Machine Learning to Improve Retrieval by Category in Big Volunteered Geodata*

Alex Sorokine, Gautam Thakur, Rachel Palumbo
Oak Ridge National Laboratory
Oak Ridge, Tennessee
{sorokine,thakurg,palumborl}@ornl.gov

ABSTRACT

Nowadays, Volunteered Geographic Information (VGI) is commonly used in research and practical applications. However, the quality assurance of such a geographic data remains a problem. In this study we use machine learning and natural language processing to improve record retrieval by category (e.g. restaurant, museum, etc.) from Wikimapia Points of Interest data. We use textual information contained in VGI records to evaluate its ability to determine the category label. The performance of the trained classifier is evaluated on the complete dataset and then is compared with its performance on regional subsets. Preliminary analysis shows significant difference in the classifier performance across the regions. Such geographic differences will have a significant effect on data enrichment efforts such as labeling entities with missing categories.

CCS CONCEPTS

• **Information systems → Relevance assessment**; • **Human-centered computing → Social tagging systems**; • **Computing methodologies → Cross-validation**;

KEYWORDS

machine learning, crowd-sourcing, natural language processing

ACM Reference Format:
Alex Sorokine, Gautam Thakur, Rachel Palumbo. 2018. Machine Learning to Improve Retrieval by Category in Big Volunteered Geodata. In *12th Workshop on Geographic Information Retrieval (GIR'18), November 6, 2018, Seattle, WA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3281354.3281358

1 INTRODUCTION

Volunteered Geographic Information (VGI) has become a ubiquitous source of data for GIScience research and multiple on-line services. VGI provides a number of advantages over traditional data collection such as large amounts of data at low cost with higher update speed and larger spatial coverage. The major drawback of the VGI is very uneven and unpredictable data quality. The challenge of assuring VGI quality has been recognized since the inception of the term itself [2]. The size of the VGI has been viewed as a key for solving this problem on the assumption that the majority of the collected data is correct and that it would significantly outweigh the erroneous data. However, the large volume of data (sometimes as high as hundreds of millions of records) and the speed of update put this task outside of the capabilities of manual processing and refinement. Here we demonstrate the use of Machine Learning (ML) and Natural Language Processing (NLP) for quality assessment of categorization of places in Wikimapia (https://wikimapia.org) — one of the largest and longest running VGI projects on the Internet.

2 VGI QUALITY CHALLENGE

Approaches to quality control differ among the VGI projects. Large commercial companies like Facebook, Google, or Yelp use a combination of paid moderators, volunteers, and more recently artificial intelligence algorithms to verify and filter the contributions. None-commercial projects like Wikipedia (https://wikipedia.org), Open Street Map (OSM, https://openstreetmap.org), and Wikimapia seldom have resources for full-time editors but instead rely on volunteers for quality control and improvement. As a result there is a significant demand for better quality control and improvement procedures as manual verification does not scale up to the size and volume of the existing VGI.

One of the common problems in VGI is the ability to retrieve the data records based on their categories or tags. Most of VGI projects are not organized into layers like traditional geographic information systems but instead each contributed feature is assigned one or several categories or tags to designate its semantics. Inconsistencies in the labeling may result from many reasons including linguistic and cultural differences among contributors, changing website policies, or fluctuations of the public interest in the specific topics. Resulting labels may show significant dissimilarities in semantics and quality between regions, acquisition time, and types of objects. The solution for the problem may be found in additional information associated with the features such as location and different kinds of free text descriptions. Recent progress in ML and NLP open the possibility to tap this textual information for assessing the quality and improving the labeling.

There is a noticeable body of research concerned with the VGI quality and offering a number of techniques for solving the problem [3, 7, others]. To our knowledge, NLP and semantics methods

*Notice: This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

Publication rights licensed to ACM. ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of the United States government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.
GIR'18, November 6, 2018, Seattle, WA, USA

https://doi.org/10.1145/3281354.3281358

have been applied for the analysis of the VGI and for tag recommending systems during edits but not in the context of data retrieval and quality assessment in the regional scope [1, 4, 8].

3 MATERIALS AND METHODS

In this study, we investigate application of ML and NLP methods to improve and assess the quality of category labeling in the subset of data collected by the Wikimapia project. As of August 2018, the Wikimapia website claims to have collected close to 30 million records of places for all parts of the world. The project was started in 2006 and its main goal is "marking all geographical objects in the world and providing a useful description of them."[1] Each object in Wikimapia is geolocated and assigned following attributes: name, description, street address, one or more category tags, ground-based photos. Categories and tags are organized in a quasi-hierarchical system in which each category may have multiple sub- and/or super-categories.

