ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

Study on Extracting Geo-Spatial Information from Open Street Map using Machine Learning

Anita Janmeda[1], Dr.Dinesh Kumar[2]

P.G. Student, Department of Computer Science & Engineering, SRCEM at Palwal, Haryana, India[1]

Associate Professor, Department of Computer Science & Engineering, SRCEM at Palwal, Haryana, India[2]

ABSTRACT: OSM is often referred to as the Wikipedia map of the world. As it is based on a large number of the same ICT structures as Wikipedia it offers its venture givers the likelihood of a) relatively quick refreshing of the guide database and also extremely visit refreshing of related altering programming and different apparatuses; b) bringing in geodata recorded from Global Positioning System (GPS)- empowered gadgets, cell phones, and other computerized maps devices; c) access to the full history of mapping exercises in OSM over its lifetime; lastly d) cooperation with other OSM clients and patrons through different correspondence channels including mailing records, exchange discussions, and physical gatherings (Mooney and Corcoran, 2013a). The gradual evolution of the OSM ecosystem has been very successful. Therefore, in this scheme we illustrate the usage of OpenStreetMaps with SVM.

KEYWORDS: Geo-Spatial Information ,Open Street Maps (OSM),Crowed Sourcing, Machine Learning, Support Vector Machine (SVM)

I. INTRODUCTION

The project got off to a slow start but since 2007 there has been an ever-increasing rate of people joining the project. In November 2014, OSM had approximately 1.85 million registered users and contributors (http://wiki.openstreetmap.org/wiki/Stats). As mentioned previously, the era of ubiquitous Internet, social media, open-source software, etc. has seen many citizen knowledge-based projects for a host of diverse purposes launched on the Internet over the last few years. OSM has been a unique case. The scholastic and mechanical networks have perceived OSM not exclusively in light of its ascent to end up an essential wholesaler of geodata yet its more extensive achievement in growing a worldwide network of individuals willing to take part in the gathering and upkeep of geodata. The OSM community is actively involved in much more than collecting geodata to build and maintain this global geodatabase. In addition, the community is involved in, for example, humanitarian work, open source software development to support OSM and the GIS community, and in building a network of support for those using and contributing to the OSM project. In recent years, several scientific disciplines (e.g. topography, GIScience, spatial arranging, cartography, software engineering, and environment) have understood the monstrous capability of OSM and it has turned into the subject of scholarly research. OSM offers specialists a novel dataset that is worldwide in scale and a group of learning made and kept up by a vast shared system of volunteers. Research on OSM has demonstrated that its geodata in a few sections of the world are more total and locationally and semantically more precise than the comparing restrictive datasets (e.g., Zielstra and Zipf, 2010; Neis and Zipf, 2012; Helbich et al., 2012), while being of high spatial heterogeneity. Incredulity among the GIS people group and industry encompassing the nature of the geodata in OSM has seen a noteworthy exertion being made on assessing the nature of the OSM geodata. This has a prompted the advancement of various programming devices and philosophies for examining the quality (Roick et al., 2011; Helbich et al., 2012; Barron et al., 2013; JokarArsanjani et al., 2013 a; JokarArsanjani and Vaz, 2015). Different methodologies even endeavor to enhance the OSM information through calculations devoted to particular question writes, for example, addresses for geocoding (Amelunxen, 2010). Examination of the advancement and development of OSM over the globe after some time has additionally risen as an exploration subject for some, scholarly investigations

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

(Mooney et al., 2012; Neis et al. 2012, 2013, JokarArsanjani et al., 2013c; Mooney and Corcoran, 2013, Fan et al. 2014)..

