p.p. 277-282.

[18] C.B. Jones, H. Alani, and D. Tudhope, (2001). "Geographic Information Retrieval with Ontologies of Place". in Spatial Information Theory, p.p. 322-335.

[19] D. Buscaldi, P. Rosso, and P. Peris, (2006). "Inferring Geographical Ontologies from Multiple Resources for GeoGraphical Information Retrieval". In proseedings of the 3rd GIR Workshop, SIGIR 2006, http://users.dsic.upv.es/~prosso/resources/BuscaldiEtAl_GIR06.pdf, 17 - 3 self.

[20] D.J. Sebald, J.A. Bucklew, (2000). "Support Vector Machine Techniques for Nonlinear Equalization". IEEE Transactions on Signal Processing, Vol. 48, No. 11, p.p. 3217-3226.