A Review on Vandalism Detection in OpenStreetMap Data and Its Emerging Trends

Nitasha Singla¹, Sukhjit Singh Sehra², Jaiteg Singh³

¹ Student, Guru Nanak Dev Engineering College, Ludhiana, Punjab, India
² Assistant Professor, Guru Nanak Dev Engineering College, Ludhiana, Punjab, India
³ Associate Professor, School of Computer Science, Chitkara University, Punjab, India

Email address:
natashasingla1987@gmail.com (N. Singla), sukhjitsehra@gmail.com (S. S. Sehra), jaitegkhaira@yahoo.co.in (J. Singh)

To cite this article:

Abstract: More than the last few period technological enlargements incorporate the increase of online mapping. Next to that, OpenStreetMap project, a standard basis of generously data composed via different cases, have experience a reliable raise in status in modern time. Solitary, key requirement to facilitate is directly connected to this status is raise in dissimilar kind of defacement. Appropriateness and Consistency of peer-production in accumulation to the accomplishment of universe are the victims of vandalism. The query still is: By what means OSM project shelter its dignity in opposition to statistics vandalism? The resolution of this query is analyzed in this paper using diverse instances and identification of vandalism enhancing by tools.

Keywords: Vandalism, OpenStreetMap, Volunteered Geographic Information (VGI), WikiMapia

1. Introduction

Volunteered Geographic Information (VGI) is the recent empowerment [26] of citizens in the mutual collection of geographic information [26]. Continuation of analytical and democratic applications is the expansions of VGI. Significantly, by inspiring those in the direction to make to some extent of your individual accord, this process are extreme economical and its artefacts are quietly accessible. These days’ geographical facts are obtainable to the public through increasingly dynamic geographical websites. The change in terminology is supported on a shift in the usage of the network, which is no longer described by the consumption of predefined content. Actually the term Web 2.0 relates to a new platform where users can adapt their own applications on the WWW to meet their individual plan, thoughts plus functionality and, most importantly, can create their own data or edit existing data. The online encyclopedia Wikipedia, established in 2001, is based precisely on this fact [7]. The newly created information is referred to as user-generated content or user-created content [7, 28]. The voluntary users, who are spread all over the world, share their data on various topics on one particular online platform [7]. This new development became admired under the term Volunteered Geographic Information [1] or crowd-sourced geodata [29]. Other websites that support alike approach allow users to share their videos and photos with others [7]. Similar efforts are the basis of geodata platforms such as OpenStreetMap. Initially look for the arrangement of an overall mesh plan, the scheme rapidly bowed distant forward. Growth in rising society and requirement for open geodata has made OSM the necessity and easily available maps. In spite of the fact, consumers of OSM have to catalogue ue preliminary contribution, the OSM replicas have been planned in an open-access loom. Still, this unwrap loom can be measured as the input to OSM’s achievement; it can also be a font of a range of troubles. Whereas the majority of offerings are valid, a little molest by unsolicited end result in vandalism. One of the most popular examples for vandalism in OSM is the case of two employees of a popular search engine deleting OSM features in the London area [33]. However, there has not yet been a general investigation on the impact and quantity of vandalism in OSM, and no different task force against vandalism has been implemented so far in OSM. The OSM Data Working Group can be contacted when serious Disputes and Vandalism are encountered [3]. Yet, Minor incidents of vandalism should be contract with by the local community [3]. Even so, pursuing the initiative of shared facts, it is unsaid that vandalism is perceived and approved by additional OSM providers in a proposed phase. Next to this time occurrence of OSM,
vandalism can happen on purpose and accidental, say the opposite to the predictable justification of the expression vandalism. Conversely, this stoutly depends on amend itself, as well as on the region in which amend has been carry out; vandalism in an urban locale will perhaps be discover and acceptable pretty rapid while vandalism in a very countryside locale will potentially stay for a very extensive phase. In general, data validation and vandalism recognition desires to be illustrious from each one. While data validation comprise dissimilar mode for superiority, vandalism center on lively data fraud. In this article, center of interest is on the last one. Even as usual method for the judgment of vandalism in OSM have to be of concentration for mutually the provider as well as the patrons of the plan, just a small number of tools or added important expansion have been consummate in this ground. Frequently tools give the prospect to catalog to an RSS feed [7], which give in turn about the newest transform in a dissimilar locale. The functionality of the tools can be contrast to watch lists in Wikipedia, which follow change that were made to preferred item. Comparing OSM and Wikis will ultimately provide the result that both are equivalent and citizens that become the sources of data [5]. Familiarity between all these tools is that they update recorded users about each solo change in a predefined region. The consumer at rest has to examine each solo edit. Thus, the key area which needs to be investigated in this current review is analysis of vandalism and expansion of programmed diagnosis of vandalism.