The sheer scale of new mapping applications is evidence of a step change in the Geographical Web (GeoWeb). Mapping has gained prominence within the range of applications known as Web 2.0, and the attention that is given to this type of application in the higher echelons of the hi-tech circles is exemplified by a series of conferences, 'Where 2.0', which were started in 2006 by O'Reilly Media –probably the leading promoters of hi-tech knowhow: 'GIS has been around for decades, but is no longer only the realm of specialists. The Web is now flush with geographic data, being harnessed in a usable and searchable format.' (Where 2.0, 2008) While Haklay, Singleton and Parker (2008) and Elwood (2009) provide an overview of this Web Mapping 2.0 landscape, one of the most interesting aspects is the emergence of crowd sourced information. Crowd sourcing is one of the most significant and potentially controversial developments in Web 2.0. This term developed from the concept of outsourcing where business operations are transferred to remote cheaper locations (Friedman, 2006). Similarly, crowdsourcing is how large groups of users can perform functions that are either difficult to automate or expensive to implement (Howe, 2006). The reason for the controversial potential of crowd sourcing is that it can be a highly exploitative activity, in which participants are encouraged to contribute to an alleged greater good when, in reality, the whole activity is contributing to an exclusive enterprise that profits from it. In such situations, crowd sourcing is the ultimate cost reduction for the enterprise, in which labour is used without any compensation or obligation between the employer and the employee

The preparation of the datasets for comparison included some manipulation. The comparison was carried out for the motorways in the London area on both datasets to ensure that they represent roughly the same area and length. Complex slip road configurations were edited in the OSM dataset to ensure that the representation was similar to the one in Meridian. The rest of the analysis was carried out by creating a buffer around each dataset, and then evaluating the overlap. As the Ordnance Survey represents the two directions as a single line, it was decided the buffer around the Meridian should be set at 20 metres (as this is the filter that the Ordnance Survey applies in the creation of the line) and, to follow Goodchild and Hunter's method, the OSM dataset was buffered with a 1-metre buffer to calculate the overlap.

Motorway	Percentage
M1	87.36%
M2	59.81%
M3	71.40%
M4	84.09%
M4 Spur	88.77%
M10	64.05%
M11	84.38%
M20	87.18%
M23	88.78%
M25	88.80%
M26	83.37%
M40	72.78%
A1(M)	85.70%
A308(M)	78.27%
A329(M)	72.11%
A404	76.65%

Table 1 – Percentage overlap between Meridian and OSM buffers Based on this analysis, we can conclude that with an average overlap of nearly 80% and variability from 60% up to 89%, the OSM dataset provides a good representation of motorways.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

SVM Classification is concerned with building a model that separates data into distinct classes. This model is built by inputting a set of training data for which the classes are pre-labeled in order for the algorithm to learn from. The model is then used by inputting a different dataset for which the classes are withheld, allowing the model to predict their class membership based on what it has learned from the training set. Well-known classification schemes include decision trees and Support Vector Machines, among a whole host of others. As this type of algorithm requires explicit class labeling, classification is a form of supervised learning.

As mentioned, Support Vector Machines (SVMs) are a particular classification strategy. SMVs work by transforming the training dataset into a higher dimension, which is then inspected for the optimal separation boundary, or boundaries, between classes. In SVMs, these boundaries are referred to as hyperplanes, which are identified by locating support vectors, or the instances that most essentially define classes, and their margins, which are the lines parallel to the hyperplane defined by the shortest distance between a hyperplane and its support vectors. Consequently, SVMs are able to classify both linear and nonlinear data.

The grand idea with SVMs is that, with a high enough number of dimensions, a hyperplane separating a particular class from all others can always be found, thereby delineating dataset member classes. When repeated a sufficient number of times, enough hyperplanes can be generated to separate all classes in n-dimensional space. Importantly, SVMs look not just for any separating hyperplane but the maximum-margin hyperplane, being that which resides equidistance from respective class support vectors.

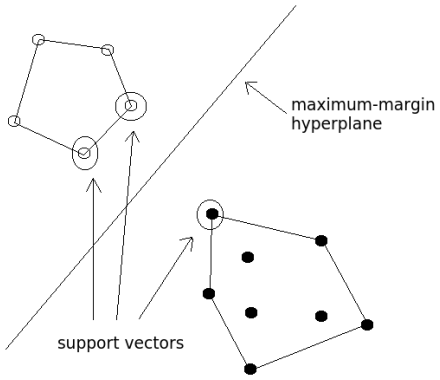

Figure.3.1: Maximum-Margin Hyperplane and the Support Vectors.