Here, we use a subset of about 4 million records for Russian Federation and Kazakhstan. These countries have the highest density of the Wikimapia features over large and diverse areas. Data preparation required several steps. About 35% of the records were removed as they did not have category labels. Then, we performed automatic language detection[2] to separate records in different languages. Most of the records in our subset were in Russian and they were used in the analysis. We have also filtered out a small number of abnormal entries that were either very long or did not look like natural language. The records were normalized by squashing all none-word characters and converting all other characters to lower case. The resulting set of 2.4 million records was split into training (80%) and testing subsets (20%).

Multinomial naïve Bayes classifier [5, http://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html] from [6] was trained to predict category labels from the entity names and then standard performance metrics (precision, recall, and F-score) were calculated. This type of classifier is based on the assumption that the probability of an entity e to belong to category c is proportional to the probability of c and the product of probabilities of occurrence of each word in e in c. The category with the highest probability is considered to be the best match.

First, the classifier was trained on a complete training subset and evaluated on the complete testing subset (row 1 in Table 1). Then the training and testing sets were each split by 94 administrative regions with at least 2,000 records in the training set. Performance of the classifier trained on the complete dataset was evaluated for testing subset of each region and are shown in rows 2–4. Finally, the classifier was separately trained on training subsets of each region and evaluated on their testing subsets (rows 5–7).

4 PRELIMINARY RESULTS

Table 1 shows significant regional differences in the classifier performance thus such differences have to be taken into account when assessing reliability of assigning missing labels from the entities names. The classifier trained on the complete training subset shows marginally better performance than the classifiers trained within separate regions. This can be explained by a wider variety of terms and categories encountered in the complete training subset. Table 2 shows the metrics for several of the most frequent categories. Some categories are detected very reliably but others show poor results. These differences will be investigated beyond the pilot stage of the project along with other improvements like using more attributes, better normalization, and training other classifiers.

Table 1: Classifier Performance in the Regional Scope

			Precision	Recall	F-score
(1)	Global		0.61	0.63	0.59
(2)	By region	best	0.79	0.79	0.78
(3)		average	0.60	0.63	0.59
(4)		worst	0.44	0.47	0.43
(5)	Local	best	0.73	0.78	0.74
(6)		average	0.52	0.61	0.53
(7)		worst	0.32	0.43	0.33

Table 2: Classifier Performance for Frequent Categories

Category	Precision	Recall	F-score
House	0.57	0.81	0.67
Apartment complex	0.60	0.64	0.62
Village	0.66	0.52	0.59
Lake	0.84	0.97	0.90
Electric pylon	0.97	1.00	0.99
Shop/Store	0.41	0.81	0.55

[1] http://wikimapia.org/docs/About_Wikimapia
[2] https://github.com/saffsd/langid.py

REFERENCES

[1] Ahmed Loai Ali, Falko Schmid, Rami Al-Salman, and Tomi Kauppinen. 2014. Ambiguity and plausibility: managing classification quality in volunteered geographic information. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL '14*. ACM Press, Dallas, Texas, 143–152. https://doi.org/10.1145/2666310.2666392

[2] Michael F. Goodchild. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69, 4 (Aug. 2007), 211–221. https://doi.org/10.1007/s10708-007-9111-y

[3] Michael F. Goodchild and Linna Li. 2012. Assuring the quality of volunteered geographic information. *Spatial Statistics* 1 (May 2012), 110–120. https://doi.org/10.1016/j.spasta.2012.03.002

[4] Yingjie Hu and Krzysztof Janowicz. 2018. An empirical study on the names of points of interest and their changes with geographic distance. *arXiv:1806.08040 [cs]* (June 2018). http://arxiv.org/abs/1806.08040 arXiv: 1806.08040.

[5] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

[6] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (Oct. 2011), 2825–2830. http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[7] Hansi Senaratne, Amin Mobasheri, Ahmed Loai Ali, Cristina Capineri, and Mordechai (Muki) Haklay. 2017. A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science* 31, 1 (Jan. 2017), 139–167. https://doi.org/10.1080/13658816.2016.1189556

[8] Arnaud Vandecasteele and Rodolphe Devillers. 2015. Improving Volunteered Geographic Information Quality Using a Tag Recommender System: The Case of OpenStreetMap. In *OpenStreetMap in GIScience: Experiences, Research, and Applications*, Jamal Jokar Arsanjani, Alexander Zipf, Peter Mooney, and Marco Helbich (Eds.). Springer International Publishing, Cham, 59–80. https://doi.org/10.1007/978-3-319-14280-7_4