The remainder of this article sheds light on these issues as follows. Section second discusses the literature review. Third section describes the different types of vandalism and the methodology that can occur in OSM. Section fourth proposes a detection of vandalism using proposed tools. The last two sections discuss results and various implications for upcoming effort and analysis.

2. Literature Review

Originated by Steve Coast in 2004 in UK, OSM project [9, 33 and 35] is a collective effort to built open records of the globe. Limitation on utilization and accessibility of record are the two main powerful strengths following the OSM project. A huge range of unusual kind of spatial facts such as roads, buildings and land use region are composed in the database. Subsequent the Blogging platforms 2.0 approach of mutual creation of huge data, any user will start contributing to the project after having a short online signing up. Essentially, every registered user has the ability to add new aspects, as well while alter or erase existing ones. In the time writing, the OSM task had almost 1.4 million registered associates, who contributed nearly 2.1 million points and 220 million lines that happen to be moderately based in 3.6 billion GPS points which are uploaded. In total, almost 200,000 members made a minimum of one edit, and roughly 3% off members made a minimum of one change per month towards database by the finish of 2011. Several different experiments in regards to the excellent and completeness connected with OSM data in contrast to other data resources are published in recent times [8, 9-12]. With Europe, OSM shows a sufficient level of data coverage for urban areas, which allowed the development and syndication of map products and services or other software, such as Place Based Serviced. Nevertheless, less populated areas usually do not show the same completeness level throughout OSM, which creates the dataset difficult to rely on in those parts. Thus, OSM data can be quite use-case dependent, and the requirements must be carefully considered [13]. The difficulty of understanding the excellence of ecological catalog was recognized diverse time before and expected concentration. Van Oort identifies various quality standards:

- Lineage- lineage covers the pat aspect of dataset, focus on its advancement and composition.
- Positional Accuracy- positional accuracy expresses the superiority estimation and synchronization of objects.
- Attribute Accuracy-attribute accuracy not only checks characteristic shapes and features but also the correctness of values.
- Completeness-checks the estimation of objects which are actually to be presents in data but are absent.
- Semantic Accuracy-semantic accuracy focus on the characterization and confinement of objects in the database.
- Usage, purpose and constraint-this is fitness-for-purpose statement which explains about its usage.
- Temporal Quality- validation of real time data.

Therefore, OpenStreetMap data accomplished scenarios reliant, and the necessities ought to be cautiously measured. Likely recognized over, each area affiliate preserves the current OSM database; conversely alteration is not sustained. There exist vital dissimilarity among Wikipedia and OSM and carry slightest a little benefits; OSM users are acknowledged by their usernames, which can be barren. In Wikipedia, users are recognized by username or IP-address and more than single user may use the same IP-address. Registered users can add data to OSM in different ways. The typical approach should be to gather data using a GPS receiver, which afterwards can be edited with on the list of various freely available editors, such since Potlatch or JOSM. Considering that November 2010, users are explicitly allowed to trace data by Bing aerial photos and add the data to OSM [14]. This allows a user to collect information without physically being at a distinct location. Regardless of how the data is collected, users can provide additional information, such as street names or building types, about the different OSM features. In the ten years since its initiation, 2,526,682,899 geo-referenced points (nodes in OSM terminology), 252,573,744 ways (both line strings and simple polygons) and 2,806,316 relations (for describing relationships, such as turn restrictions or complex polygons with holes) [32] have been collected as of today (September 2014) [2, 16]. Every entity holds account information containing identification number, name containing first, middle or last name of the one who edit its preceding and the date of previous change. Adding together, tag-value pairs
holding extra information join every feature. A few changes ended by a user accumulated in a change set. If a user wishes to implement the information of the OSM project for an application, a earth file, which contains all information with the latest database with the project, can possibly be downloaded [16]. As data need to be updated within a short interval of occasion, people keep updating almost every moment. Renewal of info is necessity otherwise it will eventually led to the exploration of OSM data consequently OSM project allow OSM Change files that includes data for position outline, 24*7 data. On average, nearly 700 new members have registered to the project each day between January and March 2012 According to [7]. 30% of newly registered contributors will become active contributors. Each day in 2012, 230 new OSM members started contributing to OSM. Table 1 contains the average number of edits per OSM object per day between January and June 2012.

<table>
<thead>
<tr>
<th>Number of ...</th>
<th>Node</th>
<th>Way</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daily created objects</td>
<td>1,200,000</td>
<td>130,000</td>
<td>1,500</td>
</tr>
<tr>
<td>Daily modified objects</td>
<td>170,000</td>
<td>70,000</td>
<td>3,500</td>
</tr>
<tr>
<td>Daily deleted objects</td>
<td>195,000</td>
<td>16,000</td>
<td>280</td>
</tr>
<tr>
<td>Users who daily edited</td>
<td>2,000</td>
<td>1,940</td>
<td>560</td>
</tr>
</tbody>
</table>

In view of these statistics for estimation prospect dispensation workloads for our optional tool, it can be resolute that every minute, 830 node creations, 190 node modifications and 135 node deletions will have to be performed. Besides, 90 new, 48 modified and 11 deleted ways and one new, two modified and 0.2 deleted relations have to be processed. Those numbers can apparently differ during the day, but they give a first warning on how much data will be edited.