When data is linearly-separable, there are many separating lines that could be chosen. Such a hyperplane can be expressed as $W \cdot X + b - 0$ where W is a vector of weights, b is a scalar bias, and X are the training data (of the form (x1, x2)). If our bias, b, is thought of as an additional weight, the equation can be expressed as $w_0 + w_1x_1 + w_2x_2 - 0$ which can, in turn, be rewritten as a pair of linear inequalities, solving to greater or less than zero, either of which satisfied indicate that a particular point lies above or below the hyperplane, respectively. Finding the maximum-margin hyperplane, or the hyperplane that resides equidistance from the support vectors, is done by combining the linear inequalities into a single equation and transforming them into a constrained quadratic optimization problem, using a Lagrangian formulation and solving using Karush-Kuhn-Tucker conditions. Following this transformation, the maximum-margin hyperplane can be expressed in the form $x = h + \sum_{i=1}^{n} \alpha_i y_i a(i) \cdot a$ where b and αi are learned parameters, n is the number of support vectors, i is a support vector instance, t is the vector of training instances, yi is the class value of a particular training instance of vector t, and a(i) is the vector of support vectors. Once the maximum-margin hyperplane is identified and training is complete, only the support vectors are relevant to the model as they define the maximum-margin hyperplane; all of the other training instances can be ignored.

When data is not linearly-separable, the data is first transformed into a higher dimensional space using some function, and this space is then searched for the hyperplane, another quadratic optimization problem. In order to

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

circumvent the additional computational complexity that increased dimensionality brings, all calculations can be performed on the original input data, which is quite likely of lower dimensionality. This higher computational complexity is based on the fact that, in higher dimensionality, the dot product of an instance and each of the support vectors would need to be calculated for each instance classification. A kernel function, which maps an instance to feature space created by a particular function it applies instances to, is used on the original, lower-dimensionality data. A common kernel function is the polynomial kernel, which expressed computes the dot product of 2 vectors, and raises that result to the power of n. SVMs are powerful, well-used classification algorithms, which garnered a substantial amount of research attention prior to the deep learning boom we are currently in. Despite the fact that, to the onlooker, they may no longer be bleeding edge algorithms, they certainly have had great success in particular domains, and remain some of the most popular classification algorithms in the toolkits of machine learning practitioners and data scientists.

II. RELATED WORK

In this paper, three novel algorithms have been proposed for processing multi-source relative skyline queries in road networks. It is not only the first effort to process relative skyline queries in road networks, but also the first study on skyline queries by considering relative network distances to multiple query points at the same time. LBC is proven to be instance optimal in terms of the network search space over all algorithms where network distances are computed by expanding the searching region from query points without using pre-computed distance information. Our experiments confirmed that LBC has the best performance consistently for various test settings. The path distance lower bound approach, based on which LBC is designed, can be applied to benefit other types of road network queries where network distance comparison is needed.[10]

The R-tree structure has been shown to be useful for indexing spatial data objects that have non-zero size Nodes corresponding to disk pages of reasonable size (e g 1024 bytes) have values of A4 that produce good performance With smaller nodes the structure should also be effective as a main-memory index, CPU performance would be comparable but there would be no I/O cost. The linear node-split algorithm proved to be as good as more expensive techniques It was fast, and the slightly worse quality of the splits did not affect search performance.OSM has its own geology crosswise over time and space. At the end of the day, seldom observe indistinguishable examples of commitments in two distinct locales/nations. When talking about OSM quality and commitments arranges the significance of concentrate different contextual investigations has been featured. Thus, in this area, two distinct maps are produced from the Open Street Map insights, which exhibit the heterogeneity of Open Street Map in various nations. This guide shows a topical arrangement of made hubs, which is one of the key components in estimating Open Street Map commitments. It ought to be noticed that in this similar report, the span of the nation, populace, total national output Gross Domestic Product and various other physical qualities of the nations are not contemplated. In any case, they are of incredible significance in performing further top to bottom investigation.[14]