Figure 1. Example of Graffiti vandalism in OSM in Zwijndrecht (The Netherlands) [18].
3. Vandalism

The term vandalism refers to a variety of socially constructed phenomena, and no clear academic consent has been established about its scope. Reviewing its much definition, Moser suggested that vandalism is a Hodge podge concept that covers behavior for which motivations are extremely different. The definitions of vandalism differ as they take into account the caused damage, the motivation of the human actor, the context of the incident. Vandalism is an ever-present and visible social phenomenon, in which intentional damage was performed on a variety of objects, including buildings, public toilets, vehicles, furniture, infrastructures, and works of art such as paintings, monuments, and sculptures. Most vandalism incidents are far from being random, absurd acts, and several competing theories to explain its reasons have been proposed. A consensus exists around the general meaningfulness of vandalism as a form of social communication between the offender and an imagined ordeal audience. The open approach connected with data collection inside OSM project can cause a number of types of vandalism. It is also possible that a factor purposely or unexpectedly makes changes to the dataset that usually are harming the projects absolute goal. Common vandalism types that can be found in the actual OSM geodetic data source are (based on [17])

- A new object with no commonly used attributes.
- A non-regular geometrical modification of an object.
- A non-common modification of the attributes of an object.
- Randomly deleting existing objects.
- An overall abnormal behavior by a contributor.
- Generating fictional and non-existing objects.
- Inappropriate use of automated edits in the database.
- Application of mechanic edits.

Previous research has identified many common types of vandalism. Viégas et al. [22] identified five common types of vandalism: mass deletion, offensive copy, phony copy, phony redirection, and idiosyncratic copy. Priedhorsky et al. [23] categorized Wikipedia damaged edits2 into seven types: misinformation, mass delete, partial delete, offensive, spam, nonsense, and other. The categories proposed in these papers were not developed systematically, and can be made more comprehensive. Potthast et al. [11] organized vandalism edits according to the “Edit content” (text, structure, link, and media) and the “Editing category” (insertion, replacement, and deletion). This organization does not consider the scale of editing, which correlates with the difficulty level of vandalism detection.

Figure 1 explains a paradigm of Graffiti vandalism in 2011 in Zwijndrecht (The Netherlands). The consumer who origin this chaos of the characteristics use the Potlatch OSM editor, which straightly affect the transform to the exist OSM database.

Potthast et al. [19] and West et al. [20] by hand examined vandalism in Wikipedia to study in relation to the exact uniqueness. Equivalent approach is hand analyzed 204 user blocks of the project to gather better information about vandalism inside OSM. Members of the actual OSM Data Doing the job Group or moderators are allowed to block other OSM members for a brief time period of time (between 0 and 96 h) [21]. Table 2 summarizes results of by hand composed vandalism cases and their characteristics.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fictional Data</td>
<td>33.3%</td>
<td>The user created some fictional data</td>
</tr>
<tr>
<td>Editing Data</td>
<td>33.3%</td>
<td>The user modified some existing data, e.g.,</td>
</tr>
<tr>
<td>Deleting Data</td>
<td>43.1%</td>
<td>The user deleted some existing data</td>
</tr>
<tr>
<td>New User</td>
<td>76.4%</td>
<td>The data was vandalized by a new project member</td>
</tr>
<tr>
<td>Potlatch Editor</td>
<td>82.4%</td>
<td>OSM editor which was used during the vandalism</td>
</tr>
</tbody>
</table>

4. Vandalism Detection

Figure 2. OSM & OSMPatrol Architecture [2]
The revealing of the found kind of defacement in the catalog permitted for the ideal implementation of system which is based on deriving rules known as OSM Patrol. Notion behind developing this type of model is to trace vandalism at early stages. As illustrated above modification to database is made at each step so for these OSM-Diff files is the main necessity. Any kind of extra information tagging the specific OSM characteristic with key value pairs can be given by user. OSM Map Features, help to judge odds against extra information in vandalism. Table index information is as followed:

- Number of nodes, number of ways and number of relations OSM contributor creates.
- Registration of date
- Common tags

Following tables are accumulated in OSM PostgreSQL database. To help retrieve the OSM Diff-files that incorporate the changes which are made to the database each and every minute, the OSMOSIS [25] application is applied. OSMOSIS is usually an open-source command line JAVA tool, which processes OSM data in lots of different ways. Figure 2 shows the entire architecture of the developed prototype in terms of the OSM task architecture [2].