This paper thinks about the combined total closest neighbor(MANN) inquiry. Weiweisun build up a calculation for handling this question, the Quick Pruning calculation. It utilizes the Euclidean total separation between an objective point and the question set as the pruning separation to prune away superfluous target focuses, the analysis comes about demonstrate that it can dispose of a significant piece of target focuses which thusly spare the execution time and I/O cost., The ideal area inquiry issue in view of street systems. unambiguous to a street arrange on which a few customers and servers are found. Every customer find out the server that is nearest to her for administration and her cost of getting served is equivalent to the(network) separate between the customer and the server serving her duplicated by her weight or significance. The ideal area inquiry issue is to find an area for setting up another server with the end goal that the greatest cost of customers being served by the servers (including the new server) is limited. This issue has been contemplated some time recently, however the cutting edge is as yet not sufficiently proficient. In this paper, creator propose an effective calculation for the ideal area question issue, which depends on an original thought of closest area part. They additionally talk about three augmentations of the ideal area question issue, in particular the ideal different area inquiry issue, the ideal area inquiry issue on 3D street systems, and the ideal area question issue with another target. Broad examinations were directed which demonstrated that our calculations are speedier than the best in class by no less than a request of extent on vast genuine benchmark datasets. For instance, on our biggest genuine datasets, the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

cutting edge kept running for over 10 hours however our calculation kept running inside 3 minutes just (i.e., >=400 times quicker).[13]

III. PROPOSED SYSTEM

The common approach to semantic parsing is a manual annotation of a corpus with natural language utterances and machine readable formulae which are then used to learn the structure and weights of a semantic parser using SVM.

While GEOQUERY queries are restricted to the closed domain of US geography, the structural complexity of the questions is higher than for FREE917, which focuses on open domain queries. Seminal work on building semantic parsers from the GEOQUERY meaning representations are Zettlemoyer and Collins (2005) or Wong and Mooney (2006). Later approaches try to learn semantic parsers from question-answer pairs only, for example, Liang et al. (2009) for GEOQUERY, or Kwiatkowski et al. (2013) or Berant et al. (2013) for FREE917. Newer research attempts to close the gap between lexical variability and structural complexity (Vlachos and Clark, 2014; Artzi et al., 2015; Pasupat and Liang, 2015), however, answer retrieval accuracy is low if semantic parsers cannot be bootstrapped from a corpus of queries and MRLs (Wang et al., 2015; Pasupat and Liang, 2015). Their approach treats semantic parsing as a monolingual machine translation problem in which natural language is translated into the machine readable language. This approach is convenient because one can make use of the efficient and robust decoders that are freely available for SMT. Despite the simplicity of the approach, Andreas et al. (2013) have shown that highly accurate semantic parsers can be trained from annotated data. OSM has previously been used by Boye et al. (2014) for pedestrian routing using a dialogue system, however, no details on semantic parsing and no resource are provided. Our SMT tuning experiment builds on the work of Riezler et al. (2014) and Haas and Riezler (2015) who applied response-based learning for SMT to the GEOQUERY and FREE917 domains, respectively.

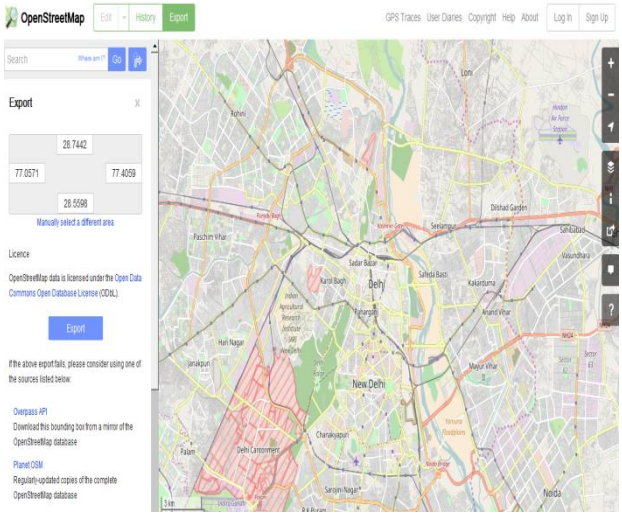

Figure 1: Open Street Map

users 2,389,374
objects 3,464,399,738
nodes 3,139,787,926
ways 320,775,580
relations 3,836,232
tags 1,259,132,137
distinct tags 76,204,309
distinct keys 57,159

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

Above : Statistics of OSM as of December 14th, 2016, Open Street Map is presented as another information base that has not, to the best of our insight, been utilized for question replying, and offer another corpus to the exploration group. Work assembles the premise of a characteristic dialect interface to Open Street Map that will empower for fascinating bearings of future research, e.g., reaction based figuring out how to enhance parsing and multilingual database get to with efficient and accurate results.