In order to accomplish the procedure first need to download current OSM-Diff file using OSMOSIS. Then OSM Patrol examines the file and detects vandalism. In order to incorporate testing techniques every modification of the selected file tool demand for two lists which include information of users those are blacklisted in past, information of users whose consideration is not taken. While assessment phase of every single modification is taken reputation of user and uniqueness of attributes may be used. In the case of vandalism traced, edit will be kept in additional table. OSMOSIS tool will revise OSM data with Diff file the moment all modification done.

5. Experimental Results

As per [27], After testing the developed prototype for small areas with known heavy and light vandalism cases, conduct for our final analysis by running the prototype on a dedicated server for one week (14 August 2012–21 August 2012) was considered. During the testing phase, OSM Patrol detected about seven Mio vandalism edits of 9,200 different users for the entire week [2]. During the same time frame, around 16 Mio edits were made to the OSM database [2].

The following Figure 3 exhibits the distribution on the affected amounts regarding nodes, ways as well as relations. Additionally, the figure provides details about how many of the affected objects were detected after a creation, modification or deletion.

Seeing that described by Neis and Zipf [7], this week basically represents an average week (regarding factor behavior), meaning that the actual OSM members contribute to the project in a similar fashion every other full week. The following Physique 4 shows the actual distribution of edits that were detected as vandalism good user reputation [2]. About 50% from the users, for which often OSMPatrol detected any possible case connected with vandalism, have a user reputation larger than 66%, indicating that additionally experienced contributors' actions could possibly be recognized as vandalism. Using the collected results, users which have a reputation level larger than 66% committed 48% from the detected possible vandalism conditions. According to these kinds of values, about 43% off detected vandalism edits were being committed by new users from the project with a decreased reputation. Overall, almost 1/3 (36%) off vandalism users were being new users.

6. Discussion

During the implementation part of rule-based product few troubles arises. Few those have not projected caused by economic reasons incapable of carryout minute vandalism defacement. User popularity was significant perspective to add the challenge membership moment slot moreover the reslovance regarding user blockages. Whole integration regarding project won't highlight almost any knowledgeable solution, reason guiding this observed most consumer register without having modifying data. User come after a few years and can their really new customization. Another in question indicator pertaining to vandalism could be the evaluation on the version number of an OSM function. When developing a new function, the variation number is defined to a single. The variation is incremented having each change of the distinct item (regardless on the actual change). Although, is an

![Figure 3. Distribution of objects and edit-types in the detected vandalism (14–21 August 2012) [2].](image-url)
alteration on an object using a high variation number more probable a type of vandalism than a change with an object using a low variation number? Why don't you consider the other situation? A few investigations revealed that there are so-called heavily modified objects [26] with OSM, but it isn't known when changes on those objects will be more likely vandalism or maybe not. As opposed to the above mentioned difficulties, some vandalism linked aspects, such as the IP deal with, cannot become implemented inside client caused by missing files (it seriously isn't possible to get together the IP address of OSM contributor). Even so, having these kinds of information could possibly be a good (additional) indicator for that vandalism possibility. It can be investigated if your IP information could possibly be a good (additional) indicator for your IP address of OSM contributor. Even so, having these kinds of information could possibly be a good (additional) indicator for that vandalism possibility. It can be investigated if your IP (and the actual access point) suits the region that change continues to be performed. By way of example, if any user modifications a street inside a country which is hundreds kilometers apart, this could possibly be more most likely vandalism as compared to changes of your street name near to the admittance point of your user, comparable to what can be described by means of West et ing [20].

7. Conclusions and Future Work

This article has offered a discussion from the phenomenon salient capabilities, inspiration, and also the current approaches carry out to maintain it in cove along with social along with technological detection and handle mechanisms intended for quality guarantee and affirmation. Following review focus on the early on researches built on vandalism, existing OSM information and successes following Wikipedia vandalism detection tools. Idea driving study is not just validation but on the defacement of vandalism as you will find few confinements. In support of upcoming do the job, enhancement of API of developed prototype are going to be top-most goal. With the help of distinct user interface using obtained results new application developers perform on the particular conclusions. Enabling contributors like a Patrol intended for specific locale may be on the list of application, by using this application factor can find specific location simultaneously towards the OSM Patrol deface vandalism. To uphold well explained white-list of contributors which usually show improvements as sounding vandalism or perhaps not will also be another program.

References


[35] Mooney, P.; Sun, H.; Corcoran, P.; Yan, L,"Citizen-Generated Spatial Data and Information: Risks and Opportunities". In Proceedings of IEEE International Conference on Intelligence and Security Informatics, Beijing, China, 10–12 July 2011