IV. CONCLUSION

Under the proposed scheme we will deliver significant scientific outcomes, which will stimulate international research networking and collaboration. As outlined above, the inherent cross-disciplinary essence of OSM research combined with the emerging data quality, data mining, and patterns determination approaches to analysis of OSM using machine learning techniques i.e. SVM and Semantic Expat Parsing . We believe that, this inter-disciplinary contributions permit a deeper understanding of how the OSM works and will become the phenomenal success for future for prompt and accurate information anytime and anywhere scenario. Last, but not least, this scheme will strive to bring OSM into the core of GIScience where the diverse world of new and classical geography and cartography will meet requirements of users and customers. The above proposed system will form the effective and accurate geo-spatial extraction mechanism and will evaluated with existing schemes thereinafter to propose the better ecosystem for OSM globally.

REFERENCES

1. Z. Chen, Y. Liy, R. C.-W. Wong, J. Xiong, G. Mai, and C. Long.Efficient algorithms for optimal location queries in road networks.In SIGMOD, 2014.
2. A. Eldawy and M. F. Mokbel. A Demonstration of SpatialHadoop: An Efficient MapReduce Framework for Spatial Data (System Demo). In VLDB, Riva del Garda, Italy, Aug. 2013.
3. P. Mooney and P. Corcoran.Characteristics of heavily edited objects in openstreetmap.Future Internet, 2012.
4. L. Alarabi, A. Eldawy, R. Alghamdi, and M. F. Mokbel. TAREEG: A MapReduce-Based Web Service for Extracting Spatial Data from OpenStreetMap (System Demonstration). In SIGMOD, pages 897–900, Snowbird, UT, June 2014
5. X. Ma, S. Shekhar, and H. Xiong. Multi-type nearest neighbor queries in road networks with time window constraints. In SIGSPATIAL GIS, pages 484–487, 2009.
6. L. Zhu, Y. Jing, W. Sun, D. Mao, and P. Liu.Voronoi-based aggregate nearest neighbor query processing in road networks. In SIGSPATIAL GIS, pages 518–521, 2010.
7. L. Wu, X. Xiao, D. Deng, G. Cong, A. D. Zhu, and S. Zhou. Shortest path and distance queries on road networks: An experimental evaluation. PVLDB, 5(5):406–417, 2012.
8. S. Vanhove and V. Fack.An effective heuristic for computing many shortest path alternatives in road networks. International Journal of Geographical Information Science, 26(6):1031–1050, 2012.
9. Durgesh K. Srivastava, LekhaBhambhu Data Classification Using Support Vector Machine,Journal of Theoretical and Applied Information Technology 2005 - 2009 JATIT.
10. Graham, M.; Hale, S.; Stephens, M. Digital Divide: “The Geography of Internet Access. Environ. Plan”. A 2012, 44, 1009–1010.
11. JokarArsanjani, J.; Helbich, M.; Bakillah, M.; Loos, L. “The Emergence and Evolution of OpenStreetMap: A Cellular Automata Approach”. Int. J. Digit. Earth 2013 b, 00, 1–15
12. Weiwei Sun, Chong Chen, “Merged aggregate nearest neighbor query processing in road network”, Singapore Management University Institutional Knowledge at Singapore Management University, 10-2013.
13. Zitong Chen, Yubao Liu, Raymond Chi-Wing Wong, “Efficient Algorithms for Optimal Location Queries in Road Networks”, The Hong Kong University of Science and Technology, Hong Kong, China.
14. Aye Su Yee Win, “Fast Algorithm for Multi-type NearestNeighborQuer”, Graduate School of Science and Engineering, Saitama University, D-002.
15. Lingkun Wu, Xiaokui Xiao, “Shortest Path and Distance Queries on Road Networks: An Experimental Evaluation”, School of Computer Engineering ,Nanyang Technological University, Singapore.
16. A. Eldawy and M. F. Mokbel. Pigeon: A spatial mapreduce language. In ICDE, pages 1242–1245, 